## SPECIFIC AIMS

The necessity for understanding gene regulation in human brain development is supported by several recent discoveries. For example, most inherited common genetic variation underlying neuropsychiatric diseases lies in non-coding regions and is presumed to exert pathogenic effects via the regulation of gene expression and splicing[1-4]. Additionally, most non-inherited (*de novo*) highly penetrant ASD risk genes are enriched in co-expression modules and protein interaction networks related to chromatin remodeling and transcriptional regulation[3-8]. Moreover, a specific shared pattern of transcriptional dysregulation is observed in the cerebral cortex in slightly more than 2/3 of post-mortem ASD cases[9,10]. Taken together, these observations emphasize the importance of integrating transcriptomic and epigenomic data with higher-order chromatin interactions to better understand the putative mechanisms underlying dysregulated genes and networks in ASD and other psychiatric disorders, a fundamental goal of psychENCODE. **The primary goal of this application is to extend our ongoing analyses of healthy and ASD brains under the psychENCODE consortium with the inclusion of additional genomic features, brain regions, developmental time points and cell-type specific analyses.** By performing these analyses we will enhance this public resource and improve our understanding of the molecular processes underlying normal human neurodevelopment and ASD.

Our group has been collaborating closely for a decade[11-15], bringing together expertise in developmental neurobiology, human tissue biobanking, genetics and genomics, statistics, bioinformatics and systems biology. Several key conceptual threads have been apparent in our work together: 1) Revealed new insights into human neurodevelopment through functional genomic profiling of postmortem tissue and cell culture models[12,16]; 2) Assessed rare and *de novo* mutations for ASD association[13,17,18], based on the notion that down-stream analyses are only as good as the genes that go into them; 3) Identified the neural processes and pathways that are altered in the presence of ASD-associated mutations, as well as when and where these processes and pathways occur in the developing human brain[15,17,19]. Here, we propose to continue this highly productive collaboration and expand psychENCODE phase 1 efforts through three integrated aims.

**Aim 1. Time, region and cell type-specific molecular profiling of control and ASD brains.** In *subaim 1.1*, we will profile the transcriptome (by RNA-seq), *cis*-regulatory elements (ChIP-seq) and 3D chromatin architecture (Hi-C) in neurotypical dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR) during mid-fetal development, infancy, childhood, adolescence and adulthood. To address cellular heterogeneity and to complement the psychENCODE phase 1 tissue level data analyses, we will obtain these data from neuronal and non-neuronal nuclei collected with fluorescence-activated nuclei sorting (FANS). In *subaim 1.2.*, complementary genomic analyses will be done on the FANS nuclei from syndromic and idiopathic ASD brains and matched control brains**,** to identify transcripts, regulatory elements, and 3D chromatin structures altered in ASD in brain region and cell type-specific manners.

**Aim 2. Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains.** In *subaim 2.1*, each dataset generated in Aim 1 will be analyzed to identify differences between the developmental stages and two major cell types in healthy and ASD tissue. Furthermore, these datasets will be integrated to gain comprehensive insights into the underlying mechanisms; Hi-C defined physical intrachromosomal interactions will be intersected with ChIP-seq to identify functional interactions between regulatory sequences potentially associated with transcriptional changes. In *subaim 2.2*, we will harmonize and integrate our multi-omic datasets with other psychENCODE studies and large-scale genomic datasets, such as BrainSpan, CommonMind, ENCODE, GTEx and REMC.

**Aim 3. Spatiotemporal analysis in ASD.** Our prior work assessed the enrichment of ASD genes in spatiotemporal co-expression networks to identify the frontal cortex during mid-fetal development as a critical window in ASD etiology. In *subaim 3.1*, we will use the neurotypical gene expression data and our expanded list of ASD associated genes to increase the resolution of this spatiotemporal analysis. In *subaim 3.2* we will use whole-genome data for 5,120 individuals in ASD families to identify non-coding *de novo* mutations within the regulatory loci identified in neurotypical brains in Aims 1 and 2. In *subaim 3.3* we will use these non-coding mutations and the regulatory networks from Aim 2 to perform an independent assessment of spatiotemporal convergence in ASD to complement our gene-based analysis in subaim 3.1. Finally, in *subaim 3.4* we will use the regulatory networks that are specific to the ASD brain identified in Aim 2 to assess enrichment of ASD-associated genes and non-coding mutations thus demonstrating that such networks are causally linked to ASD rather than simply a consequence of ASD. At the completion of this aim we will have three independent assessments of spatiotemporal convergence in ASD from ASD-associated genes, ASD-associated regulatory loci, and ASD-associated networks in the *post mortem* brain.

# RESEARCH STRATEGY

## SIGNIFICANCE

Neuropsychiatric disorders such as autism spectrum disorder (ASD), bipolar disorder (BD), and schizophrenia (SCZ) are complex and devastating illnesses with considerable morbidity and mortality, as well as high personal and societal costs. Many of them are also polygenic, with multiple variants, both rare and common, spread throughout the genome influencing the disease risk[3]. Recent studies have identified rare variants contributing to psychiatric disorders that are enriched in genes involved in global gene regulation and chromatin modification, and many common risk variants are enriched in regulatory regions of the human genome, regions whose functions are poorly understood. The interpretations of these variations in regulatory regions will certainly be improved with better maps of RNA transcripts, regulatory elements, and chromatin states in the human brain. The age of onset and progression of major psychiatric disorders also varies (**Figure 1**) necessitating the study of the temporal dynamics of gene regulation during human brain development and recognizing the developmental context of psychiatric disorders. An emerging body of research indicates that many aspects of the development and physiology of the human brain are not well recapitulated in model organisms[20-24] **and therefore it is increasingly apparent that psychiatric disorders need to be understood in the broader context of human brain development and physiology.**

In recent years, considerable effort has been made by many studies, including large-scale efforts by ENCODE, NIH Roadmap (REMC) and GTEx projects to survey the diversity of *cis*-acting regulatory regions and RNA species of the human genome across different tissues and time points. However, a comprehensive catalog of transcripts, regulatory elements, epigenetic modifications, and chromatin structure from the human brain during development and in distinct brain regions and cell types is lacking. The PsychENCODE (phase 1) projects have initiated these efforts.

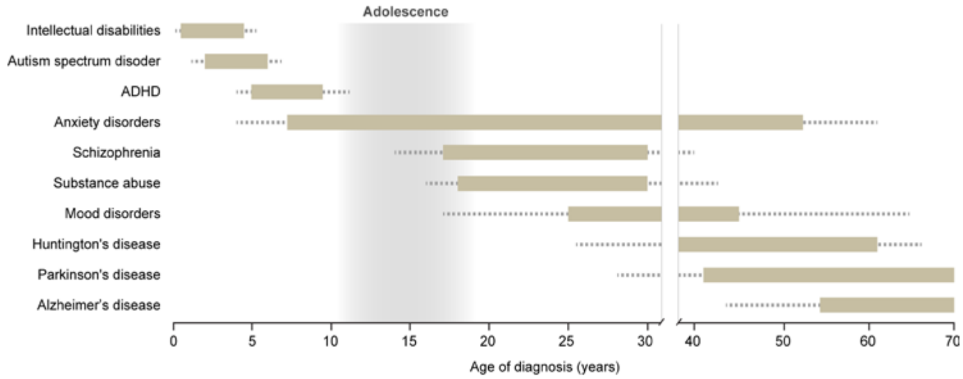


**Figure 1. Psychiatric and neurological disorders have discrete ages of onset.** *The bars indicate the age range that each disorder commonly affects, with less frequent ages of diagnosis denoted as dotted lines. This variability is indicative of dysregulation of tightly controlled developmental processes and highlights the necessity of defining the spatio-temporal molecular processes in human brain.*

**PsychENCODE consortium projects.** The key goals of the PsychENCODE project are to provide an enhanced framework of regulatory elements, catalog epigenetic modifications, and quantitate coding and non-coding RNA and protein expression in a tissue- and cell-type specific manner from neurotypical (healthy) control brains and diseased post-mortem human brains[25]. These efforts will be complemented with integrative analyses, as well as with functional characterizations of disease-associated genomic elements using human neural cell systems or the developing mouse brain. However, the human brain is heterogeneous cellularly and its development is regionally asynchronous and prolonged. *To overcome issues that hamper the potential benefits of initial psychENCODE studies, we will apply several approaches to address regional and cellular heterogeneity, prolonged development, and new genomic methods in the context of brain development and ASD.*

Here we focus on **neurotypical (control) brain and ASD**, which is a complex developmental syndrome with a significant genetic contribution. Although considerable genetic and phenotypic heterogeneity has complicated efforts to establish the biological substrates of the syndrome, the emergence of reliable genetic findings has started to shed light on potential pathogenic mechanisms, providing an extraordinary opportunity for developing a mechanistic understanding of the disorder. Recent studies suggest that **over 500 rare, *de novo* mutations contribute to ASD risk and no single genetic mutation accounts for more than 1% of ASD cases**[13,17,26-30], consistent with significant heterogeneity in this, and other neuropsychiatric disorders[3]. Despite this heterogeneity, mapping **ASD risk genes onto co-expression networks that represent normal human brain development has revealed that ASD genes coalesce in modules related to chromatin remodeling and transcriptional regulation during early fetal brain development, suggesting potential convergent pathways in the disorder**[9,15,27,31]. Another remarkable finding that parallels the convergence of genetic findings in developmental pathways is the identification and validation of shared transcriptional changes in postmortem brain in ASD[9]. This transcriptional dysregulation, coupled with the evidence that large effect size *de novo* ASD

risk genes are highly enriched in chromatin modifying genes (many of which are expressed in early fetal brain development), emphasizes the importance of understanding the nature and extent of chromatin disorganization in ASD brain and in normal brain development. **Further, since these data suggest distinct neuronal and glial gene dysregulation, it is crucial to delineate the profiles of these major cell types.** In addition to our ongoing efforts in PsychENCODE phase I project, this proposal provides critical advances in our understanding of the role(s) of non-coding functional elements in the pathophysiology of ASD and a scaffold for understanding chromatin structure and gene regulation across normal brain development. Overall, the approach proposed here will provide mechanistic insights that connect distinct transcriptional programs associated with ASD pathogenesis, and will provide a resource of the mechanisms of gene regulation across brain development to inform other neuropsychiatric disorders, a key goal of psychENCODE. **This work also leverages psychENCODE phase 1 projects by adding significant new data to expand the value of the resource and by directly addressing key areas of interest in control and ASD brains as outlined in RFA-MH-16-230**: ***1)*** *Generation of comprehensive, high resolution human brain region/cell type and age-specific maps of different classes of RNA transcripts, regulatory elements, chromatin states, chromatin conformation, and chromatin interactions;* ***2)*** *Identification of human brain region/cell type and age-specific molecular processes;* ***3)*** *Integration of these newly generated multi-omic datasets, from diseased and healthy control brains, with large-scale genomic resources;* ***4)*** *Generation and analysis of high-depth, whole genome sequencing data to allow for improved evaluation of various genetic alterations; and* ***5)*** *Development of comprehensive molecular models of disease (i.e., ASD) using systems biology approaches.*

## INNOVATION

This proposal is innovative in several aspects. First, to the best of our knowledge, the systematic discovery and functional characterization of genomic non-coding elements and 3D chromatin architecture has not been performed in healthy developing human brains or ASD brains at a cell type-specific resolution. For example, we use Hi-C, which combines chromosome conformation capture and NextGen sequencing to identify physical interactions that capture multiple levels of chromosome architecture ranging from nuclear configuration ("compartments" of about 5Mb) to TADs (domains of 500kb on average) and gene loops (often reflect enhancer promotor relationships; 40kb average), and is the only such method that spans all of these levels, genome-wide[32-34]. Second, this project will conduct direct analysis of one of the largest collection of well-characterized high quality healthy as well as syndromic and idiopathic ASD postmortem brains. Third, we will combine fluorescence-activated nuclei sorting (FANS) with advanced genomic techniques to analyze multiple genomic features in archived development control and ASD brains. Fourth, we will leverage these analyses with our ongoing psychENCODE phase 1 tissue level analyses and other recent large-scale genomic resources, such as BrainSpan, ENCODE, GTEx and Roadmap project. Therefore, our proposed data and integrated analyses has potential to improve our understanding of genomic processes and normal human brain development as well as diagnostics, neurobiology and treatment of ASD.

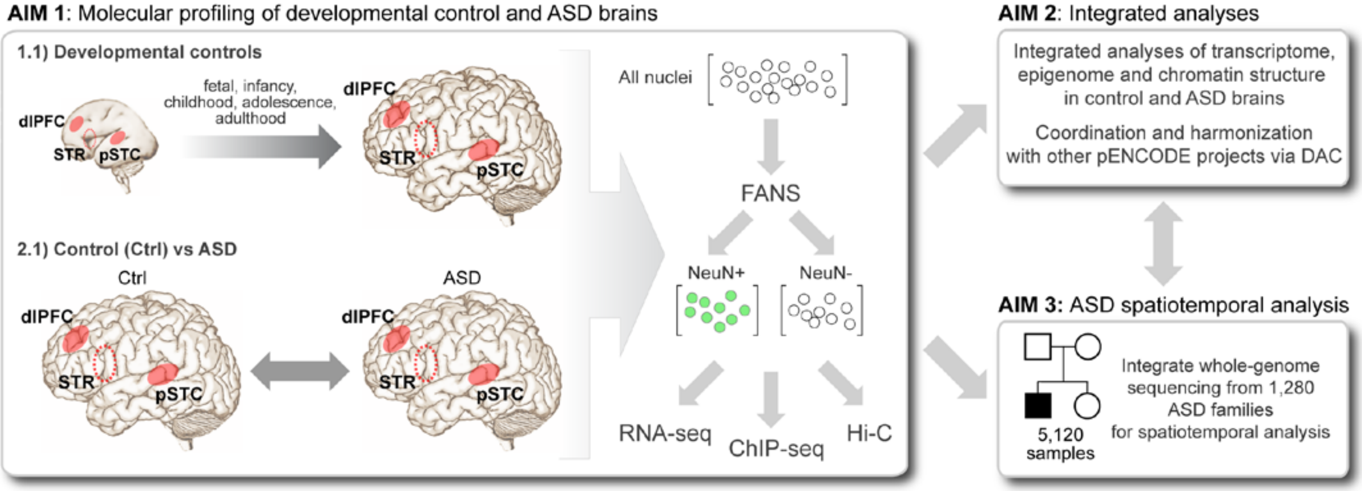
## COLLABORATION

This collaboration brings together multiple groups with long standing expertise in developmental neurobiology, psychiatry, human biobanking, genetics and genomics, statistics, bioinformatics, and systems biology that have worked closely with one another for almost a decade as evidenced by many co-publications. Several key conceptual threads have been apparent in our work together related to human brain development and neuropsychiatric disorders: **1**) Revealed new insights into human neurodevelopment through functional genomic profiling of postmortem tissue and cell culture models[12-16]; **2**) Assessed rare and *de novo* mutations for ASD association[13,17,18]; **3**) Identified the neural processes and pathways that are altered in the presence of ASD-associated mutations, as well as when and where these processes and pathways occur in the developing human brain[15,17,19]. In addition, M. Gerstein (Yale) and Z. Weng (University of Massachusetts), experts in bioinformatics and computational biology, are leaders of the PsychENCODE DAC, which will normalize the data to remove batch effects, establish uniform data processing pipelines and build calibration resources for all assays to enable comparison and integration of the data generated by all psychENCODE groups. The efforts of each group will be tightly integrated in order to communicate progress and results, design and implement analytical tools, and transfer data. Given the complexity of human neurodevelopment and genetics/neurobiology of ASD, we believe that integrating the respective expertise of these groups, and their respective collaborators at UCLA (Ernst and Geschwind), UCSF (Sanders, State and Willsey), UMass (Weng), and Yale (Gerstein and Sestan), offers the best opportunity to better understand human brain development and ASD through functional genomics. Here, we propose to leverage our expertise and continue this highly productive collaboration and expand psychENCODE phase 1.

**ELEMENTS UNIQUE TO THIS SITE** (UCLA; Geschwind, PI; Ernst, co-investigator)
The UCLA team will perform chromosome capture and deep sequencing (Aim 1) and analysis (Aim 2) of 3D chromatin architecture (Hi-C) in the neurotypical and ASD dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR) using nuclei sorted by the Yale PIs (Sestan). The UCLA group will also perform WGCNA, hQTL, and other integrative analyses of genomic data in close collaboration with Yale investigators (Aim 2). UCLA will also contribute brain tissue from ASD patients and controls to the resources at Yale for the studies (Resources). UCLA continues to generate ASD and control brain transcriptional profiles as part of psychENCODE 1, completed by the Spring of 2016, and Geschwind will contribute these data to the integrative analyses proposed, as well as future cross disorder analyses. Fetal brain is not in psychENCODE 1, but realizing its value, Geschwind has produced RNAseq and ATAC-seq data from 6 subjects that will be used for Hi-C here, and will also contribute it to psychENCODE by April 2016.

## APPROACH

The objective of this proposal is to extend our ongoing tissue level analyses of healthy and ASD brains under the psychENCODE consortium with the inclusion of additional genomic methods, brain regions, developmental time points, and cell-type specific analyses. By performing three integrated aims (**Figure 2**) we propose to enhance this public resource and improve our understanding of the molecular processes underlying normal human neurodevelopment and ASD.



AIM 1: Molecular profiling of developmental control and ASD brains
1.1) Developmental controls
dlPFC
STR
pSTC
fetal, infancy, childhood, adolescence, adulthood
dlPFC
STR
pSTC
2.1) Control (Ctrl) vs ASD
Ctrl
ASD
dlPFC
STR
pSTC
dlPFC
STR
pSTC
All nuclei
FANS
NeuN+
NeuN-
RNA-seq
ChIP-seq
Hi-C
AIM 2: Integrated analyses
Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains
Coordination and harmonization with other pENCODE projects via DAC
AIM 3: ASD spatiotemporal analysis
5,120 samples
Integrate whole-genome sequencing from 1,280 ASD families for spatiotemporal analysis

**Figure 2. Schematic workflow of three specific aims.**

### Aim 1. Time, region and cell type-specific molecular profiling of control and ASD brains.

***Rationale and preliminary supporting data:*** Three major observations provide motivation for this aim. The *first* is the recognition that genomic data, including transcriptomic, epigenetic and physical chromatin structure, from the relevant neurotypical tissue (control), spanning the key epochs of neurodevelopment and function from fetal to adult periods, provide a new and previously unobtainable view of genetic risk for psychiatric disease[10,15,16,31,35,36]. The *second* is that brain is comprised of an extremely heterogeneous mixture of cell types that exhibit distinct molecular profiles, including glia-to-neuron ratios that could show considerable fluctuations across normal development or in certain disease states. The *third* is the observations of differences in transcriptome organization via tissue-level gene co-expression network analysis conducted between ASD and normal brains[9]. Thus, here **we propose to create a region and cell type-specific normal developmental scaffolding on which to frame disease variants via transcriptional (RNA-seq), epigenetic (ChIP-seq) and chromatin architecture (HiC) profiling of neuronal and non-neuronal cells at key epochs in human brain development (subaim 1.1), as well as compare these profiles in ASD and matched control brains (subaim 1.2) to help elucidate the mechanisms by which genetic variation alters brain development and function, leading to ASD and related neuropsychiatric conditions.** While several genomic features are currently being analyzed in control and ASD brains by our and other groups in the psychENCODE consortium, cellular heterogeneity during development, other genomic features (e.g. 3D chromatin contacts), have yet to be addressed. To address these issues, we will utilize our large, high quality, phenotypically well-characterized human brain collection (see Facilities and Resources section), as well as newly implemented methods to collect molecularly defined cell type specific nuclei from archival human postmortem brains in this collection.

**Our preliminary data demonstrates a clear pattern of transcriptional dysregulation is observed in 2/3 of ASD brains[9], which we have now confirmed in our psychENCODE phase 1 projects (in a more than**

**double sized sample of cases and controls)** using tissue level RNA-seq and ChIP-seq (H3K4me3 and H3K27ac) in multiple brain regions in 43 idiopathic ASD cases, 8 cases with chromosome 15q11-13 duplication syndrome (dup15q) and ASD, and 63 controls[37,38]. We also observe that post mortem brain from patients with ASD caused by (dup)15q11-13 share this same pattern at all levels of differential protein coding gene expression, splicing and lncRNA[38]. As the first step in exploring potential mechanisms, we performed epigenetic profiling of ASD vs. control brains with H3K27ac marks, which indicate active enhancers[39]. Genes with differential H3K27ac peaks in their promoter regions (5000bp upstream of the transcription start site) were enriched with neuronal genes with changes in expression. This result demonstrates that transcriptional dysregulation in ASD is partially mediated by changes in histone/chromatin modifications. Furthermore, the two major groupings of modules derived from whole tissue gene expression analysis sort into those up-regulated and expressed in glia, and those down-regulated in neurons[9,38], strongly motivating our plan for transcriptional, epigenetic and Hi-C profiling in neurons and non-neuronal (glial) cells independently.

**Another of the important advances in methodology that will be implemented here is the assessment of 3-D chromatin structure across to different brain regions and cell types, and 5 key epochs of normal brain development and in ASD brains.** Our preliminary data strongly supports the value of these data and our ability to perform and analyze these experiments (see also[40]). We established an efficient Hi-C protocol and obtained high resolution data (10 kb resolution; via deep sequencing) from the fetal cortex from 3 individuals dissected into two zones: cortical plate (CP) and germinal zones (GZ) at post-conception week (PCW) 18 (total n = 12 samples: representative heatmap shown in **Figure 3A**). Demonstrating the data quality, principal component of the interchromosomal interaction matrix for GZ shows a high correlation with GC content ($r$=0.791, $P$<$10^{-256}$), gene number ($r$=0.384, $P$<$10^{-256}$), DNase I hypersensitivity ($r$=0.416, $P$<$10^{-256}$), and to a lesser extent, gene expression ($r$=0.129, $P$=1.11x$10^{-66}$; **Figure 3B and C**), recapitulating previous work in cell lines[41]. We next asked how chromatin interactions elicit transcriptional co-regulation. We hypothesized that highly interacting chromatin regions would be co-regulated at least in part by sharing chromatin remodelers and transcription factors (TFs). To test this, we binned chromatin interactions into top and bottom percentiles, and compared the distribution of correlation patterns for genes in the high and low interacting regions of chromatin. We observed that the high interacting regions were significantly biased toward positive correlations (**Figure 4A**), supporting the hypothesis that co-localization can predict co-expression.

We next integrated these data with the epigenomics map from the NIH Roadmap project[42]. By comparing the epigenetic mark combination matrix with the Hi-C contact matrix, we demonstrate that interacting regions exhibit shared epigenetic patterns: loci associated with transcriptional regulation and enhancers are significantly more likely to interact with each other (**Figure 4B**). Comparison of TF binding site (TFBS) combination matrix (generated from TFBS map reported in[43]) with the intrachromosomal contact matrix revealed distinct combinatorial patterns of TF binding likely to mediate chromosome interactions (**Figure 4C**), thus revealing new experimentally testable regulatory relationships.

To validate that Hi-C data can identify target genes regulated by single nucleotide polymorphisms (SNPs) in a general setting, we determined if SNPs with a significant effect on gene expression were also identified as

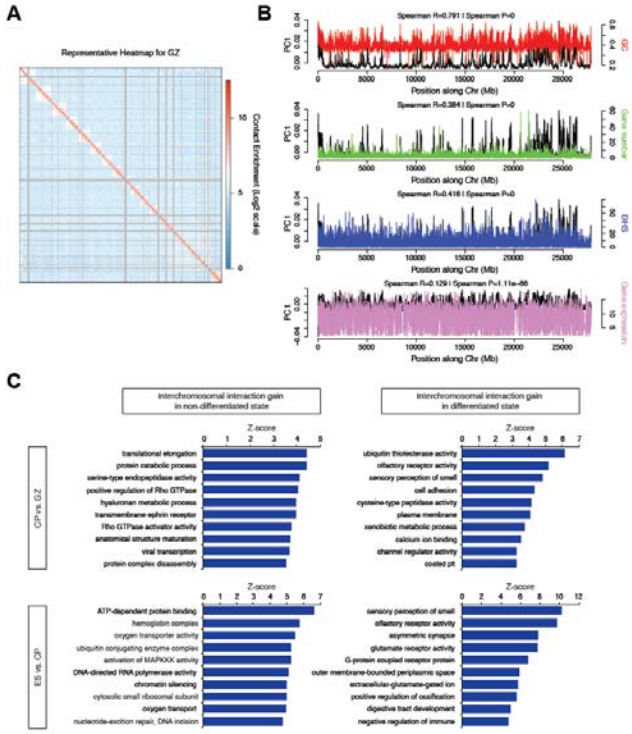


**Figure 3. Chromosome conformation in fetal brains (by Hi-C).** ***A.*** *Representative heatmap of chromosome contact matrix of GZ. Normalized contact frequency (Contact enrichment) is color-coded according to the legend on the right.* ***B.*** *Spearman correlation of PC1 of chromatin interaction profile of fetal brain (GZ) with GC content (GC), gene number, DNase I hypersensitivity (DHS), and gene expression level of fetal brains. These data show relationship of 3D structure to key known functional elements as has been previously shown in other systems.* ***C.*** *Gene ontology (GO) enrichment (GO Elite) of genes located in the top 5% of highly interacting inter-chromosomal regions specific to GZ vs. CP (top), and ES vs. CP (bottom), indicating that genes located on dynamic chromosomal regions are enriched for neuronal function in CP, which contains the more differentiated laminae. Please see Won et al. 2015 in Appendix for higher magnification figure.*

interacting by Hi-C using *cis*-expression quantitative trait loci (eQTL) data from adult frontal cortex[44]. Indeed, Hi-$C_{eQTL}$ genes were significantly over-represented with known associated genes from the eQTL study and eQTL SNP-transcript pairs exhibit significantly higher chromatin contact frequency than the null across all distance ranges measured, further supporting the utility of Hi-C to infer the gene or region of activity for regulatory variation. In addition we asked whether significant physical cis-chromosomal contacts identified with Hi-C could inform functional annotation of 108 genome-wide significant schizophrenia loci, most of which lie far outside known coding or other functional regions of the genome.

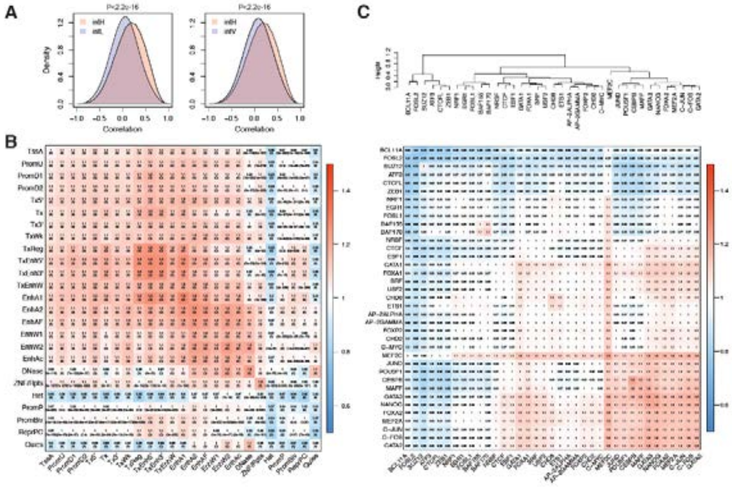


**Figure 4. Highly interacting regions share co-expression patterns, which is mediated by epigenetic regulation**. ***A.*** *The top 10,000 highest interacting regions (intH) in fetal brains both at GZ and CP show positive correlation in their gene expression patterns, while the top 10,000 lowest interacting regions (intL) and top 10,000 highly variant regions (intV) have no skew in the distribution, consistent with random interactions. P-value, Kolmogorov-Smirnov test.* ***B-C.*** *Epigenetic state combination (B) and TFBS combination (C) for intrachromosomal interacting regions. The epigenetic state matrix and TFBS combination matrix were generated by marking loci where two interacting chromosomal bins share epigenetic signature. For example, the epigenetic combination matrix between the active transcription start site (TssA) and active enhancers (EnhA1) is generated by marking where interacting loci have TssA and EnhA1. Intrachromosomal contact frequency map is compared to the epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Odds ratio (OR) and P-values are depicted in the heatmaps (Please see Won et al. 2015 in Appendix for higher magnification figure).*

Although SNPs are typically assigned to the closest genes, or those within the LD block, Hi-C indicated that about 50% of the variants were neither adjacent to the index SNPs (most-associated SNP within a locus), nor in LD. Interestingly, Hi-$C_{SCZ}$ genes significantly overlap with ASD *de novo* likely gene-disrupting (LGD) targets[26,45] (CP: OR=2.4, P=1.6x$10^{-5}$, GZ: OR=1.8, P=0.006), indicating a shared genetic etiology between ASD and schizophrenia[46]. The fact that genes with LGD mutations in ASD are associated with regulatory variants in schizophrenia suggests that complete abrogation of these genes may cause developmental defects as in ASD, while regulatory changes in these genes may cause later-onset of neuropsychiatric symptoms as in schizophrenia. *Collectively, these preliminary data demonstrate that we can conduct and analyze genome-wide Hi-C experiments, integrate these data with other epigenetic and transcriptomic data, and use chromatin architecture elucidated by Hi-C to provide novel genome-wide insights into the regulatory mechanisms occurring during neuronal differentiation and disease pathogenesis.*

***Experimental design and methods:*** In ***subaim 1.1.,*** *we will profile the transcriptome (by RNA-seq), cis-regulatory elements (ChIP-seq) and 3D chromatin architecture (Hi-C)* in the control neurotypical dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR). These regions have been implicated in the risk for ASD and schizophrenia[35] and in the cases of dlPFC and pSTC shown to have dysregulated transcriptional patterns in ASD[9]. Recent studies have also highlighted the late mid-fetal frontal cortex as most enriched for co-expression of ASD and schizophrenia *de novo* hits[15,31,35]. Brains from at least 5 key epochs of development representing mid-fetal, infancy, childhood, adolescence and adult brain, and a minimum of 6 subjects (balancing sex when possible) from each of these 5 epochs (30 brains in total) will be profiled.

Cell-type specific chromatin, epigenetic and transcriptome assays are at the core of this project. Mario Skarica, a talented research associate scientist in the Sestan lab, has developed a protocol to isolate high quality nuclei with preserved chromatin and RNA from archival fresh frozen fetal and postnatal human brains. Using this approach he has obtained, on average, 2.57 +-0.8 and 6.93+-3.3 million intact nuclei from 100 mg of the fetal or adult prefrontal gray matter (i.e., fetal CP or adult cortical layers 1 to 6 with a small part of underlying white matter), respectively (**Figure 5A**). Furthermore, we separated neuronal and non-neuronal nuclei, by immunostaining with the NeuN antisera against pan-neuronal splicing protein RBFOX3 (**Figure 5B and C**) and sorting on BD FACSAria IIU Three-Laser System. Starting with infancy and onwards, postnatal gray matter tissue corresponding to six-layered postnatal cortex and small part of adjacent white matter from dlPFC and pSTC, or STR (corresponding to the caudate-putamen with the internal capsule at the septal level) will be processed.

Tissue samples will be dissected directly from frozen tissue blocks using custom dental tools and protocol described in Kang et al., 2011[47]. These dissections will be performed by Nenad Sestan, who has over 2 decades of experience in human neuroanatomy and tissue processing and has microdissected over 1600 tissue samples for exon array profiling of the human brain transcriptome[47]. Given the high proportion of neurons in the cortical plate of the mid-fetal brain (approaching 95% or more), and relatively few neurons that are positive for NeuN at 17-20 PCW in neocortical CP or STR[48], we will not sort NeuN+ and NeuN- nuclei from mid-fetal brains, but instead analyze tissue homogenate and unsorted nuclei from CP of prospective dlPFC and pSTC as well as STR, separately, from corresponding neocortical and striatal GZ (i.e., VZ and SVZ) containing a mixed population of dividing neural stem/progenitor cells with a minor contribution of newborn neurons and glia.

Tissue samples will be pulverized and processed to release nuclei, which will be purified by ultracentrifugation and processed for RNA-seq in the case of mid-fetal samples or in the case of all postnatal specimens (infancy and onwards) sorted into a NeuN+ (predominantly neurons) and NeuN- (mostly glia) fractions. In the past year, we have obtained on average 23.45+-7.2 percentage of NeuN+ nuclei from PFC (**Figure 5C**). This approach will provide unbiased quantitative assessments of cell types in healthy and ASD brains. This approach allows us to simultaneously collect molecularly defined cell type-specific nuclei and isolate DNA, chromatin, and nuclear RNAs. Bulk tissue level RNA-seq is available for dlPFC, pSTC and STR in control and ASD brains as part of psychENCODE phase 1 studies[38], has already been added to enhance the scope of the resource. All brains necessary for this project are currently available in the Geschwind and Sestan labs (see Facilities and Resources section for the list).

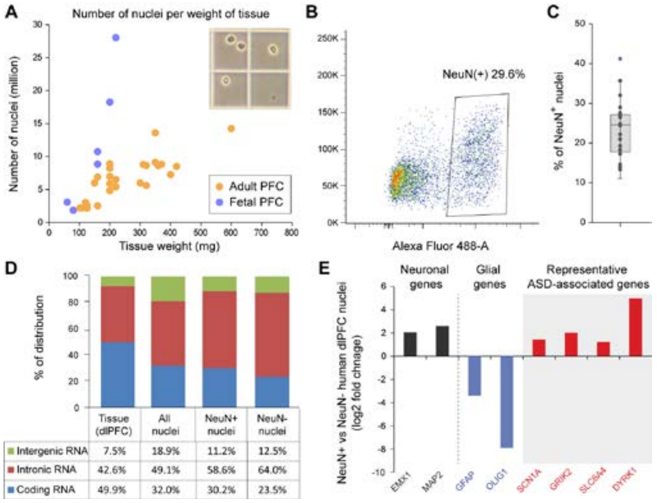


**Figure 5. Fluorescence-activated nuclei sorting (FANS) and nuclear RNA-seq of human dlPFC.** ***A.*** *Collection of single nuclei (see insert) from fetal (n=6) and adult (n=29) PFC.* ***B.*** *FANS plot for NeuN immunopositive nuclei.* ***C.*** *Percentage of NeuN+ nuclei collected across different experiments.* ***D.*** *Coverage for exon, intron and intergenic regions of different sequencing technologies.* ***E)*** *Differential expression comparison between NeuN+/- FANS nuclei for neuronal, glial and ASD-related genes.*

Total RNA will be extracted from 1 million nuclei using Norgen's Cytoplasmic & Nuclear RNA Purification Kit. RNA from tissue and cell populations will be depleted of rRNA and sequencing libraries prepared with TruSeq Stranded Total RNA with Ribo-Zero Gold and SMARTer Stranded RNA-Seq Kit, respectively. As expected, our preliminary nuclear RNA-seq analyses revealed higher percentage of unspliced primary transcripts and extensive identification of nuclear-retained long non-coding RNAs (**Figure 5D**). Importantly, we detected robust cell type-specific expression differences, including those of ASD-associated genes (**Figure 5 E)**. RNA-seq libraries will be sequenced on the Illumina HiSeq 2500 at the Yale Center for Genome Analysis (http://ycga.yale.edu/) to generate 100 bp strand specific paired-end sequence at over 40 million reads per end for each sample. For ChIP-seq, 1 million nuclei will be processed through our established protocol using well-characterized ChIP-grade H3K27ac and H3K4me3 antibodies that have been used in psychENCODE phase 1 tissue-level experiments. ChIP-seq libraries will be sequence at HiSeq 2500 at Yale at >40 million reads per sample. Using the standard pipelines developed in the Sestan and collaborating labs, we will perform QC analyses and compare the transcriptome and epigenetic data from different time points and regions to construct spatiotemporal gene and disease state profiles and co-expression networks using computational methods described in Aim 2.

For Hi-C, 2 million nuclei will be prepared from each sample and cross-linked in 1% formaldehyde for 10 min. Cross-linked DNA will then be restriction digested using HindIII, digested chromatin ends filled with biotin-14-dCTP, and resulting blunt-end fragments ligated under dilute conditions to minimize random intermolecular ligations. Following this, crosslinking will be reversed, unligated ends removed by exonuclease digestion (T4), DNA sheared by sonication, and 300-600bp fragments selected. The intermolecular ligation products containing biotin-tagged DNA will be pulled down with streptavidin beads and ligated with Illumina paired end adapters and the library sequenced by Illumina 50bp paired-end sequencing over 3 lanes of the HiSeq 25000 at UCLA, a depth necessary to facilitate sufficient hi-resolution analysis (300-500 million mapped reads), which can also be augmented by pooling samples to increase depth as needed.

In ***subaim 1.2**, complementary genomic analyses will be done on the FANS nuclei from control, and syndromic and idiopathic ASD brains*, to identify transcripts, regulatory elements, and 3D chromatin structures altered in ASD in brain region and cell type-specific manners. We will conduct RNA-seq, ChIP-seq and Hi-C on sorted neuronal and non-neuronal nuclei from 2 cortical regions, dlPFC and pSTC, and STR from 20 matched control and 20 ASD individuals, including 5 dup15q cases. We will select 10 ASD cases manifesting the shared pattern of transcriptional dysregulation observed, 10 without this pattern, and match them to controls to account for potential confounders (sex, age, postmortem interval [PMI], and RNA integrity numbers [RIN]). We will select 5 dup15q brains with most similar breakpoint structures. Hi-C will be performed on sorted nuclei using the identical experimental methods as in subaim 1.1.

***Pitfalls and alternatives:*** The techniques in these proposed experiments are commonly used in our laboratories and we do not expect complications. One potential issue is the obtainment of adequate samples. The Sestan lab has almost 200 high quality frozen human prenatal, early postnatal and adult brain specimens from clinically unremarkable (neurotypical) control donors. Control brains from this collection were used for different BrainSpan and psychENCODE phase 1 projects (see example studies[12,47,49,50] and Resources and Facilities section). Both Geschwind and Sestan labs have tissue samples from over 50 post mortem ASD cases and matched controls with good quality RNA, and have participated in the new initiative at the Simons Foundation to collect additional postmortem ASD brains. A related concern is whether the 20 ASD brains we propose to analyze are sufficient, given the heterogeneity typical of ASD, to detect robust differences between these samples and our controls. However, we were able to detect transcriptional dysregulation in 2/3rds of ASD brains in a smaller cohort17, and by directly comparing ASD brains exhibiting hallmarks of dysregulated transcription with those that do not, we expect to have sufficient statistical power to assess the extent to which 3D chromatin structure contributes to the observed transcriptional changes. Further, the use of 5 dup15q cases provides an additional homogeneous cohort, and as our preliminary results on transcriptome analysis of this cohort demonstrate (appendix), such sample sizes are sufficient. The main pitfall of Hi-C is that it averages chromosome contact population from millions of nuclei. Single-cell Hi-C can complement this limitation[51], but it can capture only one interaction for a given locus. Homogenous population of cells can be achieved by FANS and thus we propose this approach here. Additionally, Hi-C offers other benefits, including the ability to analyze interactions mediated by multiple TFs *en masse* in Hi-C, that are not easily achievable with other methods such as ChIA-PET. While our FANS approach, which follows standards accepted across the psychENCODE projects, is limited to two major groups of cells, we have been implementing the use of other cell type specific nuclear antibodies and single nuclear RNA-seq. Finally, we realize that other regions, including the thalamus, hypothalamus, and hippocampus, may be affected in ASD. We believe our work on the neocortex and STR will develop a framework for understanding of the molecular neuropathology of ASD which can then be extended to include other regions in the future.

**Aim 2. Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains.**

***Rationale:*** We will analyze the data generated in the previous aim to (1) identify *developmentally regulated and cell type specific changes* in the transcriptome, epigenome and the 3D chromatin structure (2) integrate the three types of datasets to gain comprehensive insights into the underlying mechanisms of transcriptional regulation and dysregulation in development and disease, respectively.

***Experimental design and methods*:** In ***subaim 2.1**, several first order analyses will be done for quality control and to provide the data as a processed resource in addition to the raw data*. We will use Illumina CASAVA to purify the low-quality and non-identified reads and Fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), to report fundamental quality parameters. Next, Tophat[52] will be employed to uniquely align the filtered reads to their reference genome and RSEQtools[53] to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. The R package DESeq (http://bioconductor.org/packages/release/bioc/html/DESeq.html) will be used to identify differentially expressed (DEX) genes and well established methods including MATs to identify differential splicing[10,37]. DEX genes will be detected from the reliably expressed coding and non-coding transcripts, which are defined as transcripts with RPKM ≥ 1 in at least 2 samples of different developmental period. ChIP-seq reads will be aligned to the genome by Bowtie. After filtering of low score reads, we will use the MACS platform to call peaks enriched over the input library, and peaks with high empirical FDR will be excluded from further analysis. Thus, we will catalog all potential cis-regulatory elements from our genome-wide histone modification maps in all brain regions across developmental periods.

*For Hi-C analysis*, *hiclib* (https://bitbucket.org/mirnylab/hiclib) will be used to perform all initial analysis on Hi-C data from mapping to filtering and bias correction (see also[40]). Sequenced reads will be mapped to the human

genome by *Bowtie2* (with increased stringency, *--score-min -L 0.6,0.2--very-sensitive*) through iterative mapping and read pairs allocated to HindIII restriction enzyme fragments. Self-ligated and unligated fragments, fragments from repeated regions of the genome, PCR artifacts, and genome assembly errors will be removed. Filtered reads will be binned at 10kb, 40kb, and 100kb resolution to build a genome-wide contact matrix at a given bin size. This contact map depicts contact frequency between any two genomic loci. To decompose biases from the contact matrix and yield a true contact probability map, filtered bins are subjected to iterative correction[41]. Bias correction and normalization results in a corrected heatmap of bin-level resolution. 100kb resolution bins are assessed for inter-chromosomal interactions, 40kb for TAD analysis, and 10kb for gene loop detection. For TAD-level analysis[32], we will quantify the directionality index by calculating the degree of upstream or downstream (2Mb) interaction bias of a given bin, which will be processed by a hidden Markov model (HMM) to remove hidden directionality bias. For gene loop detection, aggregate peak analysis (APA) will be performed that quantifies the aggregate enrichment of putative peak sets by calculating the sum of a series of submatrices derived from a contact matrix[34]. Resulting inter- and intra-chromosomal interaction matrices as well as genome-wide TADs and gene loops will be used for integrative analysis.

*Developmental and cell type-specific changes*: Pearson's correlations between the first principal components (PC1) from different stages and neuronal and non-neuronal cell types, as well as with our own and other published data will be calculated to compare similarities between different cell types. We will explore alternative transcriptional mechanisms or post-transcriptional modifications occurring in normal (and ASD-affected, see below) regions/cells and time points. These can include up- or down-regulating expression, altered spatiotemporal gene expression, imbalanced expression of different alleles (allele-specific expression [ASE]), aberrant splicing events, modified RNA editing sites, fusion transcripts, or loss of function due to frameshift mutations. RNA and epigenome data will also be compared with tissue level psychENCODE phase 1 and BrainSpan's RNA-seq and ChIP-seq data. We will follow up with an analysis of the relative enrichment of each cell-specific marker genes in each subpopulation and use the expression profiles of these genes to guide the identification of an expanded set of cell-type specific markers.

*Integrated analyses*: Spearman's correlations between PC1/PC2 and biological traits (gene expression, histonemark enrichment, GC content, gene density, DNase I hypersensitivity [DHS]) will be calculated. Gene expression and histone mark data generated in subaim 1.1 along with DHS of fetal brain from Epigenomic roadmap[54] will be used and average values per 100kb bin calculated. In addition to the putative cis-elements identified in the same samples, we will also use the 15 state epigenetic marks from Epigenomic Roadmap[54] in genomic regions classified based on compartments averaged across 40kb bins, as well as subject specific psychENCODE data. Epigenetic state counts[54] for one compartment category are normalized by total epigenetic mark number of that compartment category and compared between samples.

*Dysregulation in ASD brains*. Two main data analyses will be performed with the transcriptome data. We will use the same approach as in subaim 2.1 to identify DEX coding and non-coding transcripts (by DESeq) between ASD and matched controls. Gene function enrichment analysis will be performed for these DEX genes. Finally, we will also perform Weighted Gene Co-expression Network Analysis (WGCNA; http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/) to identify modules of differentially co-expressed genes in ASD cases. For ChIP-seq data, once peaks are called and filtered for quality and reproducibility, we will identify and catalog all putative enhancer and promoter sites gained or lost in ASD brains compared to matched control brains, as well as what genes they are associated with.

In Hi-C data, we will assess if compartments, TADs, and contact domain structures are abrogated in ASD brains. Interaction partners for ASD risk genes, as well as gene loops involving ASD risk gene regulatory elements will be examined. Genome-wide interchromosomal contact matrices at high resolution (approximately 10kb) will be compared between ASD and control to identify bins that exhibit the largest chromosomal interaction changes in ASD (here we refer to ASD-specific bins). Gene ontology for these genes as well as their gene expression pattern in ASD may provide novel insights on ASD mechanism. The same approach will be applied to intrachromosomal contact matrices at 10-40kb resolution. TADs in ASD vs. controls will be also compared. The directionality index around ASD-specific TAD boundaries will be calculated to check significance. Moreover, we will examine gene expression level and histone marks on TAD boundaries as well as histone marks on TADs that contain ASD risk genes. Both inter- and intra-chromosomal interaction patterns of the bins that contain differentially expressed genes in ASD or ASD risk genes will be examined in ASD vs. controls. Gene expression pattern and histone states of genomic loci that highly interact with dup15q region will be assessed. *This approach of integrating chromosome interactomes to transcriptomic and epigenetic profiles may delineate epigenetic mechanism behind gene dysregulation in ASD.*

We will also perform integrative network analyses of these multi-level genomic data with genetic variation to understand the causal mechanism of transcriptional alterations in ASD (see also Aim 3). This will include integration of DNA sequence, methylation, chromatin contacts, eQTL and hQTL by this collaborative team of investigators (e.g. to include new hQTL methods by S. Prabhakar and colleagues[55]. Gene loops detected in control and ASD will be also interrogated. Gene loops that are specific to ASD or specific to controls may directly point out aberrant enhancer-promoter interactions, TF binding, or compartmentalization of genome. We will check if ASD-specific gene loops contain any ASD-associated variants (mostly common SNV at this point, although as more whole genome sequencing (WGS) data is available over the next 12 months, we can use these data to annotate potential functions of noncoding variants (Aim 3).

In ***subaim 2.2.**, we will integrate and harmonize data across psychENCODE projects and other relevant genomic resources*. In this aim, the DAC will integrate and harmonize our datasets with other psychENCODE studies and large-scale genomic datasets, such as BrainSpan, CommonMind, ENCODE, GTEx and REMC. The PsychENCODE DAC is led by Mark Gerstein and Nenad Sestan (Yale), Zhiping Weng (University of Massachusetts), who are part of this proposal and Kevin White (University of Chicago). DAC will summarize the major analysis results produced from psychENCODE and organize them into an encyclopedia of regulatory elements in the developing and adult human brain. We are currently building such an encyclopedia for the ENCODE consortium, and we will be able to leverage the methods that we are building for ENCODE and modify them to best serve psychENCODE data. The psychENCODE encyclopedia will include several components. The first component is the raw experimental data, including the expressed transcripts in neuronal and glial cells in various brain regions, the peaks (enriched regions) of an array of histone marks, the open chromatin regions detected using ATAC-seq, the differentially enriched histone mark peaks and open chromatin regions in ASD, BD and SCZ (diseases covered by psychENCODE projects. This component will largely result from a series of uniform processing pipelines, which we will build for analyzing psychENCODE data. The second component will include results that require the integration across multiple data types, including the enhancers in each cell type, the chromatin states called using a combination of histone marks and ATAC-seq data, and the topologically associated domains and compartments called by combining histone marks, ATAC-seq and Hi-C data. The third component of the encyclopedia will provide a higher-order organization to the elements in the first two components. Specifically we will derive the target genes for enhancers in a cell type specific manner, and identify the enhancer-gene links that are disrupted in the three diseases. We will also identify the variations that are linked with difference in gene expression (eQTLs) that are within enhancers that target the corresponding genes. Finally, we will develop a portal to guide the user through the components of the psychENCODE encyclopedia, with multiple entry points, such as genes, GWAS SNPs, or a specific regulatory region in the genome.
***Pitfalls and alternatives:*** Proposed computational approaches are well established in our team and we already have a considerable expertise and collaborative history therefore we foresee no complications in performing this aim. Furthermore, Sestan, State and Geschwind have been part of the BrainSpan project and Ernst, Gerstein and Weng has been part of several other relevant genomic consortia, such ENCODE.

**Aim 3. Spatiotemporal analysis in ASD.**
***Rationale and preliminary supporting data:*** Over the past few years genomic analyses by our labs and others have made rapid progress in identifying genes associated with ASD, in particular through the identification of *de novo* mutations in ASD cases[13,17,26,30,45]. Despite the identification of these ASD-associated genes, progressing to an understanding of ASD neurobiology remains a challenge. Aims 1 and 2 described one approach to discovering this neurobiology through the identification of ASD-specific networks in *post mortem* brains. In Aim 3 we propose a complementary approach through the identification of genomic loci, brain regions, developmental stages, cell types, and neurobiological processes that are enriched for ASD mutations in genes (**subaim 3.1**) and non-coding loci (**subaims 3.2 and 3.3**) in neurotypical brains. Finally, we will test the hypothesis that ASD specific networks observed in *post mortem* brains from Aims 1 and 2 will be enriched for ASD associated mutations (**subaim 3.4**) thus demonstrating that the disruption of this network precedes the diagnosis of ASD and is therefore likely to be a cause of ASD rather than a consequence.

*1) Detection of ASD-associated genetic loci.* We identified rare and *de novo* variants in exome data from 5,563 ASD cases and 13,321 controls alongside rare and *de novo* copy number variants in microarray data from 4,687 ASD cases and 2,100 controls[17]. Comparison of these two data sets showed that small *de novo* deletions in ASD targeted the same set of genes as *de novo* loss of function point mutations in exome data. A combined analysis of exome data and small *de novo* deletions was performed using the Transmitted and *De novo* Association (TADA) method to identify ASD-associated genes. 28 ASD-associated genes were identified with very high confidence (false discovery rate (FDR) ≤ 0.01) and 65 ASD-associated genes were identified with high

confidence (FDR ≤ 0.1). These 65 genes formed a protein-protein interaction (PPI) network with two distinct subnetworks, enriched for chromatin regulatory genes and synaptic genes respectively (**Figure 6A**).

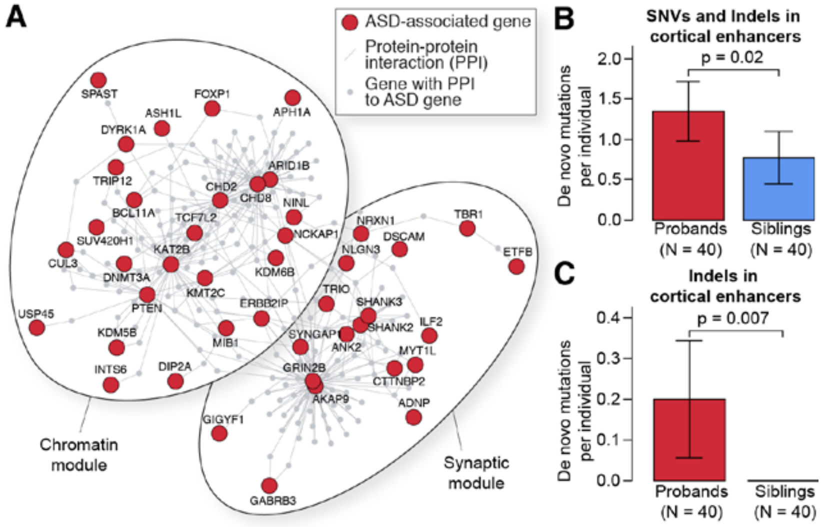


**Figure 6. ASD associated de novo mutations.** ***A.*** *65 ASD risk genes[9] (red) form a single protein-protein interaction network composed of two subnetworks. The genes in the left subnetwork are enriched for chromatin regulatory gene ontology terms. The genes in the right subnetwork are enriched for synaptic terms.* ***B.*** *De novo mutations were identified in WGS data for 40 ASD families. The median number of SNV and indel mutations per individual is shown within active enhancers that were identified by bulk tissue ChIP-Seq for H3K27ac in human dlPFC (psychENCODE phase 1 studies). P-values are calculated using linear regression with for paternal age and total de novo mutations per individual included as co-variates.* ***C.*** *The analysis was repeated for indels only.*

*__2) Detection of ASD-associated non-coding variants in whole-genome sequencing (WGS) data__*. We analyzed WGS data for 40 simplex ASD quartets composed of both parents, an affected child and an unaffected sibling control. The families were selected from the Simons Simplex Collection on the basis of no previous *de novo* loss of function or CNV mutations in exome and microarray data and high paternal age. The samples were sequenced to greater than 30x mean coverage (mean±standard 35.7±5.8). Raw data were aligned to hg19 human reference genome using BWA-mem[56]. Duplicate reads were removed with Picard (http://broadinstitute.github.io/picard/); GATK best practices[57] were used for all downstream steps including, local realignment, base quality score recalibration, SNV and indel calling, cohort-wide joint genotyping, and variant quality score recalibration. Data were normalized within families by only analyzing bases with at least 20 unique reads in all family members. A combination of PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq/) and in-house scripts were used to identify autosomal *de novo* variants based on stringent criteria designed to maximize specificity: minimum genotype likelihood (GQ) ≥20, alternate allele frequency (AB) ≤0.05 in the parents, and 0.3-0.7 in the child, minimum map quality (MQ) ≥30 in all family members, and allelic depth for the alternate allele (AD) ≥8. Approximately 7,000 *de novo* mutations were identified at a rate of 87.0±13.5 *de novo* mutations per child. Confirmation with Sanger sequencing was attempted on 10% of these variants (700) selected at random and achieved a >95% confirmation rate across both SNVs and indels, suggesting identification of *de novo* mutations with accuracy. We used tissue-level ChIP-seq for the histone modification H3K27ac from human dlPFC (psychENCODE phase 1) to identify active enhancers. We observed an increased burden of mutations in cases compared to sibling controls (p=0.02, **Figure 6B**) within these active enhancers. This association was especially strong for insertion/deletions (indels), possibly due to the greater functional impact of disrupting multiple nucleotides (p=0.007, **Figure 6C**).

*__3) Analysis of gene co-expression to identify spatiotemporal convergence of ASD-associated genes__*. We considered the convergence between 9 ASD genes[15] for gene expression data from 57 neurotypical brains that spanned 15 developmental periods and 16 brain regions[47]. To identify spatiotemporal windows whilst retaining sufficient numbers of samples for co-expression analysis we used hierarchical clustering to identify four groups of brain regions and considered each of these in 13 overlapping time periods each composed of three developmental periods (**Figure 7A**). Within each of the resulting 52 (4 x 13) spatiotemporal windows we built networks around nine high confidence ASD genes by selecting the top 20 co-expressed genes. We assessed these 52 windows for spatiotemporal convergence related to ASD etiology through the degree of enrichment for 126 independent low confidence ASD genes (**Figure 7A**). We observed strong spatiotemporal convergence between ASD risk genes in the prefrontal and primary motor-somatosensory cortex during mid-fetal development (**Figure 7A**)[15]. Analysis of cell type specific marker genes within this network showed enrichment for cortical projection neurons. This result that has been replicated by three complementary techniques: WGCNA[31], cell specific enrichment analysis[58], and NETBAG+ systems analysis[59].

*__4) Comparison of ASD-related gene sets and gene expression analysis of post-mortem ASD brains.__* Two prior analyses have identified gene co-expression WGCNA modules that are differentially expressed in the brain in ASD cases compared with controls. The microarray analysis by Voineagu et al.[9] identified a module enriched

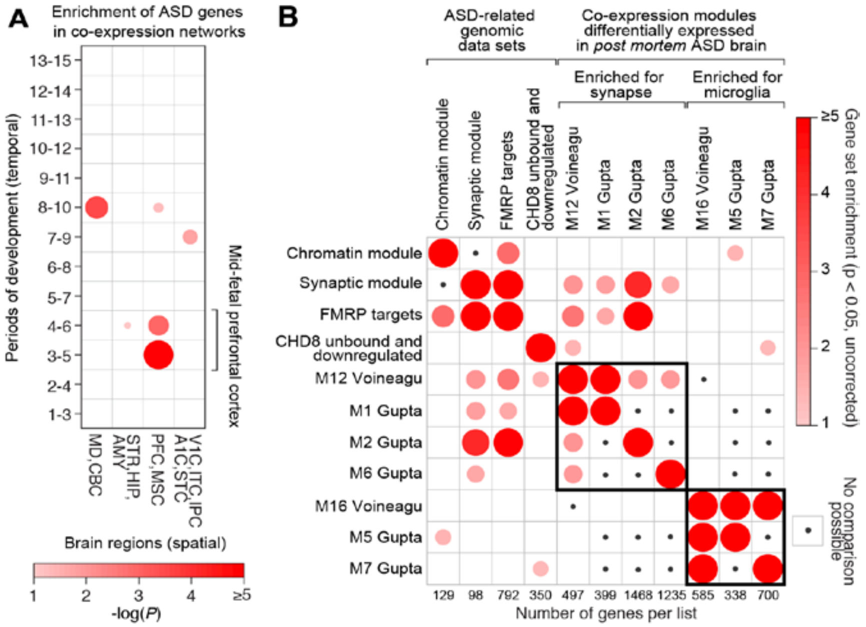


***Figure 7. Enrichment of ASD-associated genes in gene expression data. A.*** *Spatiotemporal co-expression networks were formed around nine high confidence ASD genes for 4 groups of brain regions (x-axis) and 13 overlapping developmental periods (y-axis). The –log(P) value for enrichment with 126 low confidence ASD genes is shown by the size and shade of the circle. Strong enrichment is observed in the mid-fetal PFC and primary motor-sensory cortex (PFC-MSC).* ***B.*** *Four ASD related gene sets*[9,52,53] *compared to seven WGCNA co-expression modules that are differentially expression in post mortem ASD brains (right). Fold enrichment is indicated by the size and shade of the circle. A synaptic and microglial module are indicated by the black rectangles. Small black circles show gene sets that are non-overlapping by definition (e.g. WGCNA modules in the same analysis).*

for synaptic genes (M12) that overlaps with three modules (M1, M2, and M6) identified using RNA-seq in Gupta et al[60] (**Figure 7B**). Similarly, one module enriched for microglial genes (M16) was observed in the Voineagu et al.[9] paper and this overlaps with two modules (M5 and M7) identified in the Gupta analysis (**Figure 7B**). We compared these seven WGCNA modules with four sets of ASD-related genes: the chromatin and synaptic modules from our recent analysis of exome and CNV data (**Figure 6A**)[17], RNA targets of the fragile X protein FMRP[61] that are strongly enriched within ASD-associated genes[26], and genes that are downregulated in CHD8 knockdown but not bound by CHD8 on ChIP-Seq analysis that have been described as targeting synaptic genes associated with ASD[62]. The synaptic module and FMRP targets are strongly enriched through the synaptic WGCNA modules suggesting these modules may represent causal factors that persist in the ASD brain. Further analysis is required to determine if these modules are causal or simply a consequence of ASD.

***Experimental design and methods:*** In ***subaim 3.1**, we will increase the spatiotemporal resolution of co-expression analysis of ASD neurobiology*. Our prior analysis of spatiotemporal convergence, described in detail under preliminary data[15], was based on 57 neurotypical brains, 9 high confidence ASD genes (FDR ≤0.05), and 126 low confidence ASD genes (FDR ≤0.3)[17]. These data enabled us to examine 4 groups of brain regions spanning multiple developmental periods (**Figure 7A**). The data from Aim 1 and our progress in ASD gene discovery will allow us to perform this analysis using 87 neurotypical brains, 28 high confidence ASD genes (FDR ≤0.01), and 151 low confidence ASD genes (FDR ≤0.3). As before (**Figure 7A**), the gene expression samples will be divided into spatiotemporal windows using hierarchical clustering to group related brain regions (spatial) and considering overlapping developmental windows (temporal). In each spatiotemporal window we will identify the top 20 co-expressed genes around 28 high confidence ASD genes and, following the logic that a spatiotemporal network relevant to ASD should be enriched for other ASD genes, we will assess the enrichment of the 151 low confidence ASD genes (FDR ≤0.3). The expanded number of brain samples will enable us to use small subdivisions of brain regions and developmental time regions to increase the resolution of the analysis, for example windows spanning one or two developmental periods. In addition, the larger list of high confidence ASD genes will allow us to perform the analysis by building the spatiotemporal networks around subsets of these 28 genes and improve the accuracy of the analysis through cross validation. In addition, we will divide the 28 high confidence genes by the two main functional categories observed, specifically chromatin regulators and synaptic genes, to assess the spatiotemporal dynamics of each functional category separately. The outcome of this aim will be refined gene co-expression networks that show spatial and temporal convergence among ASD risk genes. ***Pitfalls and alternatives:*** The analytical methods described here have been applied to the BrainSpan data using 9 high confidence genes resulting in the discovery of spatiotemporal convergence in the frontal cortex of the mid-fetal brain. This finding has been replicated using complementary methods[58,59]. In this aim we will be increasing the resolution through the inclusion of additional gene expression data and novel ASD-associated genes[17], therefore we do not foresee complications. An alternative 'top down' methodology such as WGCNA, in which co-expression modules are generated from the complete dataset and are then assessed for enrichment

of ASD genes, has yielded similar findings[31]. We will also apply this complementary WGCNA method across spatiotemporal windows.

In ***subaim 3.2**, we will identify ASD-associated non-coding de novo mutations in regulatory loci***. Under pre-existing funding arrangements we will have access to whole-genome sequencing (WGS) data for 5,120 individuals from 1,280 quartet families composed of two parents, an affected child, and an unaffected sibling control. We have previously reported an increased burden of *de novo* mutations between the affected and unaffected siblings[17] and we have observed this for *de novo* CNVs in microarray data and *de novo* loss of function mutations in exome data. To identify functional non-coding *de novo* mutations in regulatory loci, we will leverage the integrated RNA-Seq, ChIP-Seq, and HiC data from Aims 1 and 2 with the *de novo* mutation identification approach described in our preliminary data (**Figure 6**). To maximize our ability to discover compartments of the genome that carry risk we will assess *de novo* burden in three sets of loci: **1**) All regulatory loci identified in neurotypical brain divided by function (e.g. promoter, 3`UTR); **2**) Regulatory loci identified in neurotypical brain with a relationship to 28 high-confidence ASD genes; and **3**) Regulatory loci identified in neurotypical brain with a relationship to the points of convergence for ASD genes identified in Subaim 3.1 such as prefrontal cortex in mid-fetal development. The outcome of this aim will be non-coding mutations and regulatory loci that show association with ASD.

***Pitfalls and alternatives:*** Our methods for identifying *de novo* mutations in whole genome sequencing data are well developed and we have demonstrated a >95% confirmation rate for the mutations predicated. Additionally, our preliminary data, based on 40 families, shows evidence of ASD association for *de novo* mutations within enhancers active in human dlPFC (**Figure 6B and C**). This suggests the proposed study of 1,280 families will offer sufficient power even if the overall contribution of *de novo* mutations in the non-coding genome to ASD etiology is relatively weak. To maximize our chance of identifying ASD associated non-coding variants we will assess only the loci with the strongest evidence of functional activity, including the larger mutations, such as indels, that may carry the greatest risk. Concurrently, Dr. Sanders has an established collaboration with Mike Talkowski and the GATK CNV/SV working group to develop methods that maximize our sensitivity for detecting indels and small CNVs in whole genome sequence data.

In ***subaim 3.3**, we will identify points of spatiotemporal convergence using ASD associated non-coding mutations:* Non-coding elements such as enhancers frequently show a degree of specificity to particular developmental time points, brain regions, or cell types[63]. We will use the ASD-associated non-coding *de novo* mutations in regulatory loci and regulatory loci related to ASD associated genes to assess which integrated regulatory networks from Aim 2 show the greatest enrichment for these non-coding mutations. By considering the brain regions and developmental epochs in which these networks exist we will assess points of spatiotemporal convergence critical to ASD. The outcome of this aim will be an independent analysis of points of spatiotemporal convergence in ASD based on non-coding mutations and regulatory loci.

***Pitfalls and alternatives:*** This aim relies on the discovery of specific ASD-associated regulatory loci through the discovery of numerous *de novo* mutations in cases. Due to the small size of regulatory regions we may not see this clustering in a single regulatory element. Should this be the case we will use genomic annotation to rank the regulatory loci with a single mutation, for example considering conservation, constraint [64], and large mutations such as indels that are more likely to disrupt the element (**Figure 6C**).

In ***subaim 3.4**, we will assess regulatory networks that are observed in the post mortem ASD brain.* Aims 3.1 to 3.3 focus on neurotypical brains and their association with ASD-associated mutations. In this aim we will assess the enrichment of ASD-associated genes, non-coding mutations and regulatory networks that differ between *post mortem* ASD and neurotypical brains (**Figure 6**). Because genetic variants associated with ASD precede the onset of ASD symptoms, enrichment for these mutations will suggest that such networks are causal (**Figure 7**) to the ASD phenotype. Conversely, a lack of enrichment for these mutations in ASD-relevant networks will suggest the network is consequential to ASD. The outcome of this aim will therefore be to distinguish ASD-specific regulatory networks that are likely to be causal from those that may be consequential.

***Pitfalls and alternatives:*** Methods for assessing such enrichment are well established and we already have a large list of ASD-associated genes; we foresee no complications in performing this aim. The main challenge lies in the interpretation of a regulatory network that does not enrich for ASD-associated genes (e.g. microglia in existing *post mortem* analyses, **Figure 7B**), since this may indicate a non-causal relationship or reflect and incomplete list of ASD-associated genes. We will therefore focus on networks with positive enrichment for these genes and acknowledge the complexities of interpreting a negative result.

**TIMELINE AND MILESTONES SECTION** See Other Attachments

## BIBLIOGRAPHY

1. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature reviews. Genetics* **13**, 537-551 (2012). pmcid: 4110909.
2. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics* **45**, 984-994 (2013). pmcid: 3800159.
3. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. *Science* **349**, 1489-1494 (2015). pmcid:
4. Krystal, J. H. & State, M. W. Psychiatric disorders: diagnosis to therapy. *Cell* **157**, 201-214 (2014). pmcid: 4104191.
5. Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nature neuroscience* **17**, 764-772 (2014). pmcid: 4077789.
6. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of *de novo* mutations in the genetics of autism spectrum disorders. *Nature reviews. Genetics* **15**, 133-141 (2014). pmcid:
7. Walsh, C. A., Morrow, E. M. & Rubenstein, J. L. Autism and brain development. *Cell* **135**, 396-400 (2008). pmcid: 2701104.
8. Huguet, G., Ey, E. & Bourgeron, T. The genetic landscapes of autism spectrum disorders. *Annu Rev Genomics Hum Genet* **14**, 191-213 (2013). pmcid:
9. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384 (2011). pmcid: 3607626.
10. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature reviews. Genetics* **16**, 441-458 (2015). pmcid:
11. Abelson, J. F. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317-320 (2005). pmcid:
12. Johnson, M. B. *et al.* Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**, 494-509 (2009). pmcid: 2739738.
13. Sanders, S. J. *et al.* *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012). pmcid: 3667984.
14. State, M. W. & Sestan, N. Neuroscience. The emerging biology of autism spectrum disorders. *Science* **337**, 1301-1303 (2012). pmcid: 3657753.
15. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013). pmcid: 3995413.
16. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat Commun* **6**, 6404 (2015). pmcid: 4355952.
17. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233 (2015). pmcid: 4624267.
18. Luo, R. *et al.* Genome-wide transcriptome profiling reveals the functional impact of rare *de novo* and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet* **91**, 38-55 (2012). pmcid: 3397271.
19. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* **14**, 1109-1120 (2015). pmcid:
20. Preuss, T. M. Human brain evolution: from gene discovery to phenotype discovery. *Proc Natl Acad Sci U S A* **109 Suppl 1**, 10709-10716 (2012). pmcid: 3386880.
21. Lui, J. H. *et al.* Radial glia require PDGFD-PDGFRbeta signalling in human but not mouse neocortex. *Nature* **515**, 264-268 (2014). pmcid: 4231536.
22. Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647 (2013). pmcid: 3922239.
23. Zeng, J. *et al.* Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* **91**, 455-465 (2012). pmcid: 3511995.
24. Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213-217 (2009). pmcid: 2778075.
25. The PsychENCODE Consortium *et al.* The PsychENCODE project. *Nature neuroscience* **Manuscript accepted** (2015). pmcid:

26. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014). pmcid:

27. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246-250 (2012). pmcid: 3350576.

28. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012). pmcid: 3613847.

29. Sanders, S. J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011). pmcid: 3939065.

30. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622 (2012). pmcid: 3528801.

31. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021 (2013). pmcid:

32. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012). pmcid: 3356448.

33. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009). pmcid: 2858594.

34. Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680 (2014). pmcid:

35. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518-529 (2013). pmcid: 3894107.

36. Network & Pathway Analysis Subgroup of Psychiatric Genomics, C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209 (2015). pmcid: 4378867.

37. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523 (2014). pmcid: 4390143.

38. Parikshak, N. N. *et al.* Global changes in patterning, splicing and primate specific lncRNAs in autism brain. *Manuscript in preprint. Please see appnedix.* (2015). pmcid:

39. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010). pmcid: 3003124.

40. Won, H. *et al.* Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development. *Manuscript in preprint. Please see appnedix.* (2015). pmcid:

41. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003 (2012). pmcid: 3816492.

42. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* **28**, 1045-1048 (2010). pmcid: 3607281.

43. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nature genetics* **45**, 723-729 (2013). pmcid: 3932982.

44. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428 (2014). pmcid: 4208299.

45. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014). pmcid:

46. McCarthy, S. E. *et al. De novo* mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658 (2014). pmcid: 4031262.

47. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011). pmcid: 3566780.

48. Sarnat, H. B., Nochlin, D. & Born, D. E. Neuronal nuclear antigen (NeuN): a marker of neuronal maturation in early human fetal nervous system. *Brain Dev* **20**, 88-94 (1998). pmcid:

49. Pletikos, M. *et al.* Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron* **81**, 321-332 (2014). pmcid: 3931000.

50. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206 (2014). pmcid: 4105188.

51. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013). pmcid: 3869051.

52. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009). pmcid: 2672628.

53. Habegger, L. *et al.* RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**, 281-283 (2011). pmcid: 3018817.
54. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015). pmcid:
55. del Rosario, R. C. *et al.* Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nature methods* **12**, 458-464 (2015). pmcid:
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009). pmcid: 2705234.
57. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010). pmcid: 2928508.
58. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J Neurosci* **34**, 1420-1431 (2014). pmcid: 3898298.
59. Chang, J., Gilman, S. R., Chiang, A. H., Sanders, S. J. & Vitkup, D. Genotype to phenotype relationships in autism spectrum disorders. *Nature neuroscience* **18**, 191-198 (2015). pmcid: 4397214.
60. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* **5**, 5748 (2014). pmcid: 4270294.
61. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261 (2011). pmcid: 3232425.
62. Sugathan, A. *et al.* CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc Natl Acad Sci U S A* **111**, E4468-4477 (2014). pmcid: 4210312.
63. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521-1531 (2013). pmcid: 3989111.
64. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* **9**, e1003709 (2013). pmcid: 3749936.

## SPECIFIC AIMS

The necessity for understanding gene regulation in human brain development is supported by several recent discoveries. For example, most inherited common genetic variation underlying neuropsychiatric diseases lies in non-coding regions and is presumed to exert pathogenic effects via the regulation of gene expression and splicing[1-4]. Additionally, most non-inherited (*de novo*) highly penetrant ASD risk genes are enriched in co-expression modules and protein interaction networks related to chromatin remodeling and transcriptional regulation[3-8]. Moreover, a specific shared pattern of transcriptional dysregulation is observed in the cerebral cortex in slightly more than 2/3 of post-mortem ASD cases[9,10]. Taken together, these observations emphasize the importance of integrating transcriptomic and epigenomic data with higher-order chromatin interactions to better understand the putative mechanisms underlying dysregulated genes and networks in ASD and other psychiatric disorders, a fundamental goal of psychENCODE. **The primary goal of this application is to extend our ongoing analyses of healthy and ASD brains under the psychENCODE consortium with the inclusion of additional genomic features, brain regions, developmental time points and cell-type specific analyses.** By performing these analyses we will enhance this public resource and improve our understanding of the molecular processes underlying normal human neurodevelopment and ASD.

Our group has been collaborating closely for a decade[11-15], bringing together expertise in developmental neurobiology, human tissue biobanking, genetics and genomics, statistics, bioinformatics and systems biology. Several key conceptual threads have been apparent in our work together: 1) Revealed new insights into human neurodevelopment through functional genomic profiling of postmortem tissue and cell culture models[12,16]; 2) Assessed rare and *de novo* mutations for ASD association[13,17,18], based on the notion that down-stream analyses are only as good as the genes that go into them; 3) Identified the neural processes and pathways that are altered in the presence of ASD-associated mutations, as well as when and where these processes and pathways occur in the developing human brain[15,17,19]. Here, we propose to continue this highly productive collaboration and expand psychENCODE phase 1 efforts through three integrated aims.

**Aim 1. Time, region and cell type-specific molecular profiling of control and ASD brains.** In *subaim 1.1*, we will profile the transcriptome (by RNA-seq), *cis*-regulatory elements (ChIP-seq) and 3D chromatin architecture (Hi-C) in neurotypical dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR) during mid-fetal development, infancy, childhood, adolescence and adulthood. To address cellular heterogeneity and to complement the psychENCODE phase 1 tissue level data analyses, we will obtain these data from neuronal and non-neuronal nuclei collected with fluorescence-activated nuclei sorting (FANS). In *subaim 1.2.*, complementary genomic analyses will be done on the FANS nuclei from syndromic and idiopathic ASD brains and matched control brains**,** to identify transcripts, regulatory elements, and 3D chromatin structures altered in ASD in brain region and cell type-specific manners.

**Aim 2. Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains.** In *subaim 2.1*, each dataset generated in Aim 1 will be analyzed to identify differences between the developmental stages and two major cell types in healthy and ASD tissue. Furthermore, these datasets will be integrated to gain comprehensive insights into the underlying mechanisms; Hi-C defined physical intrachromosomal interactions will be intersected with ChIP-seq to identify functional interactions between regulatory sequences potentially associated with transcriptional changes. In *subaim 2.2*, we will harmonize and integrate our multi-omic datasets with other psychENCODE studies and large-scale genomic datasets, such as BrainSpan, CommonMind, ENCODE, GTEx and REMC.

**Aim 3. Spatiotemporal analysis in ASD.** Our prior work assessed the enrichment of ASD genes in spatiotemporal co-expression networks to identify the frontal cortex during mid-fetal development as a critical window in ASD etiology. In *subaim 3.1*, we will use the neurotypical gene expression data and our expanded list of ASD associated genes to increase the resolution of this spatiotemporal analysis. In *subaim 3.2* we will use whole-genome data for 5,120 individuals in ASD families to identify non-coding *de novo* mutations within the regulatory loci identified in neurotypical brains in Aims 1 and 2. In *subaim 3.3* we will use these non-coding mutations and the regulatory networks from Aim 2 to perform an independent assessment of spatiotemporal convergence in ASD to complement our gene-based analysis in subaim 3.1. Finally, in *subaim 3.4* we will use the regulatory networks that are specific to the ASD brain identified in Aim 2 to assess enrichment of ASD-associated genes and non-coding mutations thus demonstrating that such networks are causally linked to ASD rather than simply a consequence of ASD. At the completion of this aim we will have three independent assessments of spatiotemporal convergence in ASD from ASD-associated genes, ASD-associated regulatory loci, and ASD-associated networks in the *post mortem* brain.

## RESEARCH STRATEGY

## SIGNIFICANCE

Neuropsychiatric disorders such as autism spectrum disorder (ASD), bipolar disorder (BD), and schizophrenia (SCZ) are complex and devastating illnesses with considerable morbidity and mortality, as well as high personal and societal costs. Many of them are also polygenic, with multiple variants, both rare and common, spread throughout the genome influencing the disease risk[3]. Recent studies have identified rare variants contributing to psychiatric disorders that are enriched in genes involved in global gene regulation and chromatin modification, and many common risk variants are enriched in regulatory regions of the human genome, regions whose functions are poorly understood. The interpretations of these variations in regulatory regions will certainly be improved with better maps of RNA transcripts, regulatory elements, and chromatin states in the human brain. The age of onset and progression of major psychiatric disorders also varies (**Figure 1**) necessitating the study of the temporal dynamics of gene regulation during human brain development and recognizing the developmental context of psychiatric disorders. An emerging body of research indicates that many aspects of the development and physiology of the human brain are not well recapitulated in model organisms[20-24] **and therefore it is increasingly apparent that psychiatric disorders need to be understood in the broader context of human brain development and physiology.**

In recent years, considerable effort has been made by many studies, including large-scale efforts by ENCODE, NIH Roadmap (REMC) and GTEx projects to survey the diversity of *cis*-acting regulatory regions and RNA species of the human genome across different tissues and time points. However, a comprehensive catalog of transcripts, regulatory elements, epigenetic modifications, and chromatin structure from the human brain during development and in distinct brain regions and cell types is lacking. The PsychENCODE (phase 1) projects have initiated these efforts.

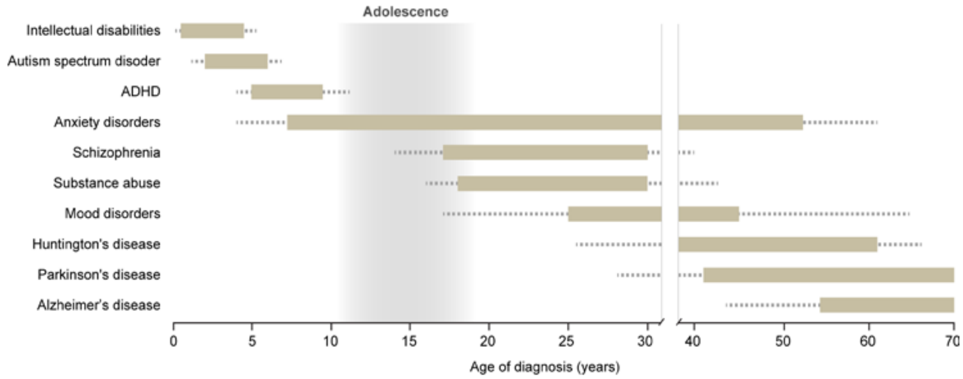


**Figure 1. Psychiatric and neurological disorders have discrete ages of onset.** *The bars indicate the age range that each disorder commonly affects, with less frequent ages of diagnosis denoted as dotted lines. This variability is indicative of dysregulation of tightly controlled developmental processes and highlights the necessity of defining the spatio-temporal molecular processes in human brain.*

**PsychENCODE consortium projects.** The key goals of the PsychENCODE project are to provide an enhanced framework of regulatory elements, catalog epigenetic modifications, and quantitate coding and non-coding RNA and protein expression in a tissue- and cell-type specific manner from neurotypical (healthy) control brains and diseased post-mortem human brains[25]. These efforts will be complemented with integrative analyses, as well as with functional characterizations of disease-associated genomic elements using human neural cell systems or the developing mouse brain. However, the human brain is heterogeneous cellularly and its development is regionally asynchronous and prolonged. *To overcome issues that hamper the potential benefits of initial psychENCODE studies, we will apply several approaches to address regional and cellular heterogeneity, prolonged development, and new genomic methods in the context of brain development and ASD.*

Here we focus on **neurotypical (control) brain and ASD**, which is a complex developmental syndrome with a significant genetic contribution. Although considerable genetic and phenotypic heterogeneity has complicated efforts to establish the biological substrates of the syndrome, the emergence of reliable genetic findings has started to shed light on potential pathogenic mechanisms, providing an extraordinary opportunity for developing a mechanistic understanding of the disorder. Recent studies suggest that **over 500 rare, *de novo* mutations contribute to ASD risk and no single genetic mutation accounts for more than 1% of ASD cases**[13,17,26-30], consistent with significant heterogeneity in this, and other neuropsychiatric disorders[3]. Despite this heterogeneity, mapping **ASD risk genes onto co-expression networks that represent normal human brain development has revealed that ASD genes coalesce in modules related to chromatin remodeling and transcriptional regulation during early fetal brain development, suggesting potential convergent pathways in the disorder**[9,15,27,31]. Another remarkable finding that parallels the convergence of genetic findings in developmental pathways is the identification and validation of shared transcriptional changes in postmortem brain in ASD[9]. This transcriptional dysregulation, coupled with the evidence that large effect size *de novo* ASD

risk genes are highly enriched in chromatin modifying genes (many of which are expressed in early fetal brain development), emphasizes the importance of understanding the nature and extent of chromatin disorganization in ASD brain and in normal brain development. **Further, since these data suggest distinct neuronal and glial gene dysregulation, it is crucial to delineate the profiles of these major cell types.** In addition to our ongoing efforts in PsychENCODE phase I project, this proposal provides critical advances in our understanding of the role(s) of non-coding functional elements in the pathophysiology of ASD and a scaffold for understanding chromatin structure and gene regulation across normal brain development. Overall, the approach proposed here will provide mechanistic insights that connect distinct transcriptional programs associated with ASD pathogenesis, and will provide a resource of the mechanisms of gene regulation across brain development to inform other neuropsychiatric disorders, a key goal of psychENCODE. **This work also leverages psychENCODE phase 1 projects by adding significant new data to expand the value of the resource and by directly addressing key areas of interest in control and ASD brains as outlined in RFA-MH-16-230**: ***1)*** *Generation of comprehensive, high resolution human brain region/cell type and age-specific maps of different classes of RNA transcripts, regulatory elements, chromatin states, chromatin conformation, and chromatin interactions;* ***2)*** *Identification of human brain region/cell type and age-specific molecular processes;* ***3)*** *Integration of these newly generated multi-omic datasets, from diseased and healthy control brains, with large-scale genomic resources;* ***4)*** *Generation and analysis of high-depth, whole genome sequencing data to allow for improved evaluation of various genetic alterations; and* ***5)*** *Development of comprehensive molecular models of disease (i.e., ASD) using systems biology approaches.*

## INNOVATION

This proposal is innovative in several aspects. First, to the best of our knowledge, the systematic discovery and functional characterization of genomic non-coding elements and 3D chromatin architecture has not been performed in healthy developing human brains or ASD brains at a cell type-specific resolution. For example, we use Hi-C, which combines chromosome conformation capture and NextGen sequencing to identify physical interactions that capture multiple levels of chromosome architecture ranging from nuclear configuration ("compartments" of about 5Mb) to TADs (domains of 500kb on average) and gene loops (often reflect enhancer promotor relationships; 40kb average), and is the only such method that spans all of these levels, genome-wide[32-34]. Second, this project will conduct direct analysis of one of the largest collection of well-characterized high quality healthy as well as syndromic and idiopathic ASD postmortem brains. Third, we will combine fluorescence-activated nuclei sorting (FANS) with advanced genomic techniques to analyze multiple genomic features in archived development control and ASD brains. Fourth, we will leverage these analyses with our ongoing psychENCODE phase 1 tissue level analyses and other recent large-scale genomic resources, such as BrainSpan, ENCODE, GTEx and Roadmap project. Therefore, our proposed data and integrated analyses has potential to improve our understanding of genomic processes and normal human brain development as well as diagnostics, neurobiology and treatment of ASD.

## COLLABORATION

This collaboration brings together multiple groups with long standing expertise in developmental neurobiology, psychiatry, human biobanking, genetics and genomics, statistics, bioinformatics, and systems biology that have worked closely with one another for almost a decade as evidenced by many co-publications. Several key conceptual threads have been apparent in our work together related to human brain development and neuropsychiatric disorders: **1**) Revealed new insights into human neurodevelopment through functional genomic profiling of postmortem tissue and cell culture models[12-16]; **2**) Assessed rare and *de novo* mutations for ASD association[13,17,18]; **3**) Identified the neural processes and pathways that are altered in the presence of ASD-associated mutations, as well as when and where these processes and pathways occur in the developing human brain[15,17,19]. In addition, M. Gerstein (Yale) and Z. Weng (University of Massachusetts), experts in bioinformatics and computational biology, are leaders of the PsychENCODE DAC, which will normalize the data to remove batch effects, establish uniform data processing pipelines and build calibration resources for all assays to enable comparison and integration of the data generated by all psychENCODE groups. The efforts of each group will be tightly integrated in order to communicate progress and results, design and implement analytical tools, and transfer data. Given the complexity of human neurodevelopment and genetics/neurobiology of ASD, we believe that integrating the respective expertise of these groups, and their respective collaborators at UCLA (Ernst and Geschwind), UCSF (Sanders, State and Willsey), UMass (Weng), and Yale (Gerstein and Sestan), offers the best opportunity to better understand human brain development and ASD through functional genomics. Here, we propose to leverage our expertise and continue this highly productive collaboration and expand psychENCODE phase 1.

**ELEMENTS UNIQUE TO THIS SITE** (UCSF; State, PI; Sanders and Willsey, co-investigators)
The UCSF team will combine the data from Aims 1 and 2 with whole-genome sequencing (WGS) data for 5,120 individuals in 1,280 ASD families to identify when, where, and in which cells the etiology of ASD occurs. In *Subaim 3.1* they will use the additional gene expression data and ASD gene discovery to increase the resolution of their prior spatiotemporal analysis that implicated prefrontal cortex during mid-fetal development. In *Subaim 3.2* they will use regulatory loci from Aims 1 and 2 to filter de novo mutations in non-coding regions in the WGS data. In *Subaim 3.3* they will use the non-coding mutations from Subaim 3.2 and the integrated analysis of regulatory networks from Aim 2 to perform an independent spatiotemporal analysis of when, where, and in which cells ASD etiology occurs. Finally in *Subaim 3.4* they will assess the enrichment of coding and non-coding mutations from WGS in the integrated regulatory, transcriptional and molecular networks in ASD brains to provide evidence for these networks being a cause rather than a consequence of ASD.

## APPROACH

The objective of this proposal is to extend our ongoing tissue level analyses of healthy and ASD brains under the psychENCODE consortium with the inclusion of additional genomic methods, brain regions, developmental time points, and cell-type specific analyses. By performing three integrated aims (**Figure 2**) we propose to enhance this public resource and improve our understanding of the molecular processes underlying normal human neurodevelopment and ASD.

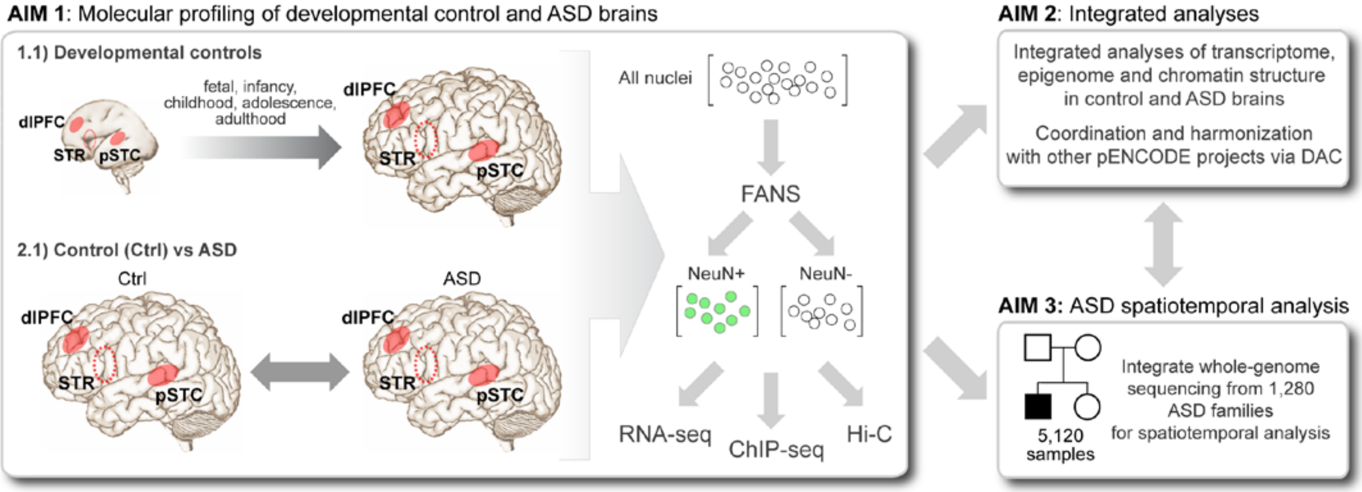

AIM 1: Molecular profiling of developmental control and ASD brains
1.1) Developmental controls
fetal, infancy, childhood, adolescence, adulthood
dlPFC
STR
pSTC
dlPFC
STR
pSTC
2.1) Control (Ctrl) vs ASD
Ctrl
ASD
dlPFC
STR
pSTC
dlPFC
STR
pSTC
All nuclei
FANS
NeuN+
NeuN-
RNA-seq
ChIP-seq
Hi-C
AIM 2: Integrated analyses
Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains
Coordination and harmonization with other pENCODE projects via DAC
AIM 3: ASD spatiotemporal analysis
5,120 samples
Integrate whole-genome sequencing from 1,280 ASD families for spatiotemporal analysis

**Figure 2. Schematic workflow of three specific aims.**

### Aim 1. Time, region and cell type-specific molecular profiling of control and ASD brains.

***Rationale and preliminary supporting data:*** Three major observations provide motivation for this aim. The *first* is the recognition that genomic data, including transcriptomic, epigenetic and physical chromatin structure, from the relevant neurotypical tissue (control), spanning the key epochs of neurodevelopment and function from fetal to adult periods, provide a new and previously unobtainable view of genetic risk for psychiatric disease[10,15,16,31,35,36]. The *second* is that brain is comprised of an extremely heterogeneous mixture of cell types that exhibit distinct molecular profiles, including glia-to-neuron ratios that could show considerable fluctuations across normal development or in certain disease states. The *third* is the observations of differences in transcriptome organization via tissue-level gene co-expression network analysis conducted between ASD and normal brains[9]. Thus, here **we propose to create a region and cell type-specific normal developmental scaffolding on which to frame disease variants via transcriptional (RNA-seq), epigenetic (ChIP-seq) and chromatin architecture (HiC) profiling of neuronal and non-neuronal cells at key epochs in human brain development (subaim 1.1), as well as compare these profiles in ASD and matched control brains (subaim 1.2) to help elucidate the mechanisms by which genetic variation alters brain development and function, leading to ASD and related neuropsychiatric conditions.** While several genomic features are currently being analyzed in control and ASD brains by our and other groups in the psychENCODE consortium, cellular heterogeneity during development, other genomic features (e.g. 3D chromatin contacts), have yet to be addressed. To address these issues, we will utilize our large, high quality, phenotypically well-characterized human brain collection (see Facilities and Resources section), as well as newly implemented methods to collect molecularly defined cell type specific nuclei from archival human postmortem brains in this collection.

**Our preliminary data demonstrates a clear pattern of transcriptional dysregulation is observed in 2/3 of ASD brains[9], which we have now confirmed in our psychENCODE phase 1 projects (in a more than**

**double sized sample of cases and controls)** using tissue level RNA-seq and ChIP-seq (H3K4me3 and H3K27ac) in multiple brain regions in 43 idiopathic ASD cases, 8 cases with chromosome 15q11-13 duplication syndrome (dup15q) and ASD, and 63 controls[37,38]. We also observe that post mortem brain from patients with ASD caused by (dup)15q11-13 share this same pattern at all levels of differential protein coding gene expression, splicing and lncRNA[38]. As the first step in exploring potential mechanisms, we performed epigenetic profiling of ASD vs. control brains with H3K27ac marks, which indicate active enhancers[39]. Genes with differential H3K27ac peaks in their promoter regions (5000bp upstream of the transcription start site) were enriched with neuronal genes with changes in expression. This result demonstrates that transcriptional dysregulation in ASD is partially mediated by changes in histone/chromatin modifications. Furthermore, the two major groupings of modules derived from whole tissue gene expression analysis sort into those up-regulated and expressed in glia, and those down-regulated in neurons[9,38], strongly motivating our plan for transcriptional, epigenetic and Hi-C profiling in neurons and non-neuronal (glial) cells independently.

**Another of the important advances in methodology that will be implemented here is the assessment of 3-D chromatin structure across to different brain regions and cell types, and 5 key epochs of normal brain development and in ASD brains.** Our preliminary data strongly supports the value of these data and our ability to perform and analyze these experiments (see also[40]). We established an efficient Hi-C protocol and obtained high resolution data (10 kb resolution; via deep sequencing) from the fetal cortex from 3 individuals dissected into two zones: cortical plate (CP) and germinal zones (GZ) at post-conception week (PCW) 18 (total n = 12 samples: representative heatmap shown in **Figure 3A**). Demonstrating the data quality, principal component of the interchromosomal interaction matrix for GZ shows a high correlation with GC content ($r$=0.791, $P<10^{-256}$), gene number ($r$=0.384, $P<10^{-256}$), DNase I hypersensitivity ($r$=0.416, $P<10^{-256}$), and to a lesser extent, gene expression ($r$=0.129, $P$=1.11x$10^{-66}$; **Figure 3B and C**), recapitulating previous work in cell lines[41]. We next asked how chromatin interactions elicit transcriptional co-regulation. We hypothesized that highly interacting chromatin regions would be co-regulated at least in part by sharing chromatin remodelers and transcription factors (TFs). To test this, we binned chromatin interactions into top and bottom percentiles, and compared the distribution of correlation patterns for genes in the high and low interacting regions of chromatin. We observed that the high interacting regions were significantly biased toward positive correlations (**Figure 4A**), supporting the hypothesis that co-localization can predict co-expression.

We next integrated these data with the epigenomics map from the NIH Roadmap project[42]. By comparing the epigenetic mark combination matrix with the Hi-C contact matrix, we demonstrate that interacting regions exhibit shared epigenetic patterns: loci associated with transcriptional regulation and enhancers are significantly more likely to interact with each other (**Figure 4B**). Comparison of TF binding site (TFBS) combination matrix (generated from TFBS map reported in[43]) with the intrachromosomal contact matrix revealed distinct combinatorial patterns of TF binding likely to mediate chromosome interactions (**Figure 4C**), thus revealing new experimentally testable regulatory relationships.

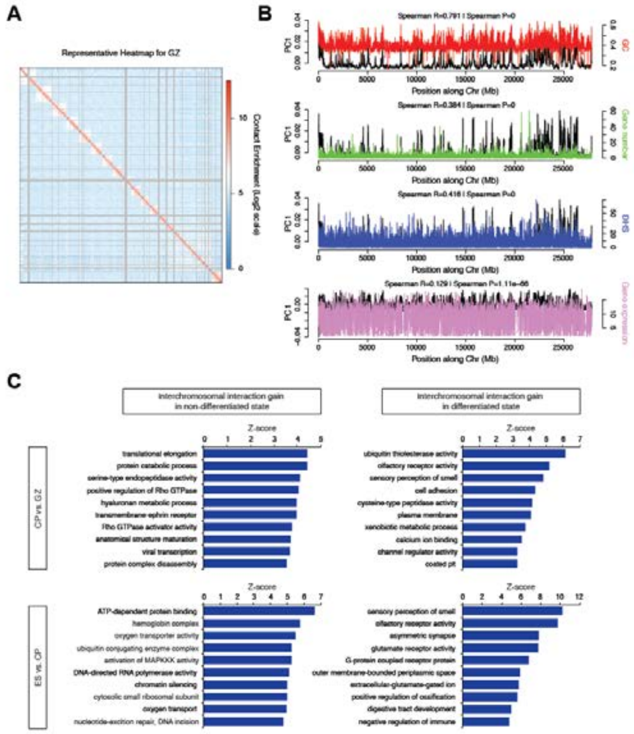


**Figure 3. Chromosome conformation in fetal brains (by Hi-C).** ***A.*** *Representative heatmap of chromosome contact matrix of GZ. Normalized contact frequency (Contact enrichment) is color-coded according to the legend on the right.* ***B.*** *Spearman correlation of PC1 of chromatin interaction profile of fetal brain (GZ) with GC content (GC), gene number, DNase I hypersensitivity (DHS), and gene expression level of fetal brains. These data show relationship of 3D structure to key known functional elements as has been previously shown in other systems.* ***C.*** *Gene ontology (GO) enrichment (GO Elite) of genes located in the top 5% of highly interacting inter-chromosomal regions specific to GZ vs. CP (top), and ES vs. CP (bottom), indicating that genes located on dynamic chromosomal regions are enriched for neuronal function in CP, which contains the more differentiated laminae. Please see Won et al. 2015 in Appendix for higher magnification figure.*

To validate that Hi-C data can identify target genes regulated by single nucleotide polymorphisms (SNPs) in a general setting, we determined if SNPs with a significant effect on gene expression were also identified as interacting by Hi-C using *cis*-expression quantitative trait loci (eQTL) data from adult frontal cortex[44]. Indeed, Hi-C$_{eQTL}$ genes were significantly over-represented with known associated genes from the eQTL study and eQTL SNP-transcript pairs exhibit significantly higher chromatin contact frequency than the null across all distance ranges measured, further supporting the utility of Hi-C to infer the gene or region of activity for regulatory variation. In addition we asked whether significant physical cis-chromosomal contacts identified with Hi-C could inform functional annotation of 108 genome-wide significant schizophrenia loci, most of which lie far outside known coding or other functional regions of the genome.

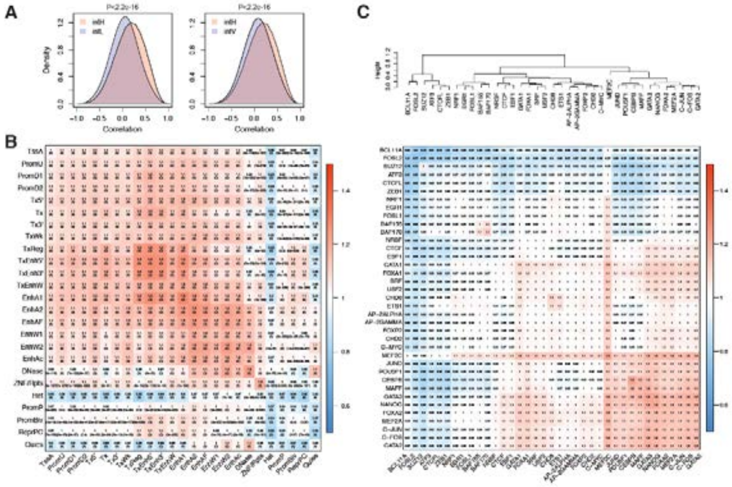


**Figure 4. Highly interacting regions share co-expression patterns, which is mediated by epigenetic regulation**. ***A.*** *The top 10,000 highest interacting regions (intH) in fetal brains both at GZ and CP show positive correlation in their gene expression patterns, while the top 10,000 lowest interacting regions (intL) and top 10,000 highly variant regions (intV) have no skew in the distribution, consistent with random interactions. P-value, Kolmogorov-Smirnov test.* ***B-C.*** *Epigenetic state combination (B) and TFBS combination (C) for intrachromosomal interacting regions. The epigenetic state matrix and TFBS combination matrix were generated by marking loci where two interacting chromosomal bins share epigenetic signature. For example, the epigenetic combination matrix between the active transcription start site (TssA) and active enhancers (EnhA1) is generated by marking where interacting loci have TssA and EnhA1. Intrachromosomal contact frequency map is compared to the epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Odds ratio (OR) and P-values are depicted in the heatmaps (Please see Won et al. 2015 in Appendix for higher magnification figure).*

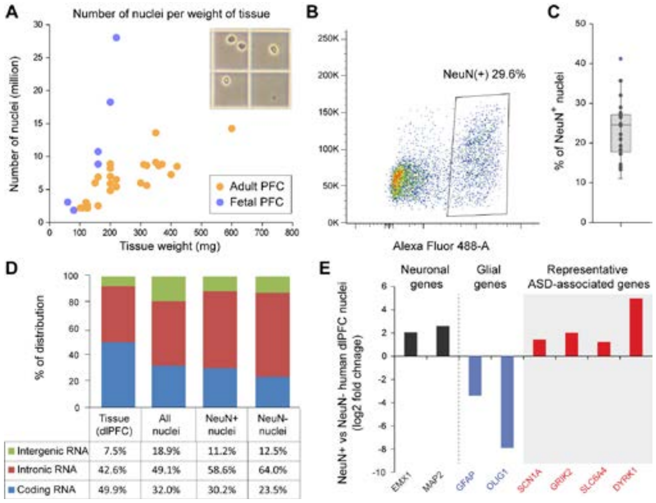
Although SNPs are typically assigned to the closest genes, or those within the LD block, Hi-C indicated that about 50% of the variants were neither adjacent to the index SNPs (most-associated SNP within a locus), nor in LD. Interestingly, Hi-C$_{SCZ}$ genes significantly overlap with ASD *de novo* likely gene-disrupting (LGD) targets[26,45] (CP: OR=2.4, P=1.6x$10^{-5}$, GZ: OR=1.8, P=0.006), indicating a shared genetic etiology between ASD and schizophrenia[46]. The fact that genes with LGD mutations in ASD are associated with regulatory variants in schizophrenia suggests that complete abrogation of these genes may cause developmental defects as in ASD, while regulatory changes in these genes may cause later-onset of neuropsychiatric symptoms as in schizophrenia. *Collectively, these preliminary data demonstrate that we can conduct and analyze genome-wide Hi-C experiments, integrate these data with other epigenetic and transcriptomic data, and use chromatin architecture elucidated by Hi-C to provide novel genome-wide insights into the regulatory mechanisms occurring during neuronal differentiation and disease pathogenesis.*

***Experimental design and methods:*** In ***subaim 1.1.,*** *we will profile the transcriptome (by RNA-seq), cis-regulatory elements (ChIP-seq) and 3D chromatin architecture (Hi-C)* in the control neurotypical dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR). These regions have been implicated in the risk for ASD and schizophrenia[35] and in the cases of dlPFC and pSTC shown to have dysregulated transcriptional patterns in ASD[9]. Recent studies have also highlighted the late mid-fetal frontal cortex as most enriched for co-expression of ASD and schizophrenia *de novo* hits[15,31,35]. Brains from at least 5 key epochs of development representing mid-fetal, infancy, childhood, adolescence and adult brain, and a minimum of 6 subjects (balancing sex when possible) from each of these 5 epochs (30 brains in total) will be profiled.

Cell-type specific chromatin, epigenetic and transcriptome assays are at the core of this project. Mario Skarica, a talented research associate scientist in the Sestan lab, has developed a protocol to isolate high quality nuclei with preserved chromatin and RNA from archival fresh frozen fetal and postnatal human brains. Using this approach he has obtained, on average, 2.57 +-0.8 and 6.93+-3.3 million intact nuclei from 100 mg of the fetal or adult prefrontal gray matter (i.e., fetal CP or adult cortical layers 1 to 6 with a small part of underlying white matter), respectively (**Figure 5A**). Furthermore, we separated neuronal and non-neuronal nuclei, by immunostaining with the NeuN antisera against pan-neuronal splicing protein RBFOX3 (**Figure 5B and C**) and sorting on BD FACSAria IIU Three-Laser System. Starting with infancy and onwards, postnatal gray matter tissue corresponding to six-layered postnatal cortex and small part of adjacent white matter from dlPFC and pSTC, or STR (corresponding to the caudate-putamen with the internal capsule at the septal level) will be processed.

Tissue samples will be dissected directly from frozen tissue blocks using custom dental tools and protocol described in Kang et al., 2011[47]. These dissections will be performed by Nenad Sestan, who has over 2 decades of experience in human neuroanatomy and tissue processing and has microdissected over 1600 tissue samples for exon array profiling of the human brain transcriptome[47]. Given the high proportion of neurons in the cortical plate of the mid-fetal brain (approaching 95% or more), and relatively few neurons that are positive for NeuN at 17-20 PCW in neocortical CP or STR[48], we will not sort NeuN+ and NeuN- nuclei from mid-fetal brains, but instead analyze tissue homogenate and unsorted nuclei from CP of prospective dlPFC and pSTC as well as STR, separately, from corresponding neocortical and striatal GZ (i.e., VZ and SVZ) containing a mixed population of dividing neural stem/progenitor cells with a minor contribution of newborn neurons and glia.

Tissue samples will be pulverized and processed to release nuclei, which will be purified by ultracentrifugation and processed for RNA-seq in the case of mid-fetal samples or in the case of all postnatal specimens (infancy and onwards) sorted into a NeuN+ (predominantly neurons) and NeuN- (mostly glia) fractions. In the past year, we have obtained on average 23.45+-7.2 percentage of NeuN+ nuclei from PFC (**Figure 5C**). This approach will provide unbiased quantitative assessments of cell types in healthy and ASD brains. This approach allows us to simultaneously collect molecularly defined cell type-specific nuclei and isolate DNA, chromatin, and nuclear RNAs. Bulk tissue level RNA-seq is available for dlPFC, pSTC and STR in control and ASD brains as part of psychENCODE phase 1 studies[38], has already been added to enhance the scope of the resource. All brains necessary for this project are currently available in the Geschwind and Sestan labs (see Facilities and Resources section for the list).



| | Tissue (dlPFC) | All nuclei | NeuN+ nuclei | NeuN- nuclei |
|---|---|---|---|---|
| 🟩 Intergenic RNA | 7.5% | 18.9% | 11.2% | 12.5% |
| 🟥 Intronic RNA | 42.6% | 49.1% | 58.6% | 64.0% |
| 🟦 Coding RNA | 49.9% | 32.0% | 30.2% | 23.5% |

**Figure 5. Fluorescence-activated nuclei sorting (FANS) and nuclear RNA-seq of human dlPFC.** ***A.*** *Collection of single nuclei (see insert) from fetal (n=6) and adult (n=29) PFC.* ***B.*** *FANS plot for NeuN immunopositive nuclei.* ***C.*** *Percentage of NeuN+ nuclei collected across different experiments.* ***D.*** *Coverage for exon, intron and intergenic regions of different sequencing technologies.* ***E)*** *Differential expression comparison between NeuN+/- FANS nuclei for neuronal, glial and ASD-related genes.*

Total RNA will be extracted from 1 million nuclei using Norgen's Cytoplasmic & Nuclear RNA Purification Kit. RNA from tissue and cell populations will be depleted of rRNA and sequencing libraries prepared with TruSeq Stranded Total RNA with Ribo-Zero Gold and SMARTer Stranded RNA-Seq Kit, respectively. As expected, our preliminary nuclear RNA-seq analyses revealed higher percentage of unspliced primary transcripts and extensive identification of nuclear-retained long non-coding RNAs (**Figure 5D**). Importantly, we detected robust cell type-specific expression differences, including those of ASD-associated genes (**Figure 5 E)**. RNA-seq libraries will be sequenced on the Illumina HiSeq 2500 at the Yale Center for Genome Analysis (http://ycga.yale.edu/) to generate 100 bp strand specific paired-end sequence at over 40 million reads per end for each sample. For ChIP-seq, 1 million nuclei will be processed through our established protocol using well-characterized ChIP-grade H3K27ac and H3K4me3 antibodies that have been used in psychENCODE phase 1 tissue-level experiments. ChIP-seq libraries will be sequence at HiSeq 2500 at Yale at >40 million reads per sample. Using the standard pipelines developed in the Sestan and collaborating labs, we will perform QC analyses and compare the transcriptome and epigenetic data from different time points and regions to construct spatiotemporal gene and disease state profiles and co-expression networks using computational methods described in Aim 2.

For Hi-C, 2 million nuclei will be prepared from each sample and cross-linked in 1% formaldehyde for 10 min. Cross-linked DNA will then be restriction digested using HindIII, digested chromatin ends filled with biotin-14-dCTP, and resulting blunt-end fragments ligated under dilute conditions to minimize random intermolecular ligations. Following this, crosslinking will be reversed, unligated ends removed by exonuclease digestion (T4), DNA sheared by sonication, and 300-600bp fragments selected. The intermolecular ligation products containing biotin-tagged DNA will be pulled down with streptavidin beads and ligated with Illumina paired end adapters and the library sequenced by Illumina 50bp paired-end sequencing over 3 lanes of the HiSeq 25000 at UCLA, a depth necessary to facilitate sufficient hi-resolution analysis (300-500 million mapped reads), which can also be augmented by pooling samples to increase depth as needed.

In ***subaim 1.2**, complementary genomic analyses will be done on the FANS nuclei from control, and syndromic and idiopathic ASD brains*, to identify transcripts, regulatory elements, and 3D chromatin structures altered in ASD in brain region and cell type-specific manners. We will conduct RNA-seq, ChIP-seq and Hi-C on sorted neuronal and non-neuronal nuclei from 2 cortical regions, dlPFC and pSTC, and STR from 20 matched control and 20 ASD individuals, including 5 dup15q cases. We will select 10 ASD cases manifesting the shared pattern of transcriptional dysregulation observed, 10 without this pattern, and match them to controls to account for potential confounders (sex, age, postmortem interval [PMI], and RNA integrity numbers [RIN]). We will select 5 dup15q brains with most similar breakpoint structures. Hi-C will be performed on sorted nuclei using the identical experimental methods as in subaim 1.1.

***Pitfalls and alternatives:*** The techniques in these proposed experiments are commonly used in our laboratories and we do not expect complications. One potential issue is the obtainment of adequate samples. The Sestan lab has almost 200 high quality frozen human prenatal, early postnatal and adult brain specimens from clinically unremarkable (neurotypical) control donors. Control brains from this collection were used for different BrainSpan and psychENCODE phase 1 projects (see example studies[12,47,49,50] and Resources and Facilities section). Both Geschwind and Sestan labs have tissue samples from over 50 post mortem ASD cases and matched controls with good quality RNA, and have participated in the new initiative at the Simons Foundation to collect additional postmortem ASD brains. A related concern is whether the 20 ASD brains we propose to analyze are sufficient, given the heterogeneity typical of ASD, to detect robust differences between these samples and our controls. However, we were able to detect transcriptional dysregulation in 2/3rds of ASD brains in a smaller cohort17, and by directly comparing ASD brains exhibiting hallmarks of dysregulated transcription with those that do not, we expect to have sufficient statistical power to assess the extent to which 3D chromatin structure contributes to the observed transcriptional changes. Further, the use of 5 dup15q cases provides an additional homogeneous cohort, and as our preliminary results on transcriptome analysis of this cohort demonstrate (appendix), such sample sizes are sufficient. The main pitfall of Hi-C is that it averages chromosome contact population from millions of nuclei. Single-cell Hi-C can complement this limitation[51], but it can capture only one interaction for a given locus. Homogenous population of cells can be achieved by FANS and thus we propose this approach here. Additionally, Hi-C offers other benefits, including the ability to analyze interactions mediated by multiple TFs *en masse* in Hi-C, that are not easily achievable with other methods such as ChIA-PET. While our FANS approach, which follows standards accepted across the psychENCODE projects, is limited to two major groups of cells, we have been implementing the use of other cell type specific nuclear antibodies and single nuclear RNA-seq. Finally, we realize that other regions, including the thalamus, hypothalamus, and hippocampus, may be affected in ASD. We believe our work on the neocortex and STR will develop a framework for understanding of the molecular neuropathology of ASD which can then be extended to include other regions in the future.

**Aim 2. Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains.**

***Rationale:*** We will analyze the data generated in the previous aim to (1) identify *developmentally regulated and cell type specific changes* in the transcriptome, epigenome and the 3D chromatin structure (2) integrate the three types of datasets to gain comprehensive insights into the underlying mechanisms of transcriptional regulation and dysregulation in development and disease, respectively.

***Experimental design and methods*:** In ***subaim 2.1**, several first order analyses will be done for quality control and to provide the data as a processed resource in addition to the raw data*. We will use Illumina CASAVA to purify the low-quality and non-identified reads and Fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), to report fundamental quality parameters. Next, Tophat[52] will be employed to uniquely align the filtered reads to their reference genome and RSEQtools[53] to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. The R package DESeq (http://bioconductor.org/packages/release/bioc/html/DESeq.html) will be used to identify differentially expressed (DEX) genes and well established methods including MATs to identify differential splicing[10,37]. DEX genes will be detected from the reliably expressed coding and non-coding transcripts, which are defined as transcripts with RPKM ≥ 1 in at least 2 samples of different developmental period. ChIP-seq reads will be aligned to the genome by Bowtie. After filtering of low score reads, we will use the MACS platform to call peaks enriched over the input library, and peaks with high empirical FDR will be excluded from further analysis. Thus, we will catalog all potential cis-regulatory elements from our genome-wide histone modification maps in all brain regions across developmental periods.

*For Hi-C analysis*, *hiclib* (https://bitbucket.org/mirnylab/hiclib) will be used to perform all initial analysis on Hi-C data from mapping to filtering and bias correction (see also[40]). Sequenced reads will be mapped to the human

genome by *Bowtie2* (with increased stringency, *--score-min -L 0.6,0.2--very-sensitive*) through iterative mapping and read pairs allocated to HindIII restriction enzyme fragments. Self-ligated and unligated fragments, fragments from repeated regions of the genome, PCR artifacts, and genome assembly errors will be removed. Filtered reads will be binned at 10kb, 40kb, and 100kb resolution to build a genome-wide contact matrix at a given bin size. This contact map depicts contact frequency between any two genomic loci. To decompose biases from the contact matrix and yield a true contact probability map, filtered bins are subjected to iterative correction[41]. Bias correction and normalization results in a corrected heatmap of bin-level resolution. 100kb resolution bins are assessed for inter-chromosomal interactions, 40kb for TAD analysis, and 10kb for gene loop detection. For TAD-level analysis[32], we will quantify the directionality index by calculating the degree of upstream or downstream (2Mb) interaction bias of a given bin, which will be processed by a hidden Markov model (HMM) to remove hidden directionality bias. For gene loop detection, aggregate peak analysis (APA) will be performed that quantifies the aggregate enrichment of putative peak sets by calculating the sum of a series of submatrices derived from a contact matrix[34]. Resulting inter- and intra-chromosomal interaction matrices as well as genome-wide TADs and gene loops will be used for integrative analysis.

*Developmental and cell type-specific changes*: Pearson's correlations between the first principal components (PC1) from different stages and neuronal and non-neuronal cell types, as well as with our own and other published data will be calculated to compare similarities between different cell types. We will explore alternative transcriptional mechanisms or post-transcriptional modifications occurring in normal (and ASD-affected, see below) regions/cells and time points. These can include up- or down-regulating expression, altered spatiotemporal gene expression, imbalanced expression of different alleles (allele-specific expression [ASE]), aberrant splicing events, modified RNA editing sites, fusion transcripts, or loss of function due to frameshift mutations. RNA and epigenome data will also be compared with tissue level psychENCODE phase 1 and BrainSpan's RNA-seq and ChIP-seq data. We will follow up with an analysis of the relative enrichment of each cell-specific marker genes in each subpopulation and use the expression profiles of these genes to guide the identification of an expanded set of cell-type specific markers.

*Integrated analyses*: Spearman's correlations between PC1/PC2 and biological traits (gene expression, histonemark enrichment, GC content, gene density, DNase I hypersensitivity [DHS]) will be calculated. Gene expression and histone mark data generated in subaim 1.1 along with DHS of fetal brain from Epigenomic roadmap[54] will be used and average values per 100kb bin calculated. In addition to the putative cis-elements identified in the same samples, we will also use the 15 state epigenetic marks from Epigenomic Roadmap[54] in genomic regions classified based on compartments averaged across 40kb bins, as well as subject specific psychENCODE data. Epigenetic state counts[54] for one compartment category are normalized by total epigenetic mark number of that compartment category and compared between samples.

*Dysregulation in ASD brains*. Two main data analyses will be performed with the transcriptome data. We will use the same approach as in subaim 2.1 to identify DEX coding and non-coding transcripts (by DESeq) between ASD and matched controls. Gene function enrichment analysis will be performed for these DEX genes. Finally, we will also perform Weighted Gene Co-expression Network Analysis (WGCNA; http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/) to identify modules of differentially co-expressed genes in ASD cases. For ChIP-seq data, once peaks are called and filtered for quality and reproducibility, we will identify and catalog all putative enhancer and promoter sites gained or lost in ASD brains compared to matched control brains, as well as what genes they are associated with.

In Hi-C data, we will assess if compartments, TADs, and contact domain structures are abrogated in ASD brains. Interaction partners for ASD risk genes, as well as gene loops involving ASD risk gene regulatory elements will be examined. Genome-wide interchromosomal contact matrices at high resolution (approximately 10kb) will be compared between ASD and control to identify bins that exhibit the largest chromosomal interaction changes in ASD (here we refer to ASD-specific bins). Gene ontology for these genes as well as their gene expression pattern in ASD may provide novel insights on ASD mechanism. The same approach will be applied to intrachromosomal contact matrices at 10-40kb resolution. TADs in ASD vs. controls will be also compared. The directionality index around ASD-specific TAD boundaries will be calculated to check significance. Moreover, we will examine gene expression level and histone marks on TAD boundaries as well as histone marks on TADs that contain ASD risk genes. Both inter- and intra-chromosomal interaction patterns of the bins that contain differentially expressed genes in ASD or ASD risk genes will be examined in ASD vs. controls. Gene expression pattern and histone states of genomic loci that highly interact with dup15q region will be assessed. *This approach of integrating chromosome interactomes to transcriptomic and epigenetic profiles may delineate epigenetic mechanism behind gene dysregulation in ASD.*

We will also perform integrative network analyses of these multi-level genomic data with genetic variation to understand the causal mechanism of transcriptional alterations in ASD (see also Aim 3). This will include integration of DNA sequence, methylation, chromatin contacts, eQTL and hQTL by this collaborative team of investigators (e.g. to include new hQTL methods by S. Prabhakar and colleagues[55]. Gene loops detected in control and ASD will be also interrogated. Gene loops that are specific to ASD or specific to controls may directly point out aberrant enhancer-promoter interactions, TF binding, or compartmentalization of genome. We will check if ASD-specific gene loops contain any ASD-associated variants (mostly common SNV at this point, although as more whole genome sequencing (WGS) data is available over the next 12 months, we can use these data to annotate potential functions of noncoding variants (Aim 3).

In ***subaim 2.2.**, we will integrate and harmonize data across psychENCODE projects and other relevant genomic resources*. In this aim, the DAC will integrate and harmonize our datasets with other psychENCODE studies and large-scale genomic datasets, such as BrainSpan, CommonMind, ENCODE, GTEx and REMC. The PsychENCODE DAC is led by Mark Gerstein and Nenad Sestan (Yale), Zhiping Weng (University of Massachusetts), who are part of this proposal and Kevin White (University of Chicago). DAC will summarize the major analysis results produced from psychENCODE and organize them into an encyclopedia of regulatory elements in the developing and adult human brain. We are currently building such an encyclopedia for the ENCODE consortium, and we will be able to leverage the methods that we are building for ENCODE and modify them to best serve psychENCODE data. The psychENCODE encyclopedia will include several components. The first component is the raw experimental data, including the expressed transcripts in neuronal and glial cells in various brain regions, the peaks (enriched regions) of an array of histone marks, the open chromatin regions detected using ATAC-seq, the differentially enriched histone mark peaks and open chromatin regions in ASD, BD and SCZ (diseases covered by psychENCODE projects. This component will largely result from a series of uniform processing pipelines, which we will build for analyzing psychENCODE data. The second component will include results that require the integration across multiple data types, including the enhancers in each cell type, the chromatin states called using a combination of histone marks and ATAC-seq data, and the topologically associated domains and compartments called by combining histone marks, ATAC-seq and Hi-C data. The third component of the encyclopedia will provide a higher-order organization to the elements in the first two components. Specifically we will derive the target genes for enhancers in a cell type specific manner, and identify the enhancer-gene links that are disrupted in the three diseases. We will also identify the variations that are linked with difference in gene expression (eQTLs) that are within enhancers that target the corresponding genes. Finally, we will develop a portal to guide the user through the components of the psychENCODE encyclopedia, with multiple entry points, such as genes, GWAS SNPs, or a specific regulatory region in the genome.

***Pitfalls and alternatives:*** Proposed computational approaches are well established in our team and we already have a considerable expertise and collaborative history therefore we foresee no complications in performing this aim. Furthermore, Sestan, State and Geschwind have been part of the BrainSpan project and Ernst, Gerstein and Weng has been part of several other relevant genomic consortia, such ENCODE.

**Aim 3. Spatiotemporal analysis in ASD.**

***Rationale and preliminary supporting data:*** Over the past few years genomic analyses by our labs and others have made rapid progress in identifying genes associated with ASD, in particular through the identification of *de novo* mutations in ASD cases[13,17,26,30,45]. Despite the identification of these ASD-associated genes, progressing to an understanding of ASD neurobiology remains a challenge. Aims 1 and 2 described one approach to discovering this neurobiology through the identification of ASD-specific networks in *post mortem* brains. In Aim 3 we propose a complementary approach through the identification of genomic loci, brain regions, developmental stages, cell types, and neurobiological processes that are enriched for ASD mutations in genes (**subaim 3.1**) and non-coding loci (**subaims 3.2 and 3.3**) in neurotypical brains. Finally, we will test the hypothesis that ASD specific networks observed in *post mortem* brains from Aims 1 and 2 will be enriched for ASD associated mutations (**subaim 3.4**) thus demonstrating that the disruption of this network precedes the diagnosis of ASD and is therefore likely to be a cause of ASD rather than a consequence.

*1) Detection of ASD-associated genetic loci.* We identified rare and *de novo* variants in exome data from 5,563 ASD cases and 13,321 controls alongside rare and *de novo* copy number variants in microarray data from 4,687 ASD cases and 2,100 controls[17]. Comparison of these two data sets showed that small *de novo* deletions in ASD targeted the same set of genes as *de novo* loss of function point mutations in exome data. A combined analysis of exome data and small *de novo* deletions was performed using the Transmitted and *De novo* Association (TADA) method to identify ASD-associated genes. 28 ASD-associated genes were identified with very high confidence (false discovery rate (FDR) ≤ 0.01) and 65 ASD-associated genes were identified with high

confidence (FDR ≤ 0.1). These 65 genes formed a protein-protein interaction (PPI) network with two distinct subnetworks, enriched for chromatin regulatory genes and synaptic genes respectively (**Figure 6A**).

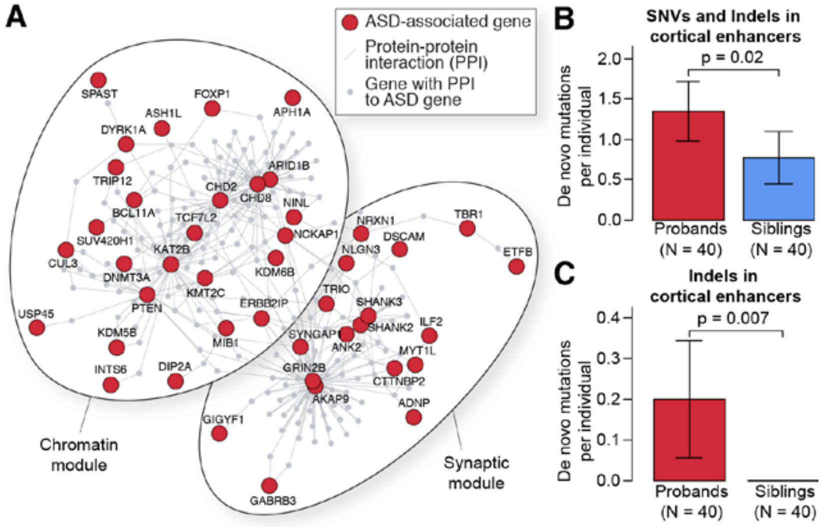


**Figure 6. ASD associated de novo mutations.** **A.** *65 ASD risk genes[9] (red) form a single protein-protein interaction network composed of two subnetworks. The genes in the left subnetwork are enriched for chromatin regulatory gene ontology terms. The genes in the right subnetwork are enriched for synaptic terms.* **B.** *De novo mutations were identified in WGS data for 40 ASD families. The median number of SNV and indel mutations per individual is shown within active enhancers that were identified by bulk tissue ChIP-Seq for H3K27ac in human dlPFC (psychENCODE phase 1 studies). P-values are calculated using linear regression with for paternal age and total de novo mutations per individual included as co-variates.* **C.** *The analysis was repeated for indels only.*

<u>*2) Detection of ASD-associated non-coding variants in whole-genome sequencing (WGS) data*</u>**.** We analyzed WGS data for 40 simplex ASD quartets composed of both parents, an affected child and an unaffected sibling control. The families were selected from the Simons Simplex Collection on the basis of no previous *de novo* loss of function or CNV mutations in exome and microarray data and high paternal age. The samples were sequenced to greater than 30x mean coverage (mean±standard 35.7±5.8). Raw data were aligned to hg19 human reference genome using BWA-mem[56]. Duplicate reads were removed with Picard (http://broadinstitute.github.io/picard/); GATK best practices[57] were used for all downstream steps including, local realignment, base quality score recalibration, SNV and indel calling, cohort-wide joint genotyping, and variant quality score recalibration. Data were normalized within families by only analyzing bases with at least 20 unique reads in all family members. A combination of PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq/) and in-house scripts were used to identify autosomal *de novo* variants based on stringent criteria designed to maximize specificity: minimum genotype likelihood (GQ) ≥20, alternate allele frequency (AB) ≤0.05 in the parents, and 0.3-0.7 in the child, minimum map quality (MQ) ≥30 in all family members, and allelic depth for the alternate allele (AD) ≥8. Approximately 7,000 *de novo* mutations were identified at a rate of 87.0±13.5 *de novo* mutations per child. Confirmation with Sanger sequencing was attempted on 10% of these variants (700) selected at random and achieved a >95% confirmation rate across both SNVs and indels, suggesting identification of *de novo* mutations with accuracy. We used tissue-level ChIP-seq for the histone modification H3K27ac from human dlPFC (psychENCODE phase 1) to identify active enhancers. We observed an increased burden of mutations in cases compared to sibling controls (p=0.02, **Figure 6B**) within these active enhancers. This association was especially strong for insertion/deletions (indels), possibly due to the greater functional impact of disrupting multiple nucleotides (p=0.007, **Figure 6C**).

<u>*3) Analysis of gene co-expression to identify spatiotemporal convergence of ASD-associated genes*</u>**.** We considered the convergence between 9 ASD genes[15] for gene expression data from 57 neurotypical brains that spanned 15 developmental periods and 16 brain regions[47]. To identify spatiotemporal windows whilst retaining sufficient numbers of samples for co-expression analysis we used hierarchical clustering to identify four groups of brain regions and considered each of these in 13 overlapping time periods each composed of three developmental periods (**Figure 7A**). Within each of the resulting 52 (4 x 13) spatiotemporal windows we built networks around nine high confidence ASD genes by selecting the top 20 co-expressed genes. We assessed these 52 windows for spatiotemporal convergence related to ASD etiology through the degree of enrichment for 126 independent low confidence ASD genes (**Figure 7A**). We observed strong spatiotemporal convergence between ASD risk genes in the prefrontal and primary motor-somatosensory cortex during mid-fetal development (**Figure 7A**)[15]. Analysis of cell type specific marker genes within this network showed enrichment for cortical projection neurons. This result that has been replicated by three complementary techniques: WGCNA[31], cell specific enrichment analysis[58], and NETBAG+ systems analysis[59].

<u>*4) Comparison of ASD-related gene sets and gene expression analysis of post-mortem ASD brains.*</u> Two prior analyses have identified gene co-expression WGCNA modules that are differentially expressed in the brain in ASD cases compared with controls. The microarray analysis by Voineagu et al.[9] identified a module enriched

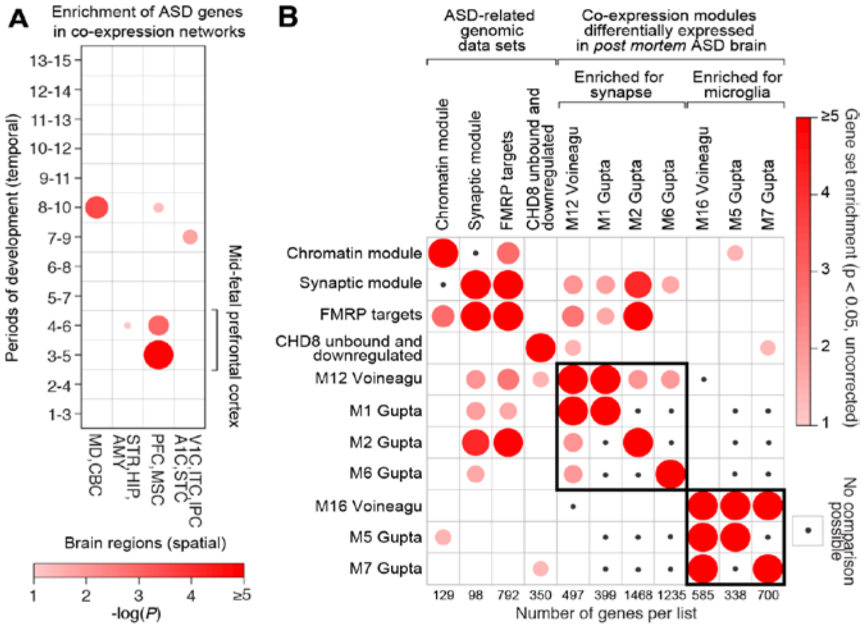


*Figure 7. Enrichment of ASD-associated genes in gene expression data. A. Spatiotemporal co-expression networks were formed around nine high confidence ASD genes for 4 groups of brain regions (x-axis) and 13 overlapping developmental periods (y-axis). The –log(P) value for enrichment with 126 low confidence ASD genes is shown by the size and shade of the circle. Strong enrichment is observed in the mid-fetal PFC and primary motor-sensory cortex (PFC-MSC). B. Four ASD related gene sets*[9,52,53] *compared to seven WGCNA co-expression modules that are differentially expression in post mortem ASD brains (right). Fold enrichment is indicated by the size and shade of the circle. A synaptic and microglial module are indicated by the black rectangles. Small black circles show gene sets that are non-overlapping by definition (e.g. WGCNA modules in the same analysis).*

for synaptic genes (M12) that overlaps with three modules (M1, M2, and M6) identified using RNA-seq in Gupta et al[60] (**Figure 7B**). Similarly, one module enriched for microglial genes (M16) was observed in the Voineagu et al.[9] paper and this overlaps with two modules (M5 and M7) identified in the Gupta analysis (**Figure 7B**). We compared these seven WGCNA modules with four sets of ASD-related genes: the chromatin and synaptic modules from our recent analysis of exome and CNV data (**Figure 6A**)[17], RNA targets of the fragile X protein FMRP[61] that are strongly enriched within ASD-associated genes[26], and genes that are downregulated in CHD8 knockdown but not bound by CHD8 on ChIP-Seq analysis that have been described as targeting synaptic genes associated with ASD[62]. The synaptic module and FMRP targets are strongly enriched through the synaptic WGCNA modules suggesting these modules may represent causal factors that persist in the ASD brain. Further analysis is required to determine if these modules are causal or simply a consequence of ASD.

***Experimental design and methods:*** In ***subaim 3.1****, we will increase the spatiotemporal resolution of co-expression analysis of ASD neurobiology*. Our prior analysis of spatiotemporal convergence, described in detail under preliminary data[15], was based on 57 neurotypical brains, 9 high confidence ASD genes (FDR ≤0.05), and 126 low confidence ASD genes (FDR ≤0.3)[17]. These data enabled us to examine 4 groups of brain regions spanning multiple developmental periods (**Figure 7A**). The data from Aim 1 and our progress in ASD gene discovery will allow us to perform this analysis using 87 neurotypical brains, 28 high confidence ASD genes (FDR ≤0.01), and 151 low confidence ASD genes (FDR ≤0.3). As before (**Figure 7A**), the gene expression samples will be divided into spatiotemporal windows using hierarchical clustering to group related brain regions (spatial) and considering overlapping developmental windows (temporal). In each spatiotemporal window we will identify the top 20 co-expressed genes around 28 high confidence ASD genes and, following the logic that a spatiotemporal network relevant to ASD should be enriched for other ASD genes, we will assess the enrichment of the 151 low confidence ASD genes (FDR ≤0.3). The expanded number of brain samples will enable us to use small subdivisions of brain regions and developmental time regions to increase the resolution of the analysis, for example windows spanning one or two developmental periods. In addition, the larger list of high confidence ASD genes will allow us to perform the analysis by building the spatiotemporal networks around subsets of these 28 genes and improve the accuracy of the analysis through cross validation. In addition, we will divide the 28 high confidence genes by the two main functional categories observed, specifically chromatin regulators and synaptic genes, to assess the spatiotemporal dynamics of each functional category separately. The outcome of this aim will be refined gene co-expression networks that show spatial and temporal convergence among ASD risk genes.
***Pitfalls and alternatives:*** The analytical methods described here have been applied to the BrainSpan data using 9 high confidence genes resulting in the discovery of spatiotemporal convergence in the frontal cortex of the mid-fetal brain. This finding has been replicated using complementary methods[58,59]. In this aim we will be increasing the resolution through the inclusion of additional gene expression data and novel ASD-associated genes[17], therefore we do not foresee complications. An alternative 'top down' methodology such as WGCNA, in which co-expression modules are generated from the complete dataset and are then assessed for enrichment

of ASD genes, has yielded similar findings[31]. We will also apply this complementary WGCNA method across spatiotemporal windows.

In ***subaim 3.2**, we will identify ASD-associated non-coding de novo mutations in regulatory loci*. Under pre-existing funding arrangements we will have access to whole-genome sequencing (WGS) data for 5,120 individuals from 1,280 quartet families composed of two parents, an affected child, and an unaffected sibling control. We have previously reported an increased burden of *de novo* mutations between the affected and unaffected siblings[17] and we have observed this for *de novo* CNVs in microarray data and *de novo* loss of function mutations in exome data. To identify functional non-coding *de novo* mutations in regulatory loci, we will leverage the integrated RNA-Seq, ChIP-Seq, and HiC data from Aims 1 and 2 with the *de novo* mutation identification approach described in our preliminary data (**Figure 6**). To maximize our ability to discover compartments of the genome that carry risk we will assess *de novo* burden in three sets of loci: **1**) All regulatory loci identified in neurotypical brain divided by function (e.g. promoter, 3`UTR); **2**) Regulatory loci identified in neurotypical brain with a relationship to 28 high-confidence ASD genes; and **3**) Regulatory loci identified in neurotypical brain with a relationship to the points of convergence for ASD genes identified in Subaim 3.1 such as prefrontal cortex in mid-fetal development. The outcome of this aim will be non-coding mutations and regulatory loci that show association with ASD.
***Pitfalls and alternatives:*** Our methods for identifying *de novo* mutations in whole genome sequencing data are well developed and we have demonstrated a >95% confirmation rate for the mutations predicated. Additionally, our preliminary data, based on 40 families, shows evidence of ASD association for *de novo* mutations within enhancers active in human dlPFC (**Figure 6B and C**). This suggests the proposed study of 1,280 families will offer sufficient power even if the overall contribution of *de novo* mutations in the non-coding genome to ASD etiology is relatively weak. To maximize our chance of identifying ASD associated non-coding variants we will assess only the loci with the strongest evidence of functional activity, including the larger mutations, such as indels, that may carry the greatest risk. Concurrently, Dr. Sanders has an established collaboration with Mike Talkowski and the GATK CNV/SV working group to develop methods that maximize our sensitivity for detecting indels and small CNVs in whole genome sequence data.

In ***subaim 3.3**, we will identify points of spatiotemporal convergence using ASD associated non-coding mutations:* Non-coding elements such as enhancers frequently show a degree of specificity to particular developmental time points, brain regions, or cell types[63]. We will use the ASD-associated non-coding *de novo* mutations in regulatory loci and regulatory loci related to ASD associated genes to assess which integrated regulatory networks from Aim 2 show the greatest enrichment for these non-coding mutations. By considering the brain regions and developmental epochs in which these networks exist we will assess points of spatiotemporal convergence critical to ASD. The outcome of this aim will be an independent analysis of points of spatiotemporal convergence in ASD based on non-coding mutations and regulatory loci.
***Pitfalls and alternatives:*** This aim relies on the discovery of specific ASD-associated regulatory loci through the discovery of numerous *de novo* mutations in cases. Due to the small size of regulatory regions we may not see this clustering in a single regulatory element. Should this be the case we will use genomic annotation to rank the regulatory loci with a single mutation, for example considering conservation, constraint [64], and large mutations such as indels that are more likely to disrupt the element (**Figure 6C**).

In ***subaim 3.4**, we will assess regulatory networks that are observed in the post mortem ASD brain.* Aims 3.1 to 3.3 focus on neurotypical brains and their association with ASD-associated mutations. In this aim we will assess the enrichment of ASD-associated genes, non-coding mutations and regulatory networks that differ between *post mortem* ASD and neurotypical brains (**Figure 6**). Because genetic variants associated with ASD precede the onset of ASD symptoms, enrichment for these mutations will suggest that such networks are causal (**Figure 7**) to the ASD phenotype. Conversely, a lack of enrichment for these mutations in ASD-relevant networks will suggest the network is consequential to ASD. The outcome of this aim will therefore be to distinguish ASD-specific regulatory networks that are likely to be causal from those that may be consequential.
***Pitfalls and alternatives:*** Methods for assessing such enrichment are well established and we already have a large list of ASD-associated genes; we foresee no complications in performing this aim. The main challenge lies in the interpretation of a regulatory network that does not enrich for ASD-associated genes (e.g. microglia in existing *post mortem* analyses, **Figure 7B**), since this may indicate a non-causal relationship or reflect and incomplete list of ASD-associated genes. We will therefore focus on networks with positive enrichment for these genes and acknowledge the complexities of interpreting a negative result.

**TIMELINE AND MILESTONES SECTION** See Other Attachments

## BIBLIOGRAPHY

1. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature reviews. Genetics* **13**, 537-551 (2012). pmcid: 4110909.
2. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics* **45**, 984-994 (2013). pmcid: 3800159.
3. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. *Science* **349**, 1489-1494 (2015). pmcid:
4. Krystal, J. H. & State, M. W. Psychiatric disorders: diagnosis to therapy. *Cell* **157**, 201-214 (2014). pmcid: 4104191.
5. Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nature neuroscience* **17**, 764-772 (2014). pmcid: 4077789.
6. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of *de novo* mutations in the genetics of autism spectrum disorders. *Nature reviews. Genetics* **15**, 133-141 (2014). pmcid:
7. Walsh, C. A., Morrow, E. M. & Rubenstein, J. L. Autism and brain development. *Cell* **135**, 396-400 (2008). pmcid: 2701104.
8. Huguet, G., Ey, E. & Bourgeron, T. The genetic landscapes of autism spectrum disorders. *Annu Rev Genomics Hum Genet* **14**, 191-213 (2013). pmcid:
9. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384 (2011). pmcid: 3607626.
10. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature reviews. Genetics* **16**, 441-458 (2015). pmcid:
11. Abelson, J. F. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317-320 (2005). pmcid:
12. Johnson, M. B. *et al.* Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**, 494-509 (2009). pmcid: 2739738.
13. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012). pmcid: 3667984.
14. State, M. W. & Sestan, N. Neuroscience. The emerging biology of autism spectrum disorders. *Science* **337**, 1301-1303 (2012). pmcid: 3657753.
15. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013). pmcid: 3995413.
16. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat Commun* **6**, 6404 (2015). pmcid: 4355952.
17. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233 (2015). pmcid: 4624267.
18. Luo, R. *et al.* Genome-wide transcriptome profiling reveals the functional impact of rare *de novo* and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet* **91**, 38-55 (2012). pmcid: 3397271.
19. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* **14**, 1109-1120 (2015). pmcid:
20. Preuss, T. M. Human brain evolution: from gene discovery to phenotype discovery. *Proc Natl Acad Sci U S A* **109 Suppl 1**, 10709-10716 (2012). pmcid: 3386880.
21. Lui, J. H. *et al.* Radial glia require PDGFD-PDGFRbeta signalling in human but not mouse neocortex. *Nature* **515**, 264-268 (2014). pmcid: 4231536.
22. Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647 (2013). pmcid: 3922239.
23. Zeng, J. *et al.* Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* **91**, 455-465 (2012). pmcid: 3511995.
24. Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213-217 (2009). pmcid: 2778075.
25. The PsychENCODE Consortium *et al.* The PsychENCODE project. *Nature neuroscience* **Manuscript accepted** (2015). pmcid:

26. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014). pmcid:
27. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246-250 (2012). pmcid: 3350576.
28. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012). pmcid: 3613847.
29. Sanders, S. J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011). pmcid: 3939065.
30. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622 (2012). pmcid: 3528801.
31. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021 (2013). pmcid:
32. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012). pmcid: 3356448.
33. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009). pmcid: 2858594.
34. Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680 (2014). pmcid:
35. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518-529 (2013). pmcid: 3894107.
36. Network & Pathway Analysis Subgroup of Psychiatric Genomics, C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209 (2015). pmcid: 4378867.
37. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523 (2014). pmcid: 4390143.
38. Parikshak, N. N. *et al.* Global changes in patterning, splicing and primate specific lncRNAs in autism brain. *Manuscript in preprint. Please see appnedix.* (2015). pmcid:
39. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010). pmcid: 3003124.
40. Won, H. *et al.* Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development. *Manuscript in preprint. Please see appnedix.* (2015). pmcid:
41. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003 (2012). pmcid: 3816492.
42. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* **28**, 1045-1048 (2010). pmcid: 3607281.
43. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nature genetics* **45**, 723-729 (2013). pmcid: 3932982.
44. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428 (2014). pmcid: 4208299.
45. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014). pmcid:
46. McCarthy, S. E. *et al. De novo* mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658 (2014). pmcid: 4031262.
47. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011). pmcid: 3566780.
48. Sarnat, H. B., Nochlin, D. & Born, D. E. Neuronal nuclear antigen (NeuN): a marker of neuronal maturation in early human fetal nervous system. *Brain Dev* **20**, 88-94 (1998). pmcid:
49. Pletikos, M. *et al.* Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron* **81**, 321-332 (2014). pmcid: 3931000.
50. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206 (2014). pmcid: 4105188.
51. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013). pmcid: 3869051.
52. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009). pmcid: 2672628.

53. Habegger, L. *et al.* RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**, 281-283 (2011). pmcid: 3018817.
54. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015). pmcid:
55. del Rosario, R. C. *et al.* Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nature methods* **12**, 458-464 (2015). pmcid:
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009). pmcid: 2705234.
57. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010). pmcid: 2928508.
58. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J Neurosci* **34**, 1420-1431 (2014). pmcid: 3898298.
59. Chang, J., Gilman, S. R., Chiang, A. H., Sanders, S. J. & Vitkup, D. Genotype to phenotype relationships in autism spectrum disorders. *Nature neuroscience* **18**, 191-198 (2015). pmcid: 4397214.
60. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* **5**, 5748 (2014). pmcid: 4270294.
61. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261 (2011). pmcid: 3232425.
62. Sugathan, A. *et al.* CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc Natl Acad Sci U S A* **111**, E4468-4477 (2014). pmcid: 4210312.
63. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521-1531 (2013). pmcid: 3989111.
64. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* **9**, e1003709 (2013). pmcid: 3749936.

## BIBLIOGRAPHY

1. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature reviews. Genetics* **13**, 537-551 (2012). pmcid: 4110909.
2. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics* **45**, 984-994 (2013). pmcid: 3800159.
3. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. *Science* **349**, 1489-1494 (2015). pmcid:
4. Krystal, J. H. & State, M. W. Psychiatric disorders: diagnosis to therapy. *Cell* **157**, 201-214 (2014). pmcid: 4104191.
5. Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nature neuroscience* **17**, 764-772 (2014). pmcid: 4077789.
6. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of *de novo* mutations in the genetics of autism spectrum disorders. *Nature reviews. Genetics* **15**, 133-141 (2014). pmcid:
7. Walsh, C. A., Morrow, E. M. & Rubenstein, J. L. Autism and brain development. *Cell* **135**, 396-400 (2008). pmcid: 2701104.
8. Huguet, G., Ey, E. & Bourgeron, T. The genetic landscapes of autism spectrum disorders. *Annu Rev Genomics Hum Genet* **14**, 191-213 (2013). pmcid:
9. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384 (2011). pmcid: 3607626.
10. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature reviews. Genetics* **16**, 441-458 (2015). pmcid:
11. Abelson, J. F. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317-320 (2005). pmcid:
12. Johnson, M. B. *et al.* Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**, 494-509 (2009). pmcid: 2739738.
13. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012). pmcid: 3667984.
14. State, M. W. & Sestan, N. Neuroscience. The emerging biology of autism spectrum disorders. *Science* **337**, 1301-1303 (2012). pmcid: 3657753.
15. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013). pmcid: 3995413.
16. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat Commun* **6**, 6404 (2015). pmcid: 4355952.
17. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233 (2015). pmcid: 4624267.
18. Luo, R. *et al.* Genome-wide transcriptome profiling reveals the functional impact of rare *de novo* and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet* **91**, 38-55 (2012). pmcid: 3397271.
19. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* **14**, 1109-1120 (2015). pmcid:
20. Preuss, T. M. Human brain evolution: from gene discovery to phenotype discovery. *Proc Natl Acad Sci U S A* **109 Suppl 1**, 10709-10716 (2012). pmcid: 3386880.
21. Lui, J. H. *et al.* Radial glia require PDGFD-PDGFRbeta signalling in human but not mouse neocortex. *Nature* **515**, 264-268 (2014). pmcid: 4231536.
22. Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647 (2013). pmcid: 3922239.
23. Zeng, J. *et al.* Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* **91**, 455-465 (2012). pmcid: 3511995.
24. Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213-217 (2009). pmcid: 2778075.
25. The PsychENCODE Consortium *et al.* The PsychENCODE project. *Nature neuroscience* **Manuscript accepted** (2015). pmcid:

26. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014). pmcid:
27. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246-250 (2012). pmcid: 3350576.
28. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012). pmcid: 3613847.
29. Sanders, S. J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011). pmcid: 3939065.
30. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622 (2012). pmcid: 3528801.
31. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021 (2013). pmcid:
32. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012). pmcid: 3356448.
33. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009). pmcid: 2858594.
34. Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680 (2014). pmcid:
35. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518-529 (2013). pmcid: 3894107.
36. Network & Pathway Analysis Subgroup of Psychiatric Genomics, C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209 (2015). pmcid: 4378867.
37. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523 (2014). pmcid: 4390143.
38. Parikshak, N. N. *et al.* Global changes in patterning, splicing and primate specific lncRNAs in autism brain. *Manuscript in preprint. Please see appnedix.* (2015). pmcid:
39. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010). pmcid: 3003124.
40. Won, H. *et al.* Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development. *Manuscript in preprint. Please see appnedix.* (2015). pmcid:
41. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003 (2012). pmcid: 3816492.
42. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* **28**, 1045-1048 (2010). pmcid: 3607281.
43. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nature genetics* **45**, 723-729 (2013). pmcid: 3932982.
44. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428 (2014). pmcid: 4208299.
45. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014). pmcid:
46. McCarthy, S. E. *et al.* *De novo* mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658 (2014). pmcid: 4031262.
47. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011). pmcid: 3566780.
48. Sarnat, H. B., Nochlin, D. & Born, D. E. Neuronal nuclear antigen (NeuN): a marker of neuronal maturation in early human fetal nervous system. *Brain Dev* **20**, 88-94 (1998). pmcid:
49. Pletikos, M. *et al.* Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron* **81**, 321-332 (2014). pmcid: 3931000.
50. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206 (2014). pmcid: 4105188.
51. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013). pmcid: 3869051.
52. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009). pmcid: 2672628.

53. Habegger, L. *et al.* RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**, 281-283 (2011). pmcid: 3018817.

54. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015). pmcid:

55. del Rosario, R. C. *et al.* Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nature methods* **12**, 458-464 (2015). pmcid:

56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009). pmcid: 2705234.

57. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010). pmcid: 2928508.

58. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J Neurosci* **34**, 1420-1431 (2014). pmcid: 3898298.

59. Chang, J., Gilman, S. R., Chiang, A. H., Sanders, S. J. & Vitkup, D. Genotype to phenotype relationships in autism spectrum disorders. *Nature neuroscience* **18**, 191-198 (2015). pmcid: 4397214.

60. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* **5**, 5748 (2014). pmcid: 4270294.

61. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261 (2011). pmcid: 3232425.

62. Sugathan, A. *et al.* CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc Natl Acad Sci U S A* **111**, E4468-4477 (2014). pmcid: 4210312.

63. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521-1531 (2013). pmcid: 3989111.

64. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* **9**, e1003709 (2013). pmcid: 3749936.

## RESEARCH STRATEGY

## SIGNIFICANCE

Neuropsychiatric disorders such as autism spectrum disorder (ASD), bipolar disorder (BD), and schizophrenia (SCZ) are complex and devastating illnesses with considerable morbidity and mortality, as well as high personal and societal costs. Many of them are also polygenic, with multiple variants, both rare and common, spread throughout the genome influencing the disease risk[3]. Recent studies have identified rare variants contributing to psychiatric disorders that are enriched in genes involved in global gene regulation and chromatin modification, and many common risk variants are enriched in regulatory regions of the human genome, regions whose functions are poorly understood. The interpretations of these variations in regulatory regions will certainly be improved with better maps of RNA transcripts, regulatory elements, and chromatin states in the human brain. The age of onset and progression of major psychiatric disorders also varies (**Figure 1**) necessitating the study of the temporal dynamics of gene regulation during human brain development and recognizing the developmental context of psychiatric disorders. An emerging body of research indicates that many aspects of the development and physiology of the human brain are not well recapitulated in model organisms[20-24] **and therefore it is increasingly apparent that psychiatric disorders need to be understood in the broader context of human brain development and physiology.**

In recent years, considerable effort has been made by many studies, including large-scale efforts by ENCODE, NIH Roadmap (REMC) and GTEx projects to survey the diversity of *cis*-acting regulatory regions and RNA species of the human genome across different tissues and time points. However, a comprehensive catalog of transcripts, regulatory elements, epigenetic modifications, and chromatin structure from the human brain during development and in distinct brain regions and cell types is lacking. The PsychENCODE (phase 1) projects have initiated these efforts.

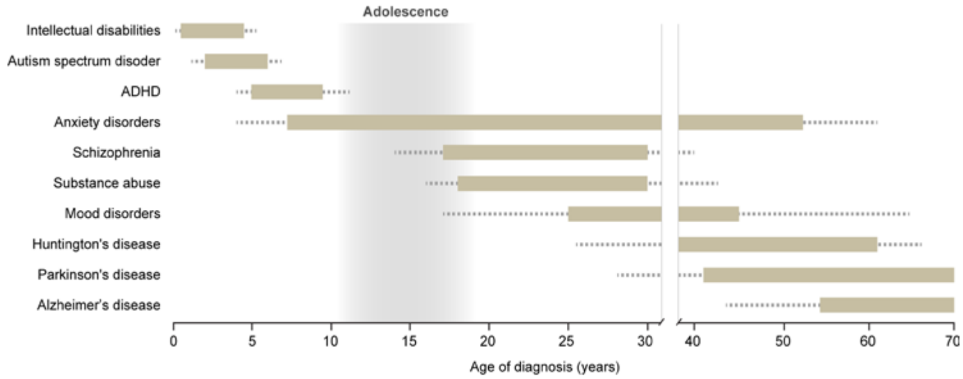


**Figure 1. Psychiatric and neurological disorders have discrete ages of onset.** *The bars indicate the age range that each disorder commonly affects, with less frequent ages of diagnosis denoted as dotted lines. This variability is indicative of dysregulation of tightly controlled developmental processes and highlights the necessity of defining the spatio-temporal molecular processes in human brain.*

**PsychENCODE consortium projects.** The key goals of the PsychENCODE project are to provide an enhanced framework of regulatory elements, catalog epigenetic modifications, and quantitate coding and non-coding RNA and protein expression in a tissue- and cell-type specific manner from neurotypical (healthy) control brains and diseased post-mortem human brains[25]. These efforts will be complemented with integrative analyses, as well as with functional characterizations of disease-associated genomic elements using human neural cell systems or the developing mouse brain. However, the human brain is heterogeneous cellularly and its development is regionally asynchronous and prolonged. *To overcome issues that hamper the potential benefits of initial psychENCODE studies, we will apply several approaches to address regional and cellular heterogeneity, prolonged development, and new genomic methods in the context of brain development and ASD.*

Here we focus on **neurotypical (control) brain and ASD**, which is a complex developmental syndrome with a significant genetic contribution. Although considerable genetic and phenotypic heterogeneity has complicated efforts to establish the biological substrates of the syndrome, the emergence of reliable genetic findings has started to shed light on potential pathogenic mechanisms, providing an extraordinary opportunity for developing a mechanistic understanding of the disorder. Recent studies suggest that **over 500 rare, *de novo* mutations contribute to ASD risk and no single genetic mutation accounts for more than 1% of ASD cases**[13,17,26-30], consistent with significant heterogeneity in this, and other neuropsychiatric disorders[3]. Despite this heterogeneity, mapping **ASD risk genes onto co-expression networks that represent normal human brain development has revealed that ASD genes coalesce in modules related to chromatin remodeling and transcriptional regulation during early fetal brain development, suggesting potential convergent pathways in the disorder**[9,15,27,31]. Another remarkable finding that parallels the convergence of genetic findings in developmental pathways is the identification and validation of shared transcriptional changes in postmortem brain in ASD[9]. This transcriptional dysregulation, coupled with the evidence that large effect size *de novo* ASD

risk genes are highly enriched in chromatin modifying genes (many of which are expressed in early fetal brain development), emphasizes the importance of understanding the nature and extent of chromatin disorganization in ASD brain and in normal brain development. **Further, since these data suggest distinct neuronal and glial gene dysregulation, it is crucial to delineate the profiles of these major cell types.** In addition to our ongoing efforts in PsychENCODE phase I project, this proposal provides critical advances in our understanding of the role(s) of non-coding functional elements in the pathophysiology of ASD and a scaffold for understanding chromatin structure and gene regulation across normal brain development. Overall, the approach proposed here will provide mechanistic insights that connect distinct transcriptional programs associated with ASD pathogenesis, and will provide a resource of the mechanisms of gene regulation across brain development to inform other neuropsychiatric disorders, a key goal of psychENCODE. **This work also leverages psychENCODE phase 1 projects by adding significant new data to expand the value of the resource and by directly addressing key areas of interest in control and ASD brains as outlined in RFA-MH-16-230**: ***1)*** *Generation of comprehensive, high resolution human brain region/cell type and age-specific maps of different classes of RNA transcripts, regulatory elements, chromatin states, chromatin conformation, and chromatin interactions;* ***2)*** *Identification of human brain region/cell type and age-specific molecular processes;* ***3)*** *Integration of these newly generated multi-omic datasets, from diseased and healthy control brains, with large-scale genomic resources;* ***4)*** *Generation and analysis of high-depth, whole genome sequencing data to allow for improved evaluation of various genetic alterations; and* ***5)*** *Development of comprehensive molecular models of disease (i.e., ASD) using systems biology approaches.*

## INNOVATION

This proposal is innovative in several aspects. First, to the best of our knowledge, the systematic discovery and functional characterization of genomic non-coding elements and 3D chromatin architecture has not been performed in healthy developing human brains or ASD brains at a cell type-specific resolution. For example, we use Hi-C, which combines chromosome conformation capture and NextGen sequencing to identify physical interactions that capture multiple levels of chromosome architecture ranging from nuclear configuration (“compartments” of about 5Mb) to TADs (domains of 500kb on average) and gene loops (often reflect enhancer promotor relationships; 40kb average), and is the only such method that spans all of these levels, genome-wide[32-34]. Second, this project will conduct direct analysis of one of the largest collection of well-characterized high quality healthy as well as syndromic and idiopathic ASD postmortem brains. Third, we will combine fluorescence-activated nuclei sorting (FANS) with advanced genomic techniques to analyze multiple genomic features in archived development control and ASD brains. Fourth, we will leverage these analyses with our ongoing psychENCODE phase 1 tissue level analyses and other recent large-scale genomic resources, such as BrainSpan, ENCODE, GTEx and Roadmap project. Therefore, our proposed data and integrated analyses has potential to improve our understanding of genomic processes and normal human brain development as well as diagnostics, neurobiology and treatment of ASD.

## COLLABORATION

This collaboration brings together multiple groups with long standing expertise in developmental neurobiology, psychiatry, human biobanking, genetics and genomics, statistics, bioinformatics, and systems biology that have worked closely with one another for almost a decade as evidenced by many co-publications. Several key conceptual threads have been apparent in our work together related to human brain development and neuropsychiatric disorders: **1**) Revealed new insights into human neurodevelopment through functional genomic profiling of postmortem tissue and cell culture models[12-16]; **2**) Assessed rare and *de novo* mutations for ASD association[13,17,18]; **3**) Identified the neural processes and pathways that are altered in the presence of ASD-associated mutations, as well as when and where these processes and pathways occur in the developing human brain[15,17,19]. In addition, M. Gerstein (Yale) and Z. Weng (University of Massachusetts), experts in bioinformatics and computational biology, are leaders of the PsychENCODE DAC, which will normalize the data to remove batch effects, establish uniform data processing pipelines and build calibration resources for all assays to enable comparison and integration of the data generated by all psychENCODE groups. The efforts of each group will be tightly integrated in order to communicate progress and results, design and implement analytical tools, and transfer data. Given the complexity of human neurodevelopment and genetics/neurobiology of ASD, we believe that integrating the respective expertise of these groups, and their respective collaborators at UCLA (Ernst and Geschwind), UCSF (Sanders, State and Willsey), UMass (Weng), and Yale (Gerstein and Sestan), offers the best opportunity to better understand human brain development and ASD through functional genomics. Here, we propose to leverage our expertise and continue this highly productive collaboration and expand psychENCODE phase 1.

**ELEMENTS UNIQUE TO THIS SITE** (UCSF; State, PI; Sanders and Willsey, co-investigators)
The UCSF team will combine the data from Aims 1 and 2 with whole-genome sequencing (WGS) data for 5,120 individuals in 1,280 ASD families to identify when, where, and in which cells the etiology of ASD occurs. In *Subaim 3.1* they will use the additional gene expression data and ASD gene discovery to increase the resolution of their prior spatiotemporal analysis that implicated prefrontal cortex during mid-fetal development. In *Subaim 3.2* they will use regulatory loci from Aims 1 and 2 to filter de novo mutations in non-coding regions in the WGS data. In *Subaim 3.3* they will use the non-coding mutations from Subaim 3.2 and the integrated analysis of regulatory networks from Aim 2 to perform an independent spatiotemporal analysis of when, where, and in which cells ASD etiology occurs. Finally in *Subaim 3.4* they will assess the enrichment of coding and non-coding mutations from WGS in the integrated regulatory, transcriptional and molecular networks in ASD brains to provide evidence for these networks being a cause rather than a consequence of ASD.

## APPROACH

The objective of this proposal is to extend our ongoing tissue level analyses of healthy and ASD brains under the psychENCODE consortium with the inclusion of additional genomic methods, brain regions, developmental time points, and cell-type specific analyses. By performing three integrated aims (**Figure 2**) we propose to enhance this public resource and improve our understanding of the molecular processes underlying normal human neurodevelopment and ASD.

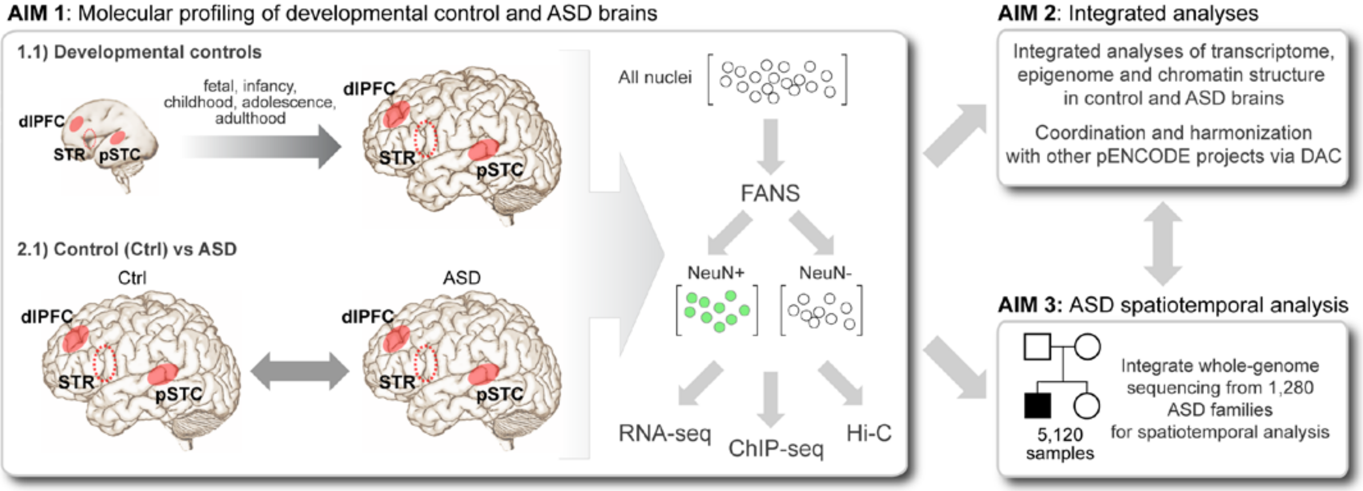

AIM 1: Molecular profiling of developmental control and ASD brains
1.1) Developmental controls
dlPFC
STR
pSTC
fetal, infancy, childhood, adolescence, adulthood
dlPFC
STR
pSTC
2.1) Control (Ctrl) vs ASD
Ctrl
ASD
dlPFC
STR
pSTC
dlPFC
STR
pSTC
All nuclei
FANS
NeuN+
NeuN-
RNA-seq
ChIP-seq
Hi-C
AIM 2: Integrated analyses
Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains
Coordination and harmonization with other pENCODE projects via DAC
AIM 3: ASD spatiotemporal analysis
5,120 samples
Integrate whole-genome sequencing from 1,280 ASD families for spatiotemporal analysis

**Figure 2. Schematic workflow of three specific aims.**

### Aim 1. Time, region and cell type-specific molecular profiling of control and ASD brains.

***Rationale and preliminary supporting data:*** Three major observations provide motivation for this aim. The *first* is the recognition that genomic data, including transcriptomic, epigenetic and physical chromatin structure, from the relevant neurotypical tissue (control), spanning the key epochs of neurodevelopment and function from fetal to adult periods, provide a new and previously unobtainable view of genetic risk for psychiatric disease[10,15,16,31,35,36]. The *second* is that brain is comprised of an extremely heterogeneous mixture of cell types that exhibit distinct molecular profiles, including glia-to-neuron ratios that could show considerable fluctuations across normal development or in certain disease states. The *third* is the observations of differences in transcriptome organization via tissue-level gene co-expression network analysis conducted between ASD and normal brains[9]. Thus, here **we propose to create a region and cell type-specific normal developmental scaffolding on which to frame disease variants via transcriptional (RNA-seq), epigenetic (ChIP-seq) and chromatin architecture (HiC) profiling of neuronal and non-neuronal cells at key epochs in human brain development (subaim 1.1), as well as compare these profiles in ASD and matched control brains (subaim 1.2) to help elucidate the mechanisms by which genetic variation alters brain development and function, leading to ASD and related neuropsychiatric conditions.** While several genomic features are currently being analyzed in control and ASD brains by our and other groups in the psychENCODE consortium, cellular heterogeneity during development, other genomic features (e.g. 3D chromatin contacts), have yet to be addressed. To address these issues, we will utilize our large, high quality, phenotypically well-characterized human brain collection (see Facilities and Resources section), as well as newly implemented methods to collect molecularly defined cell type specific nuclei from archival human postmortem brains in this collection.

**Our preliminary data demonstrates a clear pattern of transcriptional dysregulation is observed in 2/3 of ASD brains[9], which we have now confirmed in our psychENCODE phase 1 projects (in a more than**

**double sized sample of cases and controls)** using tissue level RNA-seq and ChIP-seq (H3K4me3 and H3K27ac) in multiple brain regions in 43 idiopathic ASD cases, 8 cases with chromosome 15q11-13 duplication syndrome (dup15q) and ASD, and 63 controls[37,38]. We also observe that post mortem brain from patients with ASD caused by (dup)15q11-13 share this same pattern at all levels of differential protein coding gene expression, splicing and lncRNA[38]. As the first step in exploring potential mechanisms, we performed epigenetic profiling of ASD vs. control brains with H3K27ac marks, which indicate active enhancers[39]. Genes with differential H3K27ac peaks in their promoter regions (5000bp upstream of the transcription start site) were enriched with neuronal genes with changes in expression. This result demonstrates that transcriptional dysregulation in ASD is partially mediated by changes in histone/chromatin modifications. Furthermore, the two major groupings of modules derived from whole tissue gene expression analysis sort into those up-regulated and expressed in glia, and those down-regulated in neurons[9,38], strongly motivating our plan for transcriptional, epigenetic and Hi-C profiling in neurons and non-neuronal (glial) cells independently.

**Another of the important advances in methodology that will be implemented here is the assessment of 3-D chromatin structure across to different brain regions and cell types, and 5 key epochs of normal brain development and in ASD brains.** Our preliminary data strongly supports the value of these data and our ability to perform and analyze these experiments (see also[40]). We established an efficient Hi-C protocol and obtained high resolution data (10 kb resolution; via deep sequencing) from the fetal cortex from 3 individuals dissected into two zones: cortical plate (CP) and germinal zones (GZ) at post-conception week (PCW) 18 (total n = 12 samples: representative heatmap shown in **Figure 3A**). Demonstrating the data quality, principal component of the interchromosomal interaction matrix for GZ shows a high correlation with GC content ($r$=0.791, $P$<$10^{-256}$), gene number ($r$=0.384, $P$<$10^{-256}$), DNase I hypersensitivity ($r$=0.416, $P$<$10^{-256}$), and to a lesser extent, gene expression ($r$=0.129, $P$=1.11x$10^{-66}$; **Figure 3B and C**), recapitulating previous work in cell lines[41]. We next asked how chromatin interactions elicit transcriptional co-regulation. We hypothesized that highly interacting chromatin regions would be co-regulated at least in part by sharing chromatin remodelers and transcription factors (TFs). To test this, we binned chromatin interactions into top and bottom percentiles, and compared the distribution of correlation patterns for genes in the high and low interacting regions of chromatin. We observed that the high interacting regions were significantly biased toward positive correlations (**Figure 4A**), supporting the hypothesis that co-localization can predict co-expression.

We next integrated these data with the epigenomics map from the NIH Roadmap project[42]. By comparing the epigenetic mark combination matrix with the Hi-C contact matrix, we demonstrate that interacting regions exhibit shared epigenetic patterns: loci associated with transcriptional regulation and enhancers are significantly more likely to interact with each other (**Figure 4B**). Comparison of TF binding site (TFBS) combination matrix (generated from TFBS map reported in[43]) with the intrachromosomal contact matrix revealed distinct combinatorial patterns of TF binding likely to mediate chromosome interactions (**Figure 4C**), thus revealing new experimentally testable regulatory relationships.

To validate that Hi-C data can identify target genes regulated by single nucleotide polymorphisms (SNPs) in a general setting, we determined if SNPs with a significant effect on gene expression were also identified as

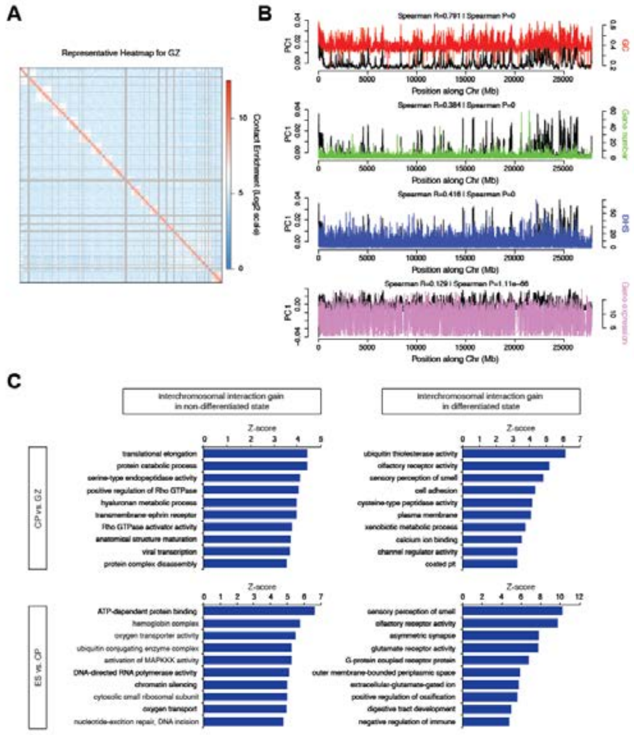


**Figure 3. Chromosome conformation in fetal brains (by Hi-C).** ***A.*** *Representative heatmap of chromosome contact matrix of GZ. Normalized contact frequency (Contact enrichment) is color-coded according to the legend on the right.* ***B.*** *Spearman correlation of PC1 of chromatin interaction profile of fetal brain (GZ) with GC content (GC), gene number, DNase I hypersensitivity (DHS), and gene expression level of fetal brains. These data show relationship of 3D structure to key known functional elements as has been previously shown in other systems.* ***C.*** *Gene ontology (GO) enrichment (GO Elite) of genes located in the top 5% of highly interacting inter-chromosomal regions specific to GZ vs. CP (top), and ES vs. CP (bottom), indicating that genes located on dynamic chromosomal regions are enriched for neuronal function in CP, which contains the more differentiated laminae. Please see Won et al. 2015 in Appendix for higher magnification figure.*

interacting by Hi-C using *cis*-expression quantitative trait loci (eQTL) data from adult frontal cortex[44]. Indeed, Hi-$C_{eQTL}$ genes were significantly over-represented with known associated genes from the eQTL study and eQTL SNP-transcript pairs exhibit significantly higher chromatin contact frequency than the null across all distance ranges measured, further supporting the utility of Hi-C to infer the gene or region of activity for regulatory variation. In addition we asked whether significant physical cis-chromosomal contacts identified with Hi-C could inform functional annotation of 108 genome-wide significant schizophrenia loci, most of which lie far outside known coding or other functional regions of the genome.

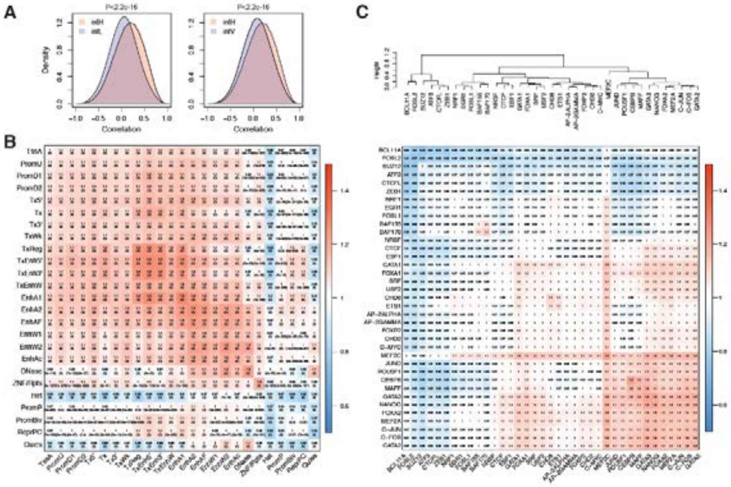


**Figure 4. Highly interacting regions share co-expression patterns, which is mediated by epigenetic regulation**. ***A.*** *The top 10,000 highest interacting regions (intH) in fetal brains both at GZ and CP show positive correlation in their gene expression patterns, while the top 10,000 lowest interacting regions (intL) and top 10,000 highly variant regions (intV) have no skew in the distribution, consistent with random interactions. P-value, Kolmogorov-Smirnov test.* ***B-C.*** *Epigenetic state combination (B) and TFBS combination (C) for intrachromosomal interacting regions. The epigenetic state matrix and TFBS combination matrix were generated by marking loci where two interacting chromosomal bins share epigenetic signature. For example, the epigenetic combination matrix between the active transcription start site (TssA) and active enhancers (EnhA1) is generated by marking where interacting loci have TssA and EnhA1. Intrachromosomal contact frequency map is compared to the epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Odds ratio (OR) and P-values are depicted in the heatmaps (Please see Won et al. 2015 in Appendix for higher magnification figure).*

Although SNPs are typically assigned to the closest genes, or those within the LD block, Hi-C indicated that about 50% of the variants were neither adjacent to the index SNPs (most-associated SNP within a locus), nor in LD. Interestingly, Hi-$C_{SCZ}$ genes significantly overlap with ASD *de novo* likely gene-disrupting (LGD) targets[26,45] (CP: OR=2.4, P=1.6x10$^{-5}$, GZ: OR=1.8, P=0.006), indicating a shared genetic etiology between ASD and schizophrenia[46]. The fact that genes with LGD mutations in ASD are associated with regulatory variants in schizophrenia suggests that complete abrogation of these genes may cause developmental defects as in ASD, while regulatory changes in these genes may cause later-onset of neuropsychiatric symptoms as in schizophrenia. *Collectively, these preliminary data demonstrate that we can conduct and analyze genome-wide Hi-C experiments, integrate these data with other epigenetic and transcriptomic data, and use chromatin architecture elucidated by Hi-C to provide novel genome-wide insights into the regulatory mechanisms occurring during neuronal differentiation and disease pathogenesis.*

***Experimental design and methods:*** In ***subaim 1.1.,*** *we will profile the transcriptome (by RNA-seq), cis-regulatory elements (ChIP-seq) and 3D chromatin architecture (Hi-C)* in the control neurotypical dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR). These regions have been implicated in the risk for ASD and schizophrenia[35] and in the cases of dlPFC and pSTC shown to have dysregulated transcriptional patterns in ASD[9]. Recent studies have also highlighted the late mid-fetal frontal cortex as most enriched for co-expression of ASD and schizophrenia *de novo* hits[15,31,35]. Brains from at least 5 key epochs of development representing mid-fetal, infancy, childhood, adolescence and adult brain, and a minimum of 6 subjects (balancing sex when possible) from each of these 5 epochs (30 brains in total) will be profiled.

Cell-type specific chromatin, epigenetic and transcriptome assays are at the core of this project. Mario Skarica, a talented research associate scientist in the Sestan lab, has developed a protocol to isolate high quality nuclei with preserved chromatin and RNA from archival fresh frozen fetal and postnatal human brains. Using this approach he has obtained, on average, 2.57 +-0.8 and 6.93+-3.3 million intact nuclei from 100 mg of the fetal or adult prefrontal gray matter (i.e., fetal CP or adult cortical layers 1 to 6 with a small part of underlying white matter), respectively (**Figure 5A**). Furthermore, we separated neuronal and non-neuronal nuclei, by immunostaining with the NeuN antisera against pan-neuronal splicing protein RBFOX3 (**Figure 5B and C**) and sorting on BD FACSAria IIU Three-Laser System. Starting with infancy and onwards, postnatal gray matter tissue corresponding to six-layered postnatal cortex and small part of adjacent white matter from dlPFC and pSTC, or STR (corresponding to the caudate-putamen with the internal capsule at the septal level) will be processed.

Tissue samples will be dissected directly from frozen tissue blocks using custom dental tools and protocol described in Kang et al., 2011[47]. These dissections will be performed by Nenad Sestan, who has over 2 decades of experience in human neuroanatomy and tissue processing and has microdissected over 1600 tissue samples for exon array profiling of the human brain transcriptome[47]. Given the high proportion of neurons in the cortical plate of the mid-fetal brain (approaching 95% or more), and relatively few neurons that are positive for NeuN at 17-20 PCW in neocortical CP or STR[48], we will not sort NeuN+ and NeuN- nuclei from mid-fetal brains, but instead analyze tissue homogenate and unsorted nuclei from CP of prospective dlPFC and pSTC as well as STR, separately, from corresponding neocortical and striatal GZ (i.e., VZ and SVZ) containing a mixed population of dividing neural stem/progenitor cells with a minor contribution of newborn neurons and glia.

Tissue samples will be pulverized and processed to release nuclei, which will be purified by ultracentrifugation and processed for RNA-seq in the case of mid-fetal samples or in the case of all postnatal specimens (infancy and onwards) sorted into a NeuN+ (predominantly neurons) and NeuN- (mostly glia) fractions. In the past year, we have obtained on average 23.45+-7.2 percentage of NeuN+ nuclei from PFC (**Figure 5C**). This approach will provide unbiased quantitative assessments of cell types in healthy and ASD brains. This approach allows us to simultaneously collect molecularly defined cell type-specific nuclei and isolate DNA, chromatin, and nuclear RNAs. Bulk tissue level RNA-seq is available for dlPFC, pSTC and STR in control and ASD brains as part of psychENCODE phase 1 studies[38], has already been added to enhance the scope of the resource. All brains necessary for this project are currently available in the Geschwind and Sestan labs (see Facilities and Resources section for the list).

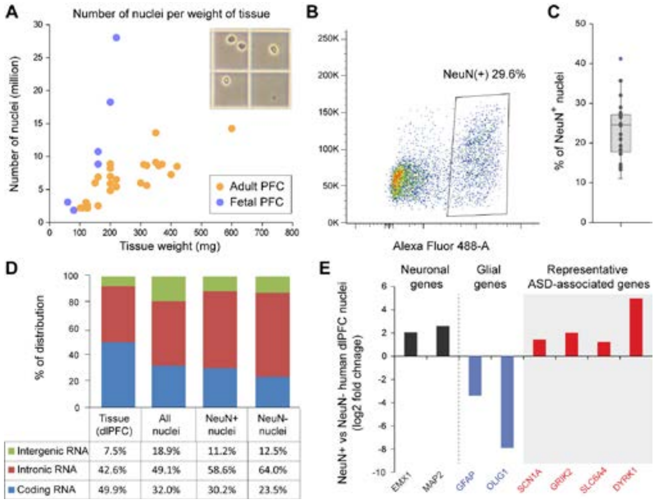


**Figure 5. Fluorescence-activated nuclei sorting (FANS) and nuclear RNA-seq of human dlPFC.** ***A.*** *Collection of single nuclei (see insert) from fetal (n=6) and adult (n=29) PFC.* ***B.*** *FANS plot for NeuN immunopositive nuclei.* ***C.*** *Percentage of NeuN+ nuclei collected across different experiments.* ***D.*** *Coverage for exon, intron and intergenic regions of different sequencing technologies.* ***E)*** *Differential expression comparison between NeuN+/- FANS nuclei for neuronal, glial and ASD-related genes.*

Total RNA will be extracted from 1 million nuclei using Norgen's Cytoplasmic & Nuclear RNA Purification Kit. RNA from tissue and cell populations will be depleted of rRNA and sequencing libraries prepared with TruSeq Stranded Total RNA with Ribo-Zero Gold and SMARTer Stranded RNA-Seq Kit, respectively. As expected, our preliminary nuclear RNA-seq analyses revealed higher percentage of unspliced primary transcripts and extensive identification of nuclear-retained long non-coding RNAs (**Figure 5D**). Importantly, we detected robust cell type-specific expression differences, including those of ASD-associated genes (**Figure 5 E)**. RNA-seq libraries will be sequenced on the Illumina HiSeq 2500 at the Yale Center for Genome Analysis (http://ycga.yale.edu/) to generate 100 bp strand specific paired-end sequence at over 40 million reads per end for each sample. For ChIP-seq, 1 million nuclei will be processed through our established protocol using well-characterized ChIP-grade H3K27ac and H3K4me3 antibodies that have been used in psychENCODE phase 1 tissue-level experiments. ChIP-seq libraries will be sequence at HiSeq 2500 at Yale at >40 million reads per sample. Using the standard pipelines developed in the Sestan and collaborating labs, we will perform QC analyses and compare the transcriptome and epigenetic data from different time points and regions to construct spatiotemporal gene and disease state profiles and co-expression networks using computational methods described in Aim 2.

For Hi-C, 2 million nuclei will be prepared from each sample and cross-linked in 1% formaldehyde for 10 min. Cross-linked DNA will then be restriction digested using HindIII, digested chromatin ends filled with biotin-14-dCTP, and resulting blunt-end fragments ligated under dilute conditions to minimize random intermolecular ligations. Following this, crosslinking will be reversed, unligated ends removed by exonuclease digestion (T4), DNA sheared by sonication, and 300-600bp fragments selected. The intermolecular ligation products containing biotin-tagged DNA will be pulled down with streptavidin beads and ligated with Illumina paired end adapters and the library sequenced by Illumina 50bp paired-end sequencing over 3 lanes of the HiSeq 25000 at UCLA, a depth necessary to facilitate sufficient hi-resolution analysis (300-500 million mapped reads), which can also be augmented by pooling samples to increase depth as needed.

In ***subaim 1.2**, complementary genomic analyses will be done on the FANS nuclei from control, and syndromic and idiopathic ASD brains***, to identify transcripts, regulatory elements, and 3D chromatin structures altered in ASD in brain region and cell type-specific manners. We will conduct RNA-seq, ChIP-seq and Hi-C on sorted neuronal and non-neuronal nuclei from 2 cortical regions, dlPFC and pSTC, and STR from 20 matched control and 20 ASD individuals, including 5 dup15q cases. We will select 10 ASD cases manifesting the shared pattern of transcriptional dysregulation observed, 10 without this pattern, and match them to controls to account for potential confounders (sex, age, postmortem interval [PMI], and RNA integrity numbers [RIN]). We will select 5 dup15q brains with most similar breakpoint structures. Hi-C will be performed on sorted nuclei using the identical experimental methods as in subaim 1.1.

***Pitfalls and alternatives:*** The techniques in these proposed experiments are commonly used in our laboratories and we do not expect complications. One potential issue is the obtainment of adequate samples. The Sestan lab has almost 200 high quality frozen human prenatal, early postnatal and adult brain specimens from clinically unremarkable (neurotypical) control donors. Control brains from this collection were used for different BrainSpan and psychENCODE phase 1 projects (see example studies[12,47,49,50] and Resources and Facilities section). Both Geschwind and Sestan labs have tissue samples from over 50 post mortem ASD cases and matched controls with good quality RNA, and have participated in the new initiative at the Simons Foundation to collect additional postmortem ASD brains. A related concern is whether the 20 ASD brains we propose to analyze are sufficient, given the heterogeneity typical of ASD, to detect robust differences between these samples and our controls. However, we were able to detect transcriptional dysregulation in 2/3rds of ASD brains in a smaller cohort17, and by directly comparing ASD brains exhibiting hallmarks of dysregulated transcription with those that do not, we expect to have sufficient statistical power to assess the extent to which 3D chromatin structure contributes to the observed transcriptional changes. Further, the use of 5 dup15q cases provides an additional homogeneous cohort, and as our preliminary results on transcriptome analysis of this cohort demonstrate (appendix), such sample sizes are sufficient. The main pitfall of Hi-C is that it averages chromosome contact population from millions of nuclei. Single-cell Hi-C can complement this limitation[51], but it can capture only one interaction for a given locus. Homogenous population of cells can be achieved by FANS and thus we propose this approach here. Additionally, Hi-C offers other benefits, including the ability to analyze interactions mediated by multiple TFs *en masse* in Hi-C, that are not easily achievable with other methods such as ChIA-PET. While our FANS approach, which follows standards accepted across the psychENCODE projects, is limited to two major groups of cells, we have been implementing the use of other cell type specific nuclear antibodies and single nuclear RNA-seq. Finally, we realize that other regions, including the thalamus, hypothalamus, and hippocampus, may be affected in ASD. We believe our work on the neocortex and STR will develop a framework for understanding of the molecular neuropathology of ASD which can then be extended to include other regions in the future.

**Aim 2. Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains.**

***Rationale:*** We will analyze the data generated in the previous aim to (1) identify *developmentally regulated and cell type specific changes* in the transcriptome, epigenome and the 3D chromatin structure (2) integrate the three types of datasets to gain comprehensive insights into the underlying mechanisms of transcriptional regulation and dysregulation in development and disease, respectively.

***Experimental design and methods***: In ***subaim 2.1**, several first order analyses will be done for quality control and to provide the data as a processed resource in addition to the raw data*. We will use Illumina CASAVA to purify the low-quality and non-identified reads and Fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), to report fundamental quality parameters. Next, Tophat[52] will be employed to uniquely align the filtered reads to their reference genome and RSEQtools[53] to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. The R package DESeq (http://bioconductor.org/packages/release/bioc/html/DESeq.html) will be used to identify differentially expressed (DEX) genes and well established methods including MATs to identify differential splicing[10,37]. DEX genes will be detected from the reliably expressed coding and non-coding transcripts, which are defined as transcripts with RPKM ≥ 1 in at least 2 samples of different developmental period. ChIP-seq reads will be aligned to the genome by Bowtie. After filtering of low score reads, we will use the MACS platform to call peaks enriched over the input library, and peaks with high empirical FDR will be excluded from further analysis. Thus, we will catalog all potential cis-regulatory elements from our genome-wide histone modification maps in all brain regions across developmental periods.

*For Hi-C analysis*, *hiclib* (https://bitbucket.org/mirnylab/hiclib) will be used to perform all initial analysis on Hi-C data from mapping to filtering and bias correction (see also[40]). Sequenced reads will be mapped to the human

genome by *Bowtie2* (with increased stringency, *--score-min -L 0.6,0.2--very-sensitive*) through iterative mapping and read pairs allocated to HindIII restriction enzyme fragments. Self-ligated and unligated fragments, fragments from repeated regions of the genome, PCR artifacts, and genome assembly errors will be removed. Filtered reads will be binned at 10kb, 40kb, and 100kb resolution to build a genome-wide contact matrix at a given bin size. This contact map depicts contact frequency between any two genomic loci. To decompose biases from the contact matrix and yield a true contact probability map, filtered bins are subjected to iterative correction[41]. Bias correction and normalization results in a corrected heatmap of bin-level resolution. 100kb resolution bins are assessed for inter-chromosomal interactions, 40kb for TAD analysis, and 10kb for gene loop detection. For TAD-level analysis[32], we will quantify the directionality index by calculating the degree of upstream or downstream (2Mb) interaction bias of a given bin, which will be processed by a hidden Markov model (HMM) to remove hidden directionality bias. For gene loop detection, aggregate peak analysis (APA) will be performed that quantifies the aggregate enrichment of putative peak sets by calculating the sum of a series of submatrices derived from a contact matrix[34]. Resulting inter- and intra-chromosomal interaction matrices as well as genome-wide TADs and gene loops will be used for integrative analysis.

*Developmental and cell type-specific changes*: Pearson's correlations between the first principal components (PC1) from different stages and neuronal and non-neuronal cell types, as well as with our own and other published data will be calculated to compare similarities between different cell types. We will explore alternative transcriptional mechanisms or post-transcriptional modifications occurring in normal (and ASD-affected, see below) regions/cells and time points. These can include up- or down-regulating expression, altered spatiotemporal gene expression, imbalanced expression of different alleles (allele-specific expression [ASE]), aberrant splicing events, modified RNA editing sites, fusion transcripts, or loss of function due to frameshift mutations. RNA and epigenome data will also be compared with tissue level psychENCODE phase 1 and BrainSpan's RNA-seq and ChIP-seq data. We will follow up with an analysis of the relative enrichment of each cell-specific marker genes in each subpopulation and use the expression profiles of these genes to guide the identification of an expanded set of cell-type specific markers.

*Integrated analyses*: Spearman's correlations between PC1/PC2 and biological traits (gene expression, histonemark enrichment, GC content, gene density, DNase I hypersensitivity [DHS]) will be calculated. Gene expression and histone mark data generated in subaim 1.1 along with DHS of fetal brain from Epigenomic roadmap[54] will be used and average values per 100kb bin calculated. In addition to the putative cis-elements identified in the same samples, we will also use the 15 state epigenetic marks from Epigenomic Roadmap[54] in genomic regions classified based on compartments averaged across 40kb bins, as well as subject specific psychENCODE data. Epigenetic state counts[54] for one compartment category are normalized by total epigenetic mark number of that compartment category and compared between samples.

*Dysregulation in ASD brains*. Two main data analyses will be performed with the transcriptome data. We will use the same approach as in subaim 2.1 to identify DEX coding and non-coding transcripts (by DESeq) between ASD and matched controls. Gene function enrichment analysis will be performed for these DEX genes. Finally, we will also perform Weighted Gene Co-expression Network Analysis (WGCNA; http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/) to identify modules of differentially co-expressed genes in ASD cases. For ChIP-seq data, once peaks are called and filtered for quality and reproducibility, we will identify and catalog all putative enhancer and promoter sites gained or lost in ASD brains compared to matched control brains, as well as what genes they are associated with.

In Hi-C data, we will assess if compartments, TADs, and contact domain structures are abrogated in ASD brains. Interaction partners for ASD risk genes, as well as gene loops involving ASD risk gene regulatory elements will be examined. Genome-wide interchromosomal contact matrices at high resolution (approximately 10kb) will be compared between ASD and control to identify bins that exhibit the largest chromosomal interaction changes in ASD (here we refer to ASD-specific bins). Gene ontology for these genes as well as their gene expression pattern in ASD may provide novel insights on ASD mechanism. The same approach will be applied to intrachromosomal contact matrices at 10-40kb resolution. TADs in ASD vs. controls will be also compared. The directionality index around ASD-specific TAD boundaries will be calculated to check significance. Moreover, we will examine gene expression level and histone marks on TAD boundaries as well as histone marks on TADs that contain ASD risk genes. Both inter- and intra-chromosomal interaction patterns of the bins that contain differentially expressed genes in ASD or ASD risk genes will be examined in ASD vs. controls. Gene expression pattern and histone states of genomic loci that highly interact with dup15q region will be assessed. *This approach of integrating chromosome interactomes to transcriptomic and epigenetic profiles may delineate epigenetic mechanism behind gene dysregulation in ASD.*

We will also perform integrative network analyses of these multi-level genomic data with genetic variation to understand the causal mechanism of transcriptional alterations in ASD (see also Aim 3). This will include integration of DNA sequence, methylation, chromatin contacts, eQTL and hQTL by this collaborative team of investigators (e.g. to include new hQTL methods by S. Prabhakar and colleagues[55]. Gene loops detected in control and ASD will be also interrogated. Gene loops that are specific to ASD or specific to controls may directly point out aberrant enhancer-promoter interactions, TF binding, or compartmentalization of genome. We will check if ASD-specific gene loops contain any ASD-associated variants (mostly common SNV at this point, although as more whole genome sequencing (WGS) data is available over the next 12 months, we can use these data to annotate potential functions of noncoding variants (Aim 3).

In ***subaim 2.2.**, we will integrate and harmonize data across psychENCODE projects and other relevant genomic resources*. In this aim, the DAC will integrate and harmonize our datasets with other psychENCODE studies and large-scale genomic datasets, such as BrainSpan, CommonMind, ENCODE, GTEx and REMC. The PsychENCODE DAC is led by Mark Gerstein and Nenad Sestan (Yale), Zhiping Weng (University of Massachusetts), who are part of this proposal and Kevin White (University of Chicago). DAC will summarize the major analysis results produced from psychENCODE and organize them into an encyclopedia of regulatory elements in the developing and adult human brain. We are currently building such an encyclopedia for the ENCODE consortium, and we will be able to leverage the methods that we are building for ENCODE and modify them to best serve psychENCODE data. The psychENCODE encyclopedia will include several components. The first component is the raw experimental data, including the expressed transcripts in neuronal and glial cells in various brain regions, the peaks (enriched regions) of an array of histone marks, the open chromatin regions detected using ATAC-seq, the differentially enriched histone mark peaks and open chromatin regions in ASD, BD and SCZ (diseases covered by psychENCODE projects. This component will largely result from a series of uniform processing pipelines, which we will build for analyzing psychENCODE data. The second component will include results that require the integration across multiple data types, including the enhancers in each cell type, the chromatin states called using a combination of histone marks and ATAC-seq data, and the topologically associated domains and compartments called by combining histone marks, ATAC-seq and Hi-C data. The third component of the encyclopedia will provide a higher-order organization to the elements in the first two components. Specifically we will derive the target genes for enhancers in a cell type specific manner, and identify the enhancer-gene links that are disrupted in the three diseases. We will also identify the variations that are linked with difference in gene expression (eQTLs) that are within enhancers that target the corresponding genes. Finally, we will develop a portal to guide the user through the components of the psychENCODE encyclopedia, with multiple entry points, such as genes, GWAS SNPs, or a specific regulatory region in the genome.

***Pitfalls and alternatives:*** Proposed computational approaches are well established in our team and we already have a considerable expertise and collaborative history therefore we foresee no complications in performing this aim. Furthermore, Sestan, State and Geschwind have been part of the BrainSpan project and Ernst, Gerstein and Weng has been part of several other relevant genomic consortia, such ENCODE.

### Aim 3. Spatiotemporal analysis in ASD.

***Rationale and preliminary supporting data:*** Over the past few years genomic analyses by our labs and others have made rapid progress in identifying genes associated with ASD, in particular through the identification of *de novo* mutations in ASD cases[13,17,26,30,45]. Despite the identification of these ASD-associated genes, progressing to an understanding of ASD neurobiology remains a challenge. Aims 1 and 2 described one approach to discovering this neurobiology through the identification of ASD-specific networks in *post mortem* brains. In Aim 3 we propose a complementary approach through the identification of genomic loci, brain regions, developmental stages, cell types, and neurobiological processes that are enriched for ASD mutations in genes (**subaim 3.1**) and non-coding loci (**subaims 3.2 and 3.3**) in neurotypical brains. Finally, we will test the hypothesis that ASD specific networks observed in *post mortem* brains from Aims 1 and 2 will be enriched for ASD associated mutations (**subaim 3.4**) thus demonstrating that the disruption of this network precedes the diagnosis of ASD and is therefore likely to be a cause of ASD rather than a consequence.

*1) Detection of ASD-associated genetic loci.* We identified rare and *de novo* variants in exome data from 5,563 ASD cases and 13,321 controls alongside rare and *de novo* copy number variants in microarray data from 4,687 ASD cases and 2,100 controls[17]. Comparison of these two data sets showed that small *de novo* deletions in ASD targeted the same set of genes as *de novo* loss of function point mutations in exome data. A combined analysis of exome data and small *de novo* deletions was performed using the Transmitted and *De novo* Association (TADA) method to identify ASD-associated genes. 28 ASD-associated genes were identified with very high confidence (false discovery rate (FDR) ≤ 0.01) and 65 ASD-associated genes were identified with high

confidence (FDR ≤ 0.1). These 65 genes formed a protein-protein interaction (PPI) network with two distinct subnetworks, enriched for chromatin regulatory genes and synaptic genes respectively (**Figure 6A**).

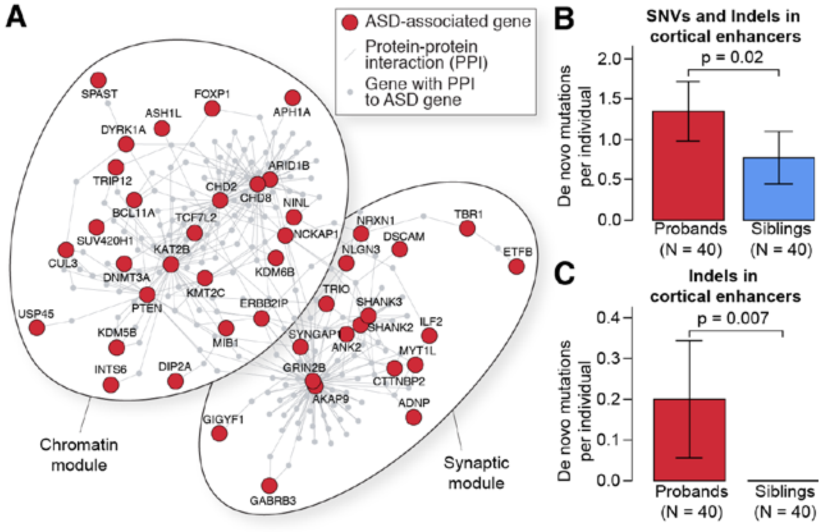


**Figure 6. ASD associated de novo mutations.** **A.** *65 ASD risk genes[9] (red) form a single protein-protein interaction network composed of two subnetworks. The genes in the left subnetwork are enriched for chromatin regulatory gene ontology terms. The genes in the right subnetwork are enriched for synaptic terms.* ***B.*** *De novo mutations were identified in WGS data for 40 ASD families. The median number of SNV and indel mutations per individual is shown within active enhancers that were identified by bulk tissue ChIP-Seq for H3K27ac in human dlPFC (psychENCODE phase 1 studies). P-values are calculated using linear regression with for paternal age and total de novo mutations per individual included as co-variates.* ***C.*** *The analysis was repeated for indels only.*

*2) Detection of ASD-associated non-coding variants in whole-genome sequencing (WGS) data*. We analyzed WGS data for 40 simplex ASD quartets composed of both parents, an affected child and an unaffected sibling control. The families were selected from the Simons Simplex Collection on the basis of no previous *de novo* loss of function or CNV mutations in exome and microarray data and high paternal age. The samples were sequenced to greater than 30x mean coverage (mean±standard 35.7±5.8). Raw data were aligned to hg19 human reference genome using BWA-mem[56]. Duplicate reads were removed with Picard (http://broadinstitute.github.io/picard/); GATK best practices[57] were used for all downstream steps including, local realignment, base quality score recalibration, SNV and indel calling, cohort-wide joint genotyping, and variant quality score recalibration. Data were normalized within families by only analyzing bases with at least 20 unique reads in all family members. A combination of PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq/) and in-house scripts were used to identify autosomal *de novo* variants based on stringent criteria designed to maximize specificity: minimum genotype likelihood (GQ) ≥20, alternate allele frequency (AB) ≤0.05 in the parents, and 0.3-0.7 in the child, minimum map quality (MQ) ≥30 in all family members, and allelic depth for the alternate allele (AD) ≥8. Approximately 7,000 *de novo* mutations were identified at a rate of 87.0±13.5 *de novo* mutations per child. Confirmation with Sanger sequencing was attempted on 10% of these variants (700) selected at random and achieved a >95% confirmation rate across both SNVs and indels, suggesting identification of *de novo* mutations with accuracy. We used tissue-level ChIP-seq for the histone modification H3K27ac from human dlPFC (psychENCODE phase 1) to identify active enhancers. We observed an increased burden of mutations in cases compared to sibling controls (p=0.02, **Figure 6B**) within these active enhancers. This association was especially strong for insertion/deletions (indels), possibly due to the greater functional impact of disrupting multiple nucleotides (p=0.007, **Figure 6C**).

*3) Analysis of gene co-expression to identify spatiotemporal convergence of ASD-associated genes*. We considered the convergence between 9 ASD genes[15] for gene expression data from 57 neurotypical brains that spanned 15 developmental periods and 16 brain regions[47]. To identify spatiotemporal windows whilst retaining sufficient numbers of samples for co-expression analysis we used hierarchical clustering to identify four groups of brain regions and considered each of these in 13 overlapping time periods each composed of three developmental periods (**Figure 7A**). Within each of the resulting 52 (4 x 13) spatiotemporal windows we built networks around nine high confidence ASD genes by selecting the top 20 co-expressed genes. We assessed these 52 windows for spatiotemporal convergence related to ASD etiology through the degree of enrichment for 126 independent low confidence ASD genes (**Figure 7A**). We observed strong spatiotemporal convergence between ASD risk genes in the prefrontal and primary motor-somatosensory cortex during mid-fetal development (**Figure 7A**)[15]. Analysis of cell type specific marker genes within this network showed enrichment for cortical projection neurons. This result that has been replicated by three complementary techniques: WGCNA[31], cell specific enrichment analysis[58], and NETBAG+ systems analysis[59].

*4) Comparison of ASD-related gene sets and gene expression analysis of post-mortem ASD brains.* Two prior analyses have identified gene co-expression WGCNA modules that are differentially expressed in the brain in ASD cases compared with controls. The microarray analysis by Voineagu et al.[9] identified a module enriched

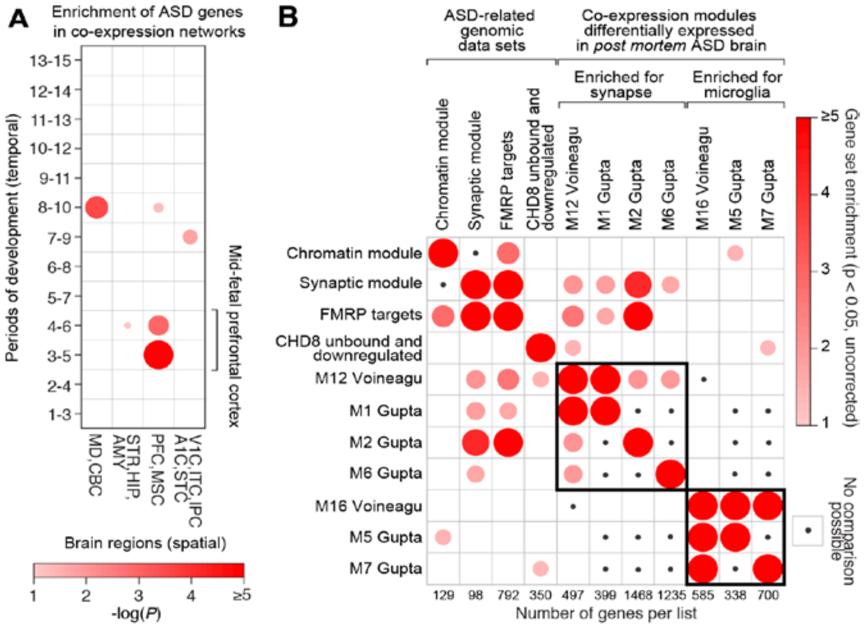


***Figure 7. Enrichment of ASD-associated genes in gene expression data. A.*** *Spatiotemporal co-expression networks were formed around nine high confidence ASD genes for 4 groups of brain regions (x-axis) and 13 overlapping developmental periods (y-axis). The –log(P) value for enrichment with 126 low confidence ASD genes is shown by the size and shade of the circle. Strong enrichment is observed in the mid-fetal PFC and primary motor-sensory cortex (PFC-MSC).* ***B.*** *Four ASD related gene sets[9,52,53] compared to seven WGCNA co-expression modules that are differentially expression in post mortem ASD brains (right). Fold enrichment is indicated by the size and shade of the circle. A synaptic and microglial module are indicated by the black rectangles. Small black circles show gene sets that are non-overlapping by definition (e.g. WGCNA modules in the same analysis).*

for synaptic genes (M12) that overlaps with three modules (M1, M2, and M6) identified using RNA-seq in Gupta et al[60] (**Figure 7B**). Similarly, one module enriched for microglial genes (M16) was observed in the Voineagu et al.[9] paper and this overlaps with two modules (M5 and M7) identified in the Gupta analysis (**Figure 7B**). We compared these seven WGCNA modules with four sets of ASD-related genes: the chromatin and synaptic modules from our recent analysis of exome and CNV data (**Figure 6A**)[17], RNA targets of the fragile X protein FMRP[61] that are strongly enriched within ASD-associated genes[26], and genes that are downregulated in CHD8 knockdown but not bound by CHD8 on ChIP-Seq analysis that have been described as targeting synaptic genes associated with ASD[62]. The synaptic module and FMRP targets are strongly enriched through the synaptic WGCNA modules suggesting these modules may represent causal factors that persist in the ASD brain. Further analysis is required to determine if these modules are causal or simply a consequence of ASD.

***Experimental design and methods:*** In ***subaim 3.1****, we will increase the spatiotemporal resolution of co-expression analysis of ASD neurobiology*. Our prior analysis of spatiotemporal convergence, described in detail under preliminary data[15], was based on 57 neurotypical brains, 9 high confidence ASD genes (FDR ≤0.05), and 126 low confidence ASD genes (FDR ≤0.3)[17]. These data enabled us to examine 4 groups of brain regions spanning multiple developmental periods (**Figure 7A**). The data from Aim 1 and our progress in ASD gene discovery will allow us to perform this analysis using 87 neurotypical brains, 28 high confidence ASD genes (FDR ≤0.01), and 151 low confidence ASD genes (FDR ≤0.3). As before (**Figure 7A**), the gene expression samples will be divided into spatiotemporal windows using hierarchical clustering to group related brain regions (spatial) and considering overlapping developmental windows (temporal). In each spatiotemporal window we will identify the top 20 co-expressed genes around 28 high confidence ASD genes and, following the logic that a spatiotemporal network relevant to ASD should be enriched for other ASD genes, we will assess the enrichment of the 151 low confidence ASD genes (FDR ≤0.3). The expanded number of brain samples will enable us to use small subdivisions of brain regions and developmental time regions to increase the resolution of the analysis, for example windows spanning one or two developmental periods. In addition, the larger list of high confidence ASD genes will allow us to perform the analysis by building the spatiotemporal networks around subsets of these 28 genes and improve the accuracy of the analysis through cross validation. In addition, we will divide the 28 high confidence genes by the two main functional categories observed, specifically chromatin regulators and synaptic genes, to assess the spatiotemporal dynamics of each functional category separately. The outcome of this aim will be refined gene co-expression networks that show spatial and temporal convergence among ASD risk genes.
***Pitfalls and alternatives:*** The analytical methods described here have been applied to the BrainSpan data using 9 high confidence genes resulting in the discovery of spatiotemporal convergence in the frontal cortex of the mid-fetal brain. This finding has been replicated using complementary methods[58,59]. In this aim we will be increasing the resolution through the inclusion of additional gene expression data and novel ASD-associated genes[17], therefore we do not foresee complications. An alternative 'top down' methodology such as WGCNA, in which co-expression modules are generated from the complete dataset and are then assessed for enrichment

of ASD genes, has yielded similar findings[31]. We will also apply this complementary WGCNA method across spatiotemporal windows.

In ***subaim 3.2**, we will identify ASD-associated non-coding de novo mutations in regulatory loci.* Under pre-existing funding arrangements we will have access to whole-genome sequencing (WGS) data for 5,120 individuals from 1,280 quartet families composed of two parents, an affected child, and an unaffected sibling control. We have previously reported an increased burden of *de novo* mutations between the affected and unaffected siblings[17] and we have observed this for *de novo* CNVs in microarray data and *de novo* loss of function mutations in exome data. To identify functional non-coding *de novo* mutations in regulatory loci, we will leverage the integrated RNA-Seq, ChIP-Seq, and HiC data from Aims 1 and 2 with the *de novo* mutation identification approach described in our preliminary data (**Figure 6**). To maximize our ability to discover compartments of the genome that carry risk we will assess *de novo* burden in three sets of loci: **1**) All regulatory loci identified in neurotypical brain divided by function (e.g. promoter, 3`UTR); **2**) Regulatory loci identified in neurotypical brain with a relationship to 28 high-confidence ASD genes; and **3**) Regulatory loci identified in neurotypical brain with a relationship to the points of convergence for ASD genes identified in Subaim 3.1 such as prefrontal cortex in mid-fetal development. The outcome of this aim will be non-coding mutations and regulatory loci that show association with ASD.

***Pitfalls and alternatives:*** Our methods for identifying *de novo* mutations in whole genome sequencing data are well developed and we have demonstrated a >95% confirmation rate for the mutations predicated. Additionally, our preliminary data, based on 40 families, shows evidence of ASD association for *de novo* mutations within enhancers active in human dlPFC (**Figure 6B and C**). This suggests the proposed study of 1,280 families will offer sufficient power even if the overall contribution of *de novo* mutations in the non-coding genome to ASD etiology is relatively weak. To maximize our chance of identifying ASD associated non-coding variants we will assess only the loci with the strongest evidence of functional activity, including the larger mutations, such as indels, that may carry the greatest risk. Concurrently, Dr. Sanders has an established collaboration with Mike Talkowski and the GATK CNV/SV working group to develop methods that maximize our sensitivity for detecting indels and small CNVs in whole genome sequence data.

In ***subaim 3.3**, we will identify points of spatiotemporal convergence using ASD associated non-coding mutations:* Non-coding elements such as enhancers frequently show a degree of specificity to particular developmental time points, brain regions, or cell types[63]. We will use the ASD-associated non-coding *de novo* mutations in regulatory loci and regulatory loci related to ASD associated genes to assess which integrated regulatory networks from Aim 2 show the greatest enrichment for these non-coding mutations. By considering the brain regions and developmental epochs in which these networks exist we will assess points of spatiotemporal convergence critical to ASD. The outcome of this aim will be an independent analysis of points of spatiotemporal convergence in ASD based on non-coding mutations and regulatory loci.

***Pitfalls and alternatives:*** This aim relies on the discovery of specific ASD-associated regulatory loci through the discovery of numerous *de novo* mutations in cases. Due to the small size of regulatory regions we may not see this clustering in a single regulatory element. Should this be the case we will use genomic annotation to rank the regulatory loci with a single mutation, for example considering conservation, constraint [64], and large mutations such as indels that are more likely to disrupt the element (**Figure 6C**).

In ***subaim 3.4**, we will assess regulatory networks that are observed in the post mortem ASD brain.* Aims 3.1 to 3.3 focus on neurotypical brains and their association with ASD-associated mutations. In this aim we will assess the enrichment of ASD-associated genes, non-coding mutations and regulatory networks that differ between *post mortem* ASD and neurotypical brains (**Figure 6**). Because genetic variants associated with ASD precede the onset of ASD symptoms, enrichment for these mutations will suggest that such networks are causal (**Figure 7**) to the ASD phenotype. Conversely, a lack of enrichment for these mutations in ASD-relevant networks will suggest the network is consequential to ASD. The outcome of this aim will therefore be to distinguish ASD-specific regulatory networks that are likely to be causal from those that may be consequential.

***Pitfalls and alternatives:*** Methods for assessing such enrichment are well established and we already have a large list of ASD-associated genes; we foresee no complications in performing this aim. The main challenge lies in the interpretation of a regulatory network that does not enrich for ASD-associated genes (e.g. microglia in existing *post mortem* analyses, **Figure 7B**), since this may indicate a non-causal relationship or reflect and incomplete list of ASD-associated genes. We will therefore focus on networks with positive enrichment for these genes and acknowledge the complexities of interpreting a negative result.

**TIMELINE AND MILESTONES SECTION** See Other Attachments

## SPECIFIC AIMS

The necessity for understanding gene regulation in human brain development is supported by several recent discoveries. For example, most inherited common genetic variation underlying neuropsychiatric diseases lies in non-coding regions and is presumed to exert pathogenic effects via the regulation of gene expression and splicing[1-4]. Additionally, most non-inherited (*de novo*) highly penetrant ASD risk genes are enriched in co-expression modules and protein interaction networks related to chromatin remodeling and transcriptional regulation[3-8]. Moreover, a specific shared pattern of transcriptional dysregulation is observed in the cerebral cortex in slightly more than 2/3 of post-mortem ASD cases[9,10]. Taken together, these observations emphasize the importance of integrating transcriptomic and epigenomic data with higher-order chromatin interactions to better understand the putative mechanisms underlying dysregulated genes and networks in ASD and other psychiatric disorders, a fundamental goal of psychENCODE. **The primary goal of this application is to extend our ongoing analyses of healthy and ASD brains under the psychENCODE consortium with the inclusion of additional genomic features, brain regions, developmental time points and cell-type specific analyses.** By performing these analyses we will enhance this public resource and improve our understanding of the molecular processes underlying normal human neurodevelopment and ASD.

Our group has been collaborating closely for a decade[11-15], bringing together expertise in developmental neurobiology, human tissue biobanking, genetics and genomics, statistics, bioinformatics and systems biology. Several key conceptual threads have been apparent in our work together: 1) Revealed new insights into human neurodevelopment through functional genomic profiling of postmortem tissue and cell culture models[12,16]; 2) Assessed rare and *de novo* mutations for ASD association[13,17,18], based on the notion that down-stream analyses are only as good as the genes that go into them; 3) Identified the neural processes and pathways that are altered in the presence of ASD-associated mutations, as well as when and where these processes and pathways occur in the developing human brain[15,17,19]. Here, we propose to continue this highly productive collaboration and expand psychENCODE phase 1 efforts through three integrated aims.

**Aim 1. Time, region and cell type-specific molecular profiling of control and ASD brains.** In *subaim 1.1*, we will profile the transcriptome (by RNA-seq), *cis*-regulatory elements (ChIP-seq) and 3D chromatin architecture (Hi-C) in neurotypical dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR) during mid-fetal development, infancy, childhood, adolescence and adulthood. To address cellular heterogeneity and to complement the psychENCODE phase 1 tissue level data analyses, we will obtain these data from neuronal and non-neuronal nuclei collected with fluorescence-activated nuclei sorting (FANS). In *subaim 1.2.*, complementary genomic analyses will be done on the FANS nuclei from syndromic and idiopathic ASD brains and matched control brains**,** to identify transcripts, regulatory elements, and 3D chromatin structures altered in ASD in brain region and cell type-specific manners.

**Aim 2. Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains.** In *subaim 2.1*, each dataset generated in Aim 1 will be analyzed to identify differences between the developmental stages and two major cell types in healthy and ASD tissue. Furthermore, these datasets will be integrated to gain comprehensive insights into the underlying mechanisms; Hi-C defined physical intrachromosomal interactions will be intersected with ChIP-seq to identify functional interactions between regulatory sequences potentially associated with transcriptional changes. In *subaim 2.2*, we will harmonize and integrate our multi-omic datasets with other psychENCODE studies and large-scale genomic datasets, such as BrainSpan, CommonMind, ENCODE, GTEx and REMC.

**Aim 3. Spatiotemporal analysis in ASD.** Our prior work assessed the enrichment of ASD genes in spatiotemporal co-expression networks to identify the frontal cortex during mid-fetal development as a critical window in ASD etiology. In *subaim 3.1*, we will use the neurotypical gene expression data and our expanded list of ASD associated genes to increase the resolution of this spatiotemporal analysis. In *subaim 3.2* we will use whole-genome data for 5,120 individuals in ASD families to identify non-coding *de novo* mutations within the regulatory loci identified in neurotypical brains in Aims 1 and 2. In *subaim 3.3* we will use these non-coding mutations and the regulatory networks from Aim 2 to perform an independent assessment of spatiotemporal convergence in ASD to complement our gene-based analysis in subaim 3.1. Finally, in *subaim 3.4* we will use the regulatory networks that are specific to the ASD brain identified in Aim 2 to assess enrichment of ASD-associated genes and non-coding mutations thus demonstrating that such networks are causally linked to ASD rather than simply a consequence of ASD. At the completion of this aim we will have three independent assessments of spatiotemporal convergence in ASD from ASD-associated genes, ASD-associated regulatory loci, and ASD-associated networks in the *post mortem* brain.

## SPECIFIC AIMS

The necessity for understanding gene regulation in human brain development is supported by several recent discoveries. For example, most inherited common genetic variation underlying neuropsychiatric diseases lies in non-coding regions and is presumed to exert pathogenic effects via the regulation of gene expression and splicing[1-4]. Additionally, most non-inherited (*de novo*) highly penetrant ASD risk genes are enriched in co-expression modules and protein interaction networks related to chromatin remodeling and transcriptional regulation[3-8]. Moreover, a specific shared pattern of transcriptional dysregulation is observed in the cerebral cortex in slightly more than 2/3 of post-mortem ASD cases[9,10]. Taken together, these observations emphasize the importance of integrating transcriptomic and epigenomic data with higher-order chromatin interactions to better understand the putative mechanisms underlying dysregulated genes and networks in ASD and other psychiatric disorders, a fundamental goal of psychENCODE. **The primary goal of this application is to extend our ongoing analyses of healthy and ASD brains under the psychENCODE consortium with the inclusion of additional genomic features, brain regions, developmental time points and cell-type specific analyses.** By performing these analyses we will enhance this public resource and improve our understanding of the molecular processes underlying normal human neurodevelopment and ASD.

Our group has been collaborating closely for a decade[11-15], bringing together expertise in developmental neurobiology, human tissue biobanking, genetics and genomics, statistics, bioinformatics and systems biology. Several key conceptual threads have been apparent in our work together: 1) Revealed new insights into human neurodevelopment through functional genomic profiling of postmortem tissue and cell culture models[12,16]; 2) Assessed rare and *de novo* mutations for ASD association[13,17,18], based on the notion that down-stream analyses are only as good as the genes that go into them; 3) Identified the neural processes and pathways that are altered in the presence of ASD-associated mutations, as well as when and where these processes and pathways occur in the developing human brain[15,17,19]. Here, we propose to continue this highly productive collaboration and expand psychENCODE phase 1 efforts through three integrated aims.

**Aim 1. Time, region and cell type-specific molecular profiling of control and ASD brains.** In *__subaim 1.1__*, we will profile the transcriptome (by RNA-seq), *cis*-regulatory elements (ChIP-seq) and 3D chromatin architecture (Hi-C) in neurotypical dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR) during mid-fetal development, infancy, childhood, adolescence and adulthood. To address cellular heterogeneity and to complement the psychENCODE phase 1 tissue level data analyses, we will obtain these data from neuronal and non-neuronal nuclei collected with fluorescence-activated nuclei sorting (FANS). In *__subaim 1.2.__*, complementary genomic analyses will be done on the FANS nuclei from syndromic and idiopathic ASD brains and matched control brains**,** to identify transcripts, regulatory elements, and 3D chromatin structures altered in ASD in brain region and cell type-specific manners.

**Aim 2. Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains.** In *__subaim 2.1__*, each dataset generated in Aim 1 will be analyzed to identify differences between the developmental stages and two major cell types in healthy and ASD tissue. Furthermore, these datasets will be integrated to gain comprehensive insights into the underlying mechanisms; Hi-C defined physical intrachromosomal interactions will be intersected with ChIP-seq to identify functional interactions between regulatory sequences potentially associated with transcriptional changes. In *__subaim 2.2__*, we will harmonize and integrate our multi-omic datasets with other psychENCODE studies and large-scale genomic datasets, such as BrainSpan, CommonMind, ENCODE, GTEx and REMC.

**Aim 3. Spatiotemporal analysis in ASD.** Our prior work assessed the enrichment of ASD genes in spatiotemporal co-expression networks to identify the frontal cortex during mid-fetal development as a critical window in ASD etiology. In *__subaim 3.1__*, we will use the neurotypical gene expression data and our expanded list of ASD associated genes to increase the resolution of this spatiotemporal analysis. In *__subaim 3.2__* we will use whole-genome data for 5,120 individuals in ASD families to identify non-coding *de novo* mutations within the regulatory loci identified in neurotypical brains in Aims 1 and 2. In *__subaim 3.3__* we will use these non-coding mutations and the regulatory networks from Aim 2 to perform an independent assessment of spatiotemporal convergence in ASD to complement our gene-based analysis in subaim 3.1. Finally, in *__subaim 3.4__* we will use the regulatory networks that are specific to the ASD brain identified in Aim 2 to assess enrichment of ASD-associated genes and non-coding mutations thus demonstrating that such networks are causally linked to ASD rather than simply a consequence of ASD. At the completion of this aim we will have three independent assessments of spatiotemporal convergence in ASD from ASD-associated genes, ASD-associated regulatory loci, and ASD-associated networks in the *post mortem* brain.

## RESEARCH STRATEGY

## SIGNIFICANCE

Neuropsychiatric disorders such as autism spectrum disorder (ASD), bipolar disorder (BD), and schizophrenia (SCZ) are complex and devastating illnesses with considerable morbidity and mortality, as well as high personal and societal costs. Many of them are also polygenic, with multiple variants, both rare and common, spread throughout the genome influencing the disease risk[3]. Recent studies have identified rare variants contributing to psychiatric disorders that are enriched in genes involved in global gene regulation and chromatin modification, and many common risk variants are enriched in regulatory regions of the human genome, regions whose functions are poorly understood. The interpretations of these variations in regulatory regions will certainly be improved with better maps of RNA transcripts, regulatory elements, and chromatin states in the human brain. The age of onset and progression of major psychiatric disorders also varies (**Figure 1**) necessitating the study of the temporal dynamics of gene regulation during human brain development and recognizing the developmental context of psychiatric disorders. An emerging body of research indicates that many aspects of the development and physiology of the human brain are not well recapitulated in model organisms[20-24] **and therefore it is increasingly apparent that psychiatric disorders need to be understood in the broader context of human brain development and physiology.**

In recent years, considerable effort has been made by many studies, including large-scale efforts by ENCODE, NIH Roadmap (REMC) and GTEx projects to survey the diversity of *cis*-acting regulatory regions and RNA species of the human genome across different tissues and time points. However, a comprehensive catalog of transcripts, regulatory elements, epigenetic modifications, and chromatin structure from the human brain during development and in distinct brain regions and cell types is lacking. The PsychENCODE (phase 1) projects have initiated these efforts.

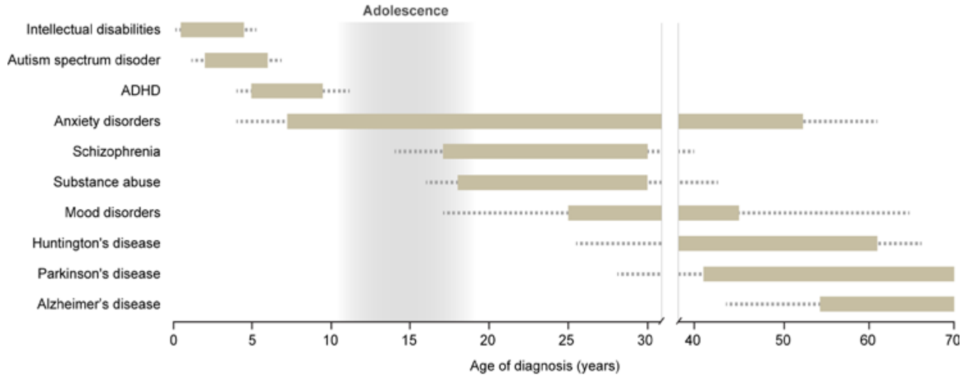


**Figure 1. Psychiatric and neurological disorders have discrete ages of onset.** *The bars indicate the age range that each disorder commonly affects, with less frequent ages of diagnosis denoted as dotted lines. This variability is indicative of dysregulation of tightly controlled developmental processes and highlights the necessity of defining the spatio-temporal molecular processes in human brain.*

**PsychENCODE consortium projects.** The key goals of the PsychENCODE project are to provide an enhanced framework of regulatory elements, catalog epigenetic modifications, and quantitate coding and non-coding RNA and protein expression in a tissue- and cell-type specific manner from neurotypical (healthy) control brains and diseased post-mortem human brains[25]. These efforts will be complemented with integrative analyses, as well as with functional characterizations of disease-associated genomic elements using human neural cell systems or the developing mouse brain. However, the human brain is heterogeneous cellularly and its development is regionally asynchronous and prolonged. *To overcome issues that hamper the potential benefits of initial psychENCODE studies, we will apply several approaches to address regional and cellular heterogeneity, prolonged development, and new genomic methods in the context of brain development and ASD.*

Here we focus on **neurotypical (control) brain and ASD**, which is a complex developmental syndrome with a significant genetic contribution. Although considerable genetic and phenotypic heterogeneity has complicated efforts to establish the biological substrates of the syndrome, the emergence of reliable genetic findings has started to shed light on potential pathogenic mechanisms, providing an extraordinary opportunity for developing a mechanistic understanding of the disorder. Recent studies suggest that **over 500 rare, *de novo* mutations contribute to ASD risk and no single genetic mutation accounts for more than 1% of ASD cases**[13,17,26-30], consistent with significant heterogeneity in this, and other neuropsychiatric disorders[3]. Despite this heterogeneity, mapping **ASD risk genes onto co-expression networks that represent normal human brain development has revealed that ASD genes coalesce in modules related to chromatin remodeling and transcriptional regulation during early fetal brain development, suggesting potential convergent pathways in the disorder**[9,15,27,31]. Another remarkable finding that parallels the convergence of genetic findings in developmental pathways is the identification and validation of shared transcriptional changes in postmortem brain in ASD[9]. This transcriptional dysregulation, coupled with the evidence that large effect size *de novo* ASD

risk genes are highly enriched in chromatin modifying genes (many of which are expressed in early fetal brain development), emphasizes the importance of understanding the nature and extent of chromatin disorganization in ASD brain and in normal brain development. **Further, since these data suggest distinct neuronal and glial gene dysregulation, it is crucial to delineate the profiles of these major cell types.** In addition to our ongoing efforts in PsychENCODE phase I project, this proposal provides critical advances in our understanding of the role(s) of non-coding functional elements in the pathophysiology of ASD and a scaffold for understanding chromatin structure and gene regulation across normal brain development. Overall, the approach proposed here will provide mechanistic insights that connect distinct transcriptional programs associated with ASD pathogenesis, and will provide a resource of the mechanisms of gene regulation across brain development to inform other neuropsychiatric disorders, a key goal of psychENCODE. **This work also leverages psychENCODE phase 1 projects by adding significant new data to expand the value of the resource and by directly addressing key areas of interest in control and ASD brains as outlined in RFA-MH-16-230**: ***1)*** *Generation of comprehensive, high resolution human brain region/cell type and age-specific maps of different classes of RNA transcripts, regulatory elements, chromatin states, chromatin conformation, and chromatin interactions;* ***2)*** *Identification of human brain region/cell type and age-specific molecular processes;* ***3)*** *Integration of these newly generated multi-omic datasets, from diseased and healthy control brains, with large-scale genomic resources;* ***4)*** *Generation and analysis of high-depth, whole genome sequencing data to allow for improved evaluation of various genetic alterations; and* ***5)*** *Development of comprehensive molecular models of disease (i.e., ASD) using systems biology approaches.*

## INNOVATION

This proposal is innovative in several aspects. First, to the best of our knowledge, the systematic discovery and functional characterization of genomic non-coding elements and 3D chromatin architecture has not been performed in healthy developing human brains or ASD brains at a cell type-specific resolution. For example, we use Hi-C, which combines chromosome conformation capture and NextGen sequencing to identify physical interactions that capture multiple levels of chromosome architecture ranging from nuclear configuration (“compartments” of about 5Mb) to TADs (domains of 500kb on average) and gene loops (often reflect enhancer promotor relationships; 40kb average), and is the only such method that spans all of these levels, genome-wide[32-34]. Second, this project will conduct direct analysis of one of the largest collection of well-characterized high quality healthy as well as syndromic and idiopathic ASD postmortem brains. Third, we will combine fluorescence-activated nuclei sorting (FANS) with advanced genomic techniques to analyze multiple genomic features in archived development control and ASD brains. Fourth, we will leverage these analyses with our ongoing psychENCODE phase 1 tissue level analyses and other recent large-scale genomic resources, such as BrainSpan, ENCODE, GTEx and Roadmap project. Therefore, our proposed data and integrated analyses has potential to improve our understanding of genomic processes and normal human brain development as well as diagnostics, neurobiology and treatment of ASD.

## COLLABORATION

This collaboration brings together multiple groups with long standing expertise in developmental neurobiology, psychiatry, human biobanking, genetics and genomics, statistics, bioinformatics, and systems biology that have worked closely with one another for almost a decade as evidenced by many co-publications. Several key conceptual threads have been apparent in our work together related to human brain development and neuropsychiatric disorders: **1**) Revealed new insights into human neurodevelopment through functional genomic profiling of postmortem tissue and cell culture models[12-16]; **2**) Assessed rare and *de novo* mutations for ASD association[13,17,18]; **3**) Identified the neural processes and pathways that are altered in the presence of ASD-associated mutations, as well as when and where these processes and pathways occur in the developing human brain[15,17,19]. In addition, M. Gerstein (Yale) and Z. Weng (University of Massachusetts), experts in bioinformatics and computational biology, are leaders of the PsychENCODE DAC, which will normalize the data to remove batch effects, establish uniform data processing pipelines and build calibration resources for all assays to enable comparison and integration of the data generated by all psychENCODE groups. The efforts of each group will be tightly integrated in order to communicate progress and results, design and implement analytical tools, and transfer data. Given the complexity of human neurodevelopment and genetics/neurobiology of ASD, we believe that integrating the respective expertise of these groups, and their respective collaborators at UCLA (Ernst and Geschwind), UCSF (Sanders, State and Willsey), UMass (Weng), and Yale (Gerstein and Sestan), offers the best opportunity to better understand human brain development and ASD through functional genomics. Here, we propose to leverage our expertise and continue this highly productive collaboration and expand psychENCODE phase 1.

**ELEMENTS UNIQUE TO THIS SITE** (Yale; Sestan, PI; Gerstein and Weng, co-investigators)
In *subaim 1.1*, the Sestan lab will apply RNA-seq to profile of neuronal and non-neuronal transcriptomes of the developing and adult dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR) using FANS. ChIP-seq will identify putative enhancers and promoters in the same samples.
In *subaim 1.2*, the Sestan lab will perform complementary studies in ASD and matched control brains.
In *Aim 2*, the Yale and University of Massachusetts members of the psychENCODE data analysis center (DAC), Mark Gesrtein, Nenad Sestan and Zhiping Weng, will implement computational approaches to elucidate common and cell type specific regulatory, transcriptional and molecular networks that are active during brain development and are compromised in ASD. We will also leverage the tissue level transcriptome and epigenome datasets generated in the BrainSpan and PsychENCODE 1 as well as integrate these newly generated multi-omic datasets, from ASD and healthy control brains, with several large-scale genomic resources.

## APPROACH

The objective of this proposal is to extend our ongoing tissue level analyses of healthy and ASD brains under the psychENCODE consortium with the inclusion of additional genomic methods, brain regions, developmental time points, and cell-type specific analyses. By performing three integrated aims (**Figure 2**) we propose to enhance this public resource and improve our understanding of the molecular processes underlying normal human neurodevelopment and ASD.

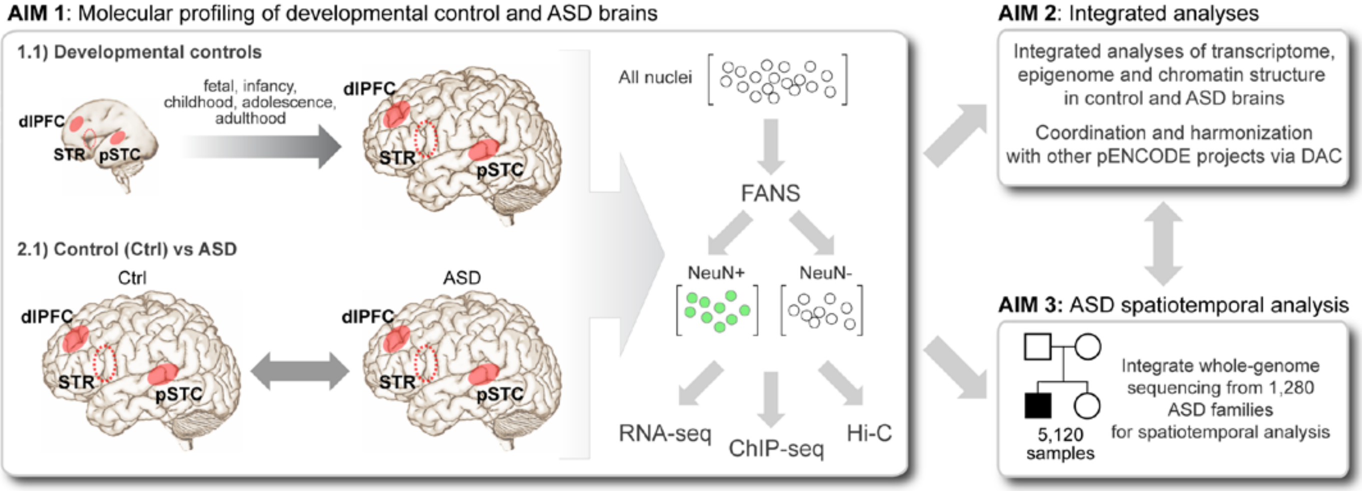

AIM 1: Molecular profiling of developmental control and ASD brains
1.1) Developmental controls
dlPFC
STR
pSTC
fetal, infancy, childhood, adolescence, adulthood
dlPFC
STR
pSTC
2.1) Control (Ctrl) vs ASD
Ctrl
ASD
dlPFC
STR
pSTC
dlPFC
STR
pSTC
All nuclei
FANS
NeuN+
NeuN-
RNA-seq
ChIP-seq
Hi-C
AIM 2: Integrated analyses
Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains
Coordination and harmonization with other pENCODE projects via DAC
AIM 3: ASD spatiotemporal analysis
5,120 samples
Integrate whole-genome sequencing from 1,280 ASD families for spatiotemporal analysis

**Figure 2. Schematic workflow of three specific aims.**

### Aim 1. Time, region and cell type-specific molecular profiling of control and ASD brains.

***Rationale and preliminary supporting data:*** Three major observations provide motivation for this aim. The *first* is the recognition that genomic data, including transcriptomic, epigenetic and physical chromatin structure, from the relevant neurotypical tissue (control), spanning the key epochs of neurodevelopment and function from fetal to adult periods, provide a new and previously unobtainable view of genetic risk for psychiatric disease[10,15,16,31,35,36]. The *second* is that brain is comprised of an extremely heterogeneous mixture of cell types that exhibit distinct molecular profiles, including glia-to-neuron ratios that could show considerable fluctuations across normal development or in certain disease states. The *third* is the observations of differences in transcriptome organization via tissue-level gene co-expression network analysis conducted between ASD and normal brains[9]. Thus, here **we propose to create a region and cell type-specific normal developmental scaffolding on which to frame disease variants via transcriptional (RNA-seq), epigenetic (ChIP-seq) and chromatin architecture (HiC) profiling of neuronal and non-neuronal cells at key epochs in human brain development (subaim 1.1), as well as compare these profiles in ASD and matched control brains (subaim 1.2) to help elucidate the mechanisms by which genetic variation alters brain development and function, leading to ASD and related neuropsychiatric conditions.** While several genomic features are currently being analyzed in control and ASD brains by our and other groups in the psychENCODE consortium, cellular heterogeneity during development, other genomic features (e.g. 3D chromatin contacts), have yet to be addressed. To address these issues, we will utilize our large, high quality, phenotypically well-characterized human brain collection (see Facilities and Resources section), as well as newly implemented methods to collect molecularly defined cell type specific nuclei from archival human postmortem brains in this collection.

**Our preliminary data demonstrates a clear pattern of transcriptional dysregulation is observed in 2/3 of ASD brains[9], which we have now confirmed in our psychENCODE phase 1 projects (in a more than**

**double sized sample of cases and controls)** using tissue level RNA-seq and ChIP-seq (H3K4me3 and H3K27ac) in multiple brain regions in 43 idiopathic ASD cases, 8 cases with chromosome 15q11-13 duplication syndrome (dup15q) and ASD, and 63 controls[37,38]. We also observe that post mortem brain from patients with ASD caused by (dup)15q11-13 share this same pattern at all levels of differential protein coding gene expression, splicing and lncRNA[38]. As the first step in exploring potential mechanisms, we performed epigenetic profiling of ASD vs. control brains with H3K27ac marks, which indicate active enhancers[39]. Genes with differential H3K27ac peaks in their promoter regions (5000bp upstream of the transcription start site) were enriched with neuronal genes with changes in expression. This result demonstrates that transcriptional dysregulation in ASD is partially mediated by changes in histone/chromatin modifications. Furthermore, the two major groupings of modules derived from whole tissue gene expression analysis sort into those up-regulated and expressed in glia, and those down-regulated in neurons[9,38], strongly motivating our plan for transcriptional, epigenetic and Hi-C profiling in neurons and non-neuronal (glial) cells independently.

**Another of the important advances in methodology that will be implemented here is the assessment of 3-D chromatin structure across to different brain regions and cell types, and 5 key epochs of normal brain development and in ASD brains.** Our preliminary data strongly supports the value of these data and our ability to perform and analyze these experiments (see also[40]). We established an efficient Hi-C protocol and obtained high resolution data (10 kb resolution; via deep sequencing) from the fetal cortex from 3 individuals dissected into two zones: cortical plate (CP) and germinal zones (GZ) at post-conception week (PCW) 18 (total n = 12 samples: representative heatmap shown in **Figure 3A**). Demonstrating the data quality, principal component of the interchromosomal interaction matrix for GZ shows a high correlation with GC content ($r$=0.791, $P$<$10^{-256}$), gene number ($r$=0.384, $P$<$10^{-256}$), DNase I hypersensitivity ($r$=0.416, $P$<$10^{-256}$), and to a lesser extent, gene expression ($r$=0.129, $P$=1.11x$10^{-66}$; **Figure 3B and C**), recapitulating previous work in cell lines[41]. We next asked how chromatin interactions elicit transcriptional co-regulation. We hypothesized that highly interacting chromatin regions would be co-regulated at least in part by sharing chromatin remodelers and transcription factors (TFs). To test this, we binned chromatin interactions into top and bottom percentiles, and compared the distribution of correlation patterns for genes in the high and low interacting regions of chromatin. We observed that the high interacting regions were significantly biased toward positive correlations (**Figure 4A**), supporting the hypothesis that co-localization can predict co-expression.

We next integrated these data with the epigenomics map from the NIH Roadmap project[42]. By comparing the epigenetic mark combination matrix with the Hi-C contact matrix, we demonstrate that interacting regions exhibit shared epigenetic patterns: loci associated with transcriptional regulation and enhancers are significantly more likely to interact with each other (**Figure 4B**). Comparison of TF binding site (TFBS) combination matrix (generated from TFBS map reported in[43]) with the intrachromosomal contact matrix revealed distinct combinatorial patterns of TF binding likely to mediate chromosome interactions (**Figure 4C**), thus revealing new experimentally testable regulatory relationships.

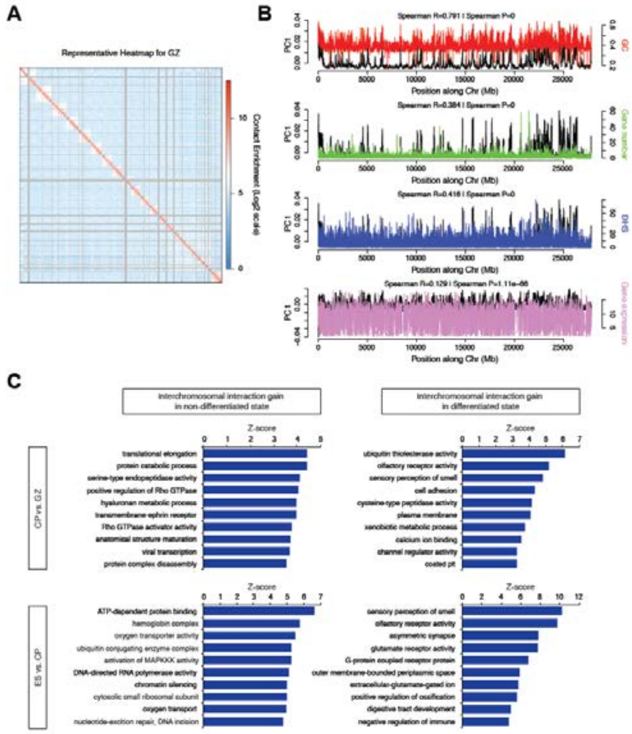


**Figure 3. Chromosome conformation in fetal brains (by Hi-C).** ***A.*** *Representative heatmap of chromosome contact matrix of GZ. Normalized contact frequency (Contact enrichment) is color-coded according to the legend on the right.* ***B.*** *Spearman correlation of PC1 of chromatin interaction profile of fetal brain (GZ) with GC content (GC), gene number, DNase I hypersensitivity (DHS), and gene expression level of fetal brains. These data show relationship of 3D structure to key known functional elements as has been previously shown in other systems.* ***C.*** *Gene ontology (GO) enrichment (GO Elite) of genes located in the top 5% of highly interacting inter-chromosomal regions specific to GZ vs. CP (top), and ES vs. CP (bottom), indicating that genes located on dynamic chromosomal regions are enriched for neuronal function in CP, which contains the more differentiated laminae. Please see Won et al. 2015 in Appendix for higher magnification figure.*

To validate that Hi-C data can identify target genes regulated by single nucleotide polymorphisms (SNPs) in a general setting, we determined if SNPs with a significant effect on gene expression were also identified as

interacting by Hi-C using *cis*-expression quantitative trait loci (eQTL) data from adult frontal cortex[44]. Indeed, Hi-$C_{eQTL}$ genes were significantly over-represented with known associated genes from the eQTL study and eQTL SNP-transcript pairs exhibit significantly higher chromatin contact frequency than the null across all distance ranges measured, further supporting the utility of Hi-C to infer the gene or region of activity for regulatory variation. In addition we asked whether significant physical cis-chromosomal contacts identified with Hi-C could inform functional annotation of 108 genome-wide significant schizophrenia loci, most of which lie far outside known coding or other functional regions of the genome.

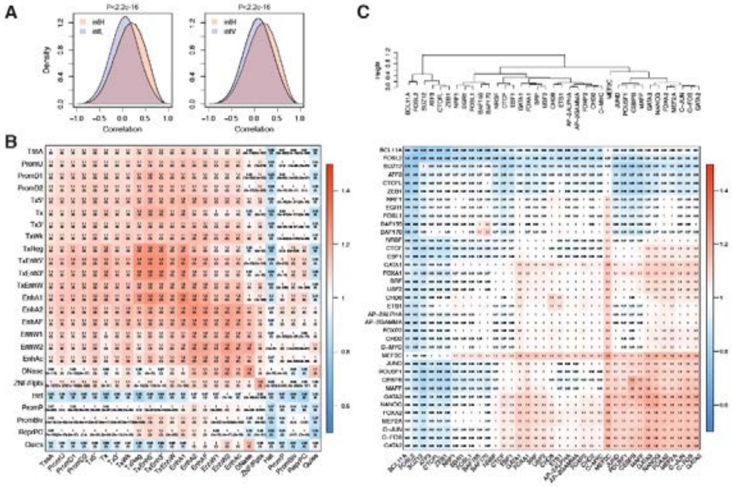


**Figure 4. Highly interacting regions share co-expression patterns, which is mediated by epigenetic regulation**. ***A.*** *The top 10,000 highest interacting regions (intH) in fetal brains both at GZ and CP show positive correlation in their gene expression patterns, while the top 10,000 lowest interacting regions (intL) and top 10,000 highly variant regions (intV) have no skew in the distribution, consistent with random interactions. P-value, Kolmogorov-Smirnov test.* ***B-C.*** *Epigenetic state combination (B) and TFBS combination (C) for intrachromosomal interacting regions. The epigenetic state matrix and TFBS combination matrix were generated by marking loci where two interacting chromosomal bins share epigenetic signature. For example, the epigenetic combination matrix between the active transcription start site (TssA) and active enhancers (EnhA1) is generated by marking where interacting loci have TssA and EnhA1. Intrachromosomal contact frequency map is compared to the epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Odds ratio (OR) and P-values are depicted in the heatmaps (Please see Won et al. 2015 in Appendix for higher magnification figure).*

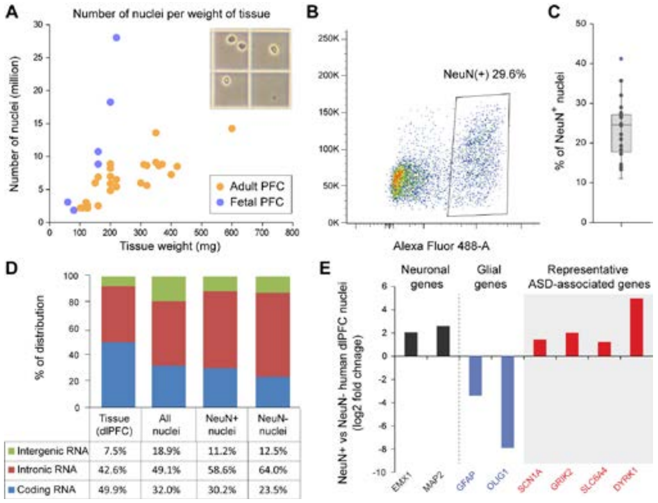
Although SNPs are typically assigned to the closest genes, or those within the LD block, Hi-C indicated that about 50% of the variants were neither adjacent to the index SNPs (most-associated SNP within a locus), nor in LD. Interestingly, Hi-$C_{SCZ}$ genes significantly overlap with ASD *de novo* likely gene-disrupting (LGD) targets[26,45] (CP: OR=2.4, P=1.6x$10^{-5}$, GZ: OR=1.8, P=0.006), indicating a shared genetic etiology between ASD and schizophrenia[46]. The fact that genes with LGD mutations in ASD are associated with regulatory variants in schizophrenia suggests that complete abrogation of these genes may cause developmental defects as in ASD, while regulatory changes in these genes may cause later-onset of neuropsychiatric symptoms as in schizophrenia. *Collectively, these preliminary data demonstrate that we can conduct and analyze genome-wide Hi-C experiments, integrate these data with other epigenetic and transcriptomic data, and use chromatin architecture elucidated by Hi-C to provide novel genome-wide insights into the regulatory mechanisms occurring during neuronal differentiation and disease pathogenesis.*

***Experimental design and methods:*** In ***subaim 1.1.,*** *we will profile the transcriptome (by RNA-seq), cis-regulatory elements (ChIP-seq) and 3D chromatin architecture (Hi-C)* in the control neurotypical dorsolateral prefrontal cortex (dlPFC), posterior superior temporal cortex (pSTC) and striatum (STR). These regions have been implicated in the risk for ASD and schizophrenia[35] and in the cases of dlPFC and pSTC shown to have dysregulated transcriptional patterns in ASD[9]. Recent studies have also highlighted the late mid-fetal frontal cortex as most enriched for co-expression of ASD and schizophrenia *de novo* hits[15,31,35]. Brains from at least 5 key epochs of development representing mid-fetal, infancy, childhood, adolescence and adult brain, and a minimum of 6 subjects (balancing sex when possible) from each of these 5 epochs (30 brains in total) will be profiled.

Cell-type specific chromatin, epigenetic and transcriptome assays are at the core of this project. Mario Skarica, a talented research associate scientist in the Sestan lab, has developed a protocol to isolate high quality nuclei with preserved chromatin and RNA from archival fresh frozen fetal and postnatal human brains. Using this approach he has obtained, on average, 2.57 +-0.8 and 6.93+-3.3 million intact nuclei from 100 mg of the fetal or adult prefrontal gray matter (i.e., fetal CP or adult cortical layers 1 to 6 with a small part of underlying white matter), respectively (**Figure 5A**). Furthermore, we separated neuronal and non-neuronal nuclei, by immunostaining with the NeuN antisera against pan-neuronal splicing protein RBFOX3 (**Figure 5B and C**) and sorting on BD FACSAria IIU Three-Laser System. Starting with infancy and onwards, postnatal gray matter tissue corresponding to six-layered postnatal cortex and small part of adjacent white matter from dlPFC and pSTC, or STR (corresponding to the caudate-putamen with the internal capsule at the septal level) will be processed.

Tissue samples will be dissected directly from frozen tissue blocks using custom dental tools and protocol described in Kang et al., 2011[47]. These dissections will be performed by Nenad Sestan, who has over 2 decades of experience in human neuroanatomy and tissue processing and has microdissected over 1600 tissue samples for exon array profiling of the human brain transcriptome[47]. Given the high proportion of neurons in the cortical plate of the mid-fetal brain (approaching 95% or more), and relatively few neurons that are positive for NeuN at 17-20 PCW in neocortical CP or STR[48], we will not sort NeuN+ and NeuN- nuclei from mid-fetal brains, but instead analyze tissue homogenate and unsorted nuclei from CP of prospective dlPFC and pSTC as well as STR, separately, from corresponding neocortical and striatal GZ (i.e., VZ and SVZ) containing a mixed population of dividing neural stem/progenitor cells with a minor contribution of newborn neurons and glia.

Tissue samples will be pulverized and processed to release nuclei, which will be purified by ultracentrifugation and processed for RNA-seq in the case of mid-fetal samples or in the case of all postnatal specimens (infancy and onwards) sorted into a NeuN+ (predominantly neurons) and NeuN- (mostly glia) fractions. In the past year, we have obtained on average 23.45+-7.2 percentage of NeuN+ nuclei from PFC (**Figure 5C**). This approach will provide unbiased quantitative assessments of cell types in healthy and ASD brains. This approach allows us to simultaneously collect molecularly defined cell type-specific nuclei and isolate DNA, chromatin, and nuclear RNAs. Bulk tissue level RNA-seq is available for dlPFC, pSTC and STR in control and ASD brains as part of psychENCODE phase 1 studies[38], has already been added to enhance the scope of the resource. All brains necessary for this project are currently available in the Geschwind and Sestan labs (see Facilities and Resources section for the list).



| | Tissue (dlPFC) | All nuclei | NeuN+ nuclei | NeuN- nuclei |
|---|---|---|---|---|
| 🟩 Intergenic RNA | 7.5% | 18.9% | 11.2% | 12.5% |
| 🟥 Intronic RNA | 42.6% | 49.1% | 58.6% | 64.0% |
| 🟦 Coding RNA | 49.9% | 32.0% | 30.2% | 23.5% |

**Figure 5. Fluorescence-activated nuclei sorting (FANS) and nuclear RNA-seq of human dlPFC.** ***A.*** *Collection of single nuclei (see insert) from fetal (n=6) and adult (n=29) PFC.* ***B.*** *FANS plot for NeuN immunopositive nuclei.* ***C.*** *Percentage of NeuN+ nuclei collected across different experiments.* ***D.*** *Coverage for exon, intron and intergenic regions of different sequencing technologies.* ***E)*** *Differential expression comparison between NeuN+/- FANS nuclei for neuronal, glial and ASD-related genes.*

Total RNA will be extracted from 1 million nuclei using Norgen's Cytoplasmic & Nuclear RNA Purification Kit. RNA from tissue and cell populations will be depleted of rRNA and sequencing libraries prepared with TruSeq Stranded Total RNA with Ribo-Zero Gold and SMARTer Stranded RNA-Seq Kit, respectively. As expected, our preliminary nuclear RNA-seq analyses revealed higher percentage of unspliced primary transcripts and extensive identification of nuclear-retained long non-coding RNAs (**Figure 5D**). Importantly, we detected robust cell type-specific expression differences, including those of ASD-associated genes (**Figure 5 E)**. RNA-seq libraries will be sequenced on the Illumina HiSeq 2500 at the Yale Center for Genome Analysis (http://ycga.yale.edu/) to generate 100 bp strand specific paired-end sequence at over 40 million reads per end for each sample. For ChIP-seq, 1 million nuclei will be processed through our established protocol using well-characterized ChIP-grade H3K27ac and H3K4me3 antibodies that have been used in psychENCODE phase 1 tissue-level experiments. ChIP-seq libraries will be sequence at HiSeq 2500 at Yale at >40 million reads per sample. Using the standard pipelines developed in the Sestan and collaborating labs, we will perform QC analyses and compare the transcriptome and epigenetic data from different time points and regions to construct spatiotemporal gene and disease state profiles and co-expression networks using computational methods described in Aim 2.

For Hi-C, 2 million nuclei will be prepared from each sample and cross-linked in 1% formaldehyde for 10 min. Cross-linked DNA will then be restriction digested using HindIII, digested chromatin ends filled with biotin-14-dCTP, and resulting blunt-end fragments ligated under dilute conditions to minimize random intermolecular ligations. Following this, crosslinking will be reversed, unligated ends removed by exonuclease digestion (T4), DNA sheared by sonication, and 300-600bp fragments selected. The intermolecular ligation products containing biotin-tagged DNA will be pulled down with streptavidin beads and ligated with Illumina paired end adapters and the library sequenced by Illumina 50bp paired-end sequencing over 3 lanes of the HiSeq 25000 at UCLA, a depth necessary to facilitate sufficient hi-resolution analysis (300-500 million mapped reads), which can also be augmented by pooling samples to increase depth as needed.

In ***subaim 1.2****, complementary genomic analyses will be done on the FANS nuclei from control, and syndromic and idiopathic ASD brains*, to identify transcripts, regulatory elements, and 3D chromatin structures altered in ASD in brain region and cell type-specific manners. We will conduct RNA-seq, ChIP-seq and Hi-C on sorted neuronal and non-neuronal nuclei from 2 cortical regions, dlPFC and pSTC, and STR from 20 matched control and 20 ASD individuals, including 5 dup15q cases. We will select 10 ASD cases manifesting the shared pattern of transcriptional dysregulation observed, 10 without this pattern, and match them to controls to account for potential confounders (sex, age, postmortem interval [PMI], and RNA integrity numbers [RIN]). We will select 5 dup15q brains with most similar breakpoint structures. Hi-C will be performed on sorted nuclei using the identical experimental methods as in subaim 1.1.

***Pitfalls and alternatives:*** The techniques in these proposed experiments are commonly used in our laboratories and we do not expect complications. One potential issue is the obtainment of adequate samples. The Sestan lab has almost 200 high quality frozen human prenatal, early postnatal and adult brain specimens from clinically unremarkable (neurotypical) control donors. Control brains from this collection were used for different BrainSpan and psychENCODE phase 1 projects (see example studies[12,47,49,50] and Resources and Facilities section). Both Geschwind and Sestan labs have tissue samples from over 50 post mortem ASD cases and matched controls with good quality RNA, and have participated in the new initiative at the Simons Foundation to collect additional postmortem ASD brains. A related concern is whether the 20 ASD brains we propose to analyze are sufficient, given the heterogeneity typical of ASD, to detect robust differences between these samples and our controls. However, we were able to detect transcriptional dysregulation in 2/3rds of ASD brains in a smaller cohort17, and by directly comparing ASD brains exhibiting hallmarks of dysregulated transcription with those that do not, we expect to have sufficient statistical power to assess the extent to which 3D chromatin structure contributes to the observed transcriptional changes. Further, the use of 5 dup15q cases provides an additional homogeneous cohort, and as our preliminary results on transcriptome analysis of this cohort demonstrate (appendix), such sample sizes are sufficient. The main pitfall of Hi-C is that it averages chromosome contact population from millions of nuclei. Single-cell Hi-C can complement this limitation[51], but it can capture only one interaction for a given locus. Homogenous population of cells can be achieved by FANS and thus we propose this approach here. Additionally, Hi-C offers other benefits, including the ability to analyze interactions mediated by multiple TFs *en masse* in Hi-C, that are not easily achievable with other methods such as ChIA-PET. While our FANS approach, which follows standards accepted across the psychENCODE projects, is limited to two major groups of cells, we have been implementing the use of other cell type specific nuclear antibodies and single nuclear RNA-seq. Finally, we realize that other regions, including the thalamus, hypothalamus, and hippocampus, may be affected in ASD. We believe our work on the neocortex and STR will develop a framework for understanding of the molecular neuropathology of ASD which can then be extended to include other regions in the future.

**Aim 2. Integrated analyses of transcriptome, epigenome and chromatin structure in control and ASD brains.**

***Rationale:*** We will analyze the data generated in the previous aim to (1) identify *developmentally regulated and cell type specific changes* in the transcriptome, epigenome and the 3D chromatin structure (2) integrate the three types of datasets to gain comprehensive insights into the underlying mechanisms of transcriptional regulation and dysregulation in development and disease, respectively.

***Experimental design and methods*:** In ***subaim 2.1****, several first order analyses will be done for quality control and to provide the data as a processed resource in addition to the raw data*. We will use Illumina CASAVA to purify the low-quality and non-identified reads and Fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), to report fundamental quality parameters. Next, Tophat[52] will be employed to uniquely align the filtered reads to their reference genome and RSEQtools[53] to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. The R package DESeq (http://bioconductor.org/packages/release/bioc/html/DESeq.html) will be used to identify differentially expressed (DEX) genes and well established methods including MATs to identify differential splicing[10,37]. DEX genes will be detected from the reliably expressed coding and non-coding transcripts, which are defined as transcripts with RPKM ≥ 1 in at least 2 samples of different developmental period. ChIP-seq reads will be aligned to the genome by Bowtie. After filtering of low score reads, we will use the MACS platform to call peaks enriched over the input library, and peaks with high empirical FDR will be excluded from further analysis. Thus, we will catalog all potential cis-regulatory elements from our genome-wide histone modification maps in all brain regions across developmental periods.

*For Hi-C analysis*, *hiclib* (https://bitbucket.org/mirnylab/hiclib) will be used to perform all initial analysis on Hi-C data from mapping to filtering and bias correction (see also[40]). Sequenced reads will be mapped to the human

genome by *Bowtie2* (with increased stringency, *--score-min -L 0.6,0.2--very-sensitive*) through iterative mapping and read pairs allocated to HindIII restriction enzyme fragments. Self-ligated and unligated fragments, fragments from repeated regions of the genome, PCR artifacts, and genome assembly errors will be removed. Filtered reads will be binned at 10kb, 40kb, and 100kb resolution to build a genome-wide contact matrix at a given bin size. This contact map depicts contact frequency between any two genomic loci. To decompose biases from the contact matrix and yield a true contact probability map, filtered bins are subjected to iterative correction[41]. Bias correction and normalization results in a corrected heatmap of bin-level resolution. 100kb resolution bins are assessed for inter-chromosomal interactions, 40kb for TAD analysis, and 10kb for gene loop detection. For TAD-level analysis[32], we will quantify the directionality index by calculating the degree of upstream or downstream (2Mb) interaction bias of a given bin, which will be processed by a hidden Markov model (HMM) to remove hidden directionality bias. For gene loop detection, aggregate peak analysis (APA) will be performed that quantifies the aggregate enrichment of putative peak sets by calculating the sum of a series of submatrices derived from a contact matrix[34]. Resulting inter- and intra-chromosomal interaction matrices as well as genome-wide TADs and gene loops will be used for integrative analysis.

*Developmental and cell type-specific changes*: Pearson's correlations between the first principal components (PC1) from different stages and neuronal and non-neuronal cell types, as well as with our own and other published data will be calculated to compare similarities between different cell types. We will explore alternative transcriptional mechanisms or post-transcriptional modifications occurring in normal (and ASD-affected, see below) regions/cells and time points. These can include up- or down-regulating expression, altered spatiotemporal gene expression, imbalanced expression of different alleles (allele-specific expression [ASE]), aberrant splicing events, modified RNA editing sites, fusion transcripts, or loss of function due to frameshift mutations. RNA and epigenome data will also be compared with tissue level psychENCODE phase 1 and BrainSpan's RNA-seq and ChIP-seq data. We will follow up with an analysis of the relative enrichment of each cell-specific marker genes in each subpopulation and use the expression profiles of these genes to guide the identification of an expanded set of cell-type specific markers.

*Integrated analyses*: Spearman's correlations between PC1/PC2 and biological traits (gene expression, histonemark enrichment, GC content, gene density, DNase I hypersensitivity [DHS]) will be calculated. Gene expression and histone mark data generated in subaim 1.1 along with DHS of fetal brain from Epigenomic roadmap[54] will be used and average values per 100kb bin calculated. In addition to the putative cis-elements identified in the same samples, we will also use the 15 state epigenetic marks from Epigenomic Roadmap[54] in genomic regions classified based on compartments averaged across 40kb bins, as well as subject specific psychENCODE data. Epigenetic state counts[54] for one compartment category are normalized by total epigenetic mark number of that compartment category and compared between samples.

*Dysregulation in ASD brains*. Two main data analyses will be performed with the transcriptome data. We will use the same approach as in subaim 2.1 to identify DEX coding and non-coding transcripts (by DESeq) between ASD and matched controls. Gene function enrichment analysis will be performed for these DEX genes. Finally, we will also perform Weighted Gene Co-expression Network Analysis (WGCNA; http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/) to identify modules of differentially co-expressed genes in ASD cases. For ChIP-seq data, once peaks are called and filtered for quality and reproducibility, we will identify and catalog all putative enhancer and promoter sites gained or lost in ASD brains compared to matched control brains, as well as what genes they are associated with.

In Hi-C data, we will assess if compartments, TADs, and contact domain structures are abrogated in ASD brains. Interaction partners for ASD risk genes, as well as gene loops involving ASD risk gene regulatory elements will be examined. Genome-wide interchromosomal contact matrices at high resolution (approximately 10kb) will be compared between ASD and control to identify bins that exhibit the largest chromosomal interaction changes in ASD (here we refer to ASD-specific bins). Gene ontology for these genes as well as their gene expression pattern in ASD may provide novel insights on ASD mechanism. The same approach will be applied to intrachromosomal contact matrices at 10-40kb resolution. TADs in ASD vs. controls will be also compared. The directionality index around ASD-specific TAD boundaries will be calculated to check significance. Moreover, we will examine gene expression level and histone marks on TAD boundaries as well as histone marks on TADs that contain ASD risk genes. Both inter- and intra-chromosomal interaction patterns of the bins that contain differentially expressed genes in ASD or ASD risk genes will be examined in ASD vs. controls. Gene expression pattern and histone states of genomic loci that highly interact with dup15q region will be assessed. *This approach of integrating chromosome interactomes to transcriptomic and epigenetic profiles may delineate epigenetic mechanism behind gene dysregulation in ASD.*

We will also perform integrative network analyses of these multi-level genomic data with genetic variation to understand the causal mechanism of transcriptional alterations in ASD (see also Aim 3). This will include integration of DNA sequence, methylation, chromatin contacts, eQTL and hQTL by this collaborative team of investigators (e.g. to include new hQTL methods by S. Prabhakar and colleagues[55]. Gene loops detected in control and ASD will be also interrogated. Gene loops that are specific to ASD or specific to controls may directly point out aberrant enhancer-promoter interactions, TF binding, or compartmentalization of genome. We will check if ASD-specific gene loops contain any ASD-associated variants (mostly common SNV at this point, although as more whole genome sequencing (WGS) data is available over the next 12 months, we can use these data to annotate potential functions of noncoding variants (Aim 3).

In ***subaim 2.2.**, we will integrate and harmonize data across psychENCODE projects and other relevant genomic resources*. In this aim, the DAC will integrate and harmonize our datasets with other psychENCODE studies and large-scale genomic datasets, such as BrainSpan, CommonMind, ENCODE, GTEx and REMC. The PsychENCODE DAC is led by Mark Gerstein and Nenad Sestan (Yale), Zhiping Weng (University of Massachusetts), who are part of this proposal and Kevin White (University of Chicago). DAC will summarize the major analysis results produced from psychENCODE and organize them into an encyclopedia of regulatory elements in the developing and adult human brain. We are currently building such an encyclopedia for the ENCODE consortium, and we will be able to leverage the methods that we are building for ENCODE and modify them to best serve psychENCODE data. The psychENCODE encyclopedia will include several components. The first component is the raw experimental data, including the expressed transcripts in neuronal and glial cells in various brain regions, the peaks (enriched regions) of an array of histone marks, the open chromatin regions detected using ATAC-seq, the differentially enriched histone mark peaks and open chromatin regions in ASD, BD and SCZ (diseases covered by psychENCODE projects. This component will largely result from a series of uniform processing pipelines, which we will build for analyzing psychENCODE data. The second component will include results that require the integration across multiple data types, including the enhancers in each cell type, the chromatin states called using a combination of histone marks and ATAC-seq data, and the topologically associated domains and compartments called by combining histone marks, ATAC-seq and Hi-C data. The third component of the encyclopedia will provide a higher-order organization to the elements in the first two components. Specifically we will derive the target genes for enhancers in a cell type specific manner, and identify the enhancer-gene links that are disrupted in the three diseases. We will also identify the variations that are linked with difference in gene expression (eQTLs) that are within enhancers that target the corresponding genes. Finally, we will develop a portal to guide the user through the components of the psychENCODE encyclopedia, with multiple entry points, such as genes, GWAS SNPs, or a specific regulatory region in the genome.
***Pitfalls and alternatives:*** Proposed computational approaches are well established in our team and we already have a considerable expertise and collaborative history therefore we foresee no complications in performing this aim. Furthermore, Sestan, State and Geschwind have been part of the BrainSpan project and Ernst, Gerstein and Weng has been part of several other relevant genomic consortia, such ENCODE.

**Aim 3. Spatiotemporal analysis in ASD.**
***Rationale and preliminary supporting data:*** Over the past few years genomic analyses by our labs and others have made rapid progress in identifying genes associated with ASD, in particular through the identification of *de novo* mutations in ASD cases[13,17,26,30,45]. Despite the identification of these ASD-associated genes, progressing to an understanding of ASD neurobiology remains a challenge. Aims 1 and 2 described one approach to discovering this neurobiology through the identification of ASD-specific networks in *post mortem* brains. In Aim 3 we propose a complementary approach through the identification of genomic loci, brain regions, developmental stages, cell types, and neurobiological processes that are enriched for ASD mutations in genes (**subaim 3.1**) and non-coding loci (**subaims 3.2 and 3.3**) in neurotypical brains. Finally, we will test the hypothesis that ASD specific networks observed in *post mortem* brains from Aims 1 and 2 will be enriched for ASD associated mutations (**subaim 3.4**) thus demonstrating that the disruption of this network precedes the diagnosis of ASD and is therefore likely to be a cause of ASD rather than a consequence.

*1) Detection of ASD-associated genetic loci.* We identified rare and *de novo* variants in exome data from 5,563 ASD cases and 13,321 controls alongside rare and *de novo* copy number variants in microarray data from 4,687 ASD cases and 2,100 controls[17]. Comparison of these two data sets showed that small *de novo* deletions in ASD targeted the same set of genes as *de novo* loss of function point mutations in exome data. A combined analysis of exome data and small *de novo* deletions was performed using the Transmitted and *De novo* Association (TADA) method to identify ASD-associated genes. 28 ASD-associated genes were identified with very high confidence (false discovery rate (FDR) ≤ 0.01) and 65 ASD-associated genes were identified with high

confidence (FDR ≤ 0.1). These 65 genes formed a protein-protein interaction (PPI) network with two distinct subnetworks, enriched for chromatin regulatory genes and synaptic genes respectively (**Figure 6A**).

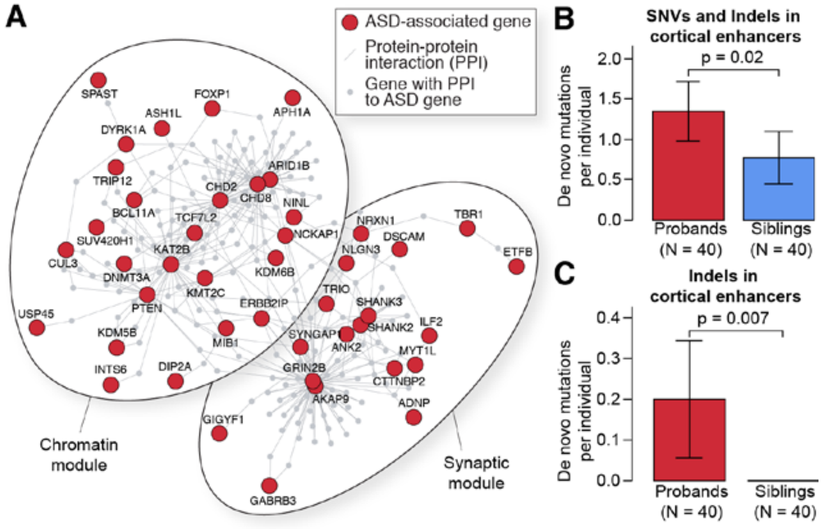


**Figure 6. ASD associated de novo mutations. A.** *65 ASD risk genes[9] (red) form a single protein-protein interaction network composed of two subnetworks. The genes in the left subnetwork are enriched for chromatin regulatory gene ontology terms. The genes in the right subnetwork are enriched for synaptic terms.* ***B.*** *De novo mutations were identified in WGS data for 40 ASD families. The median number of SNV and indel mutations per individual is shown within active enhancers that were identified by bulk tissue ChIP-Seq for H3K27ac in human dlPFC (psychENCODE phase 1 studies). P-values are calculated using linear regression with for paternal age and total de novo mutations per individual included as co-variates.* ***C.*** *The analysis was repeated for indels only.*

*2) Detection of ASD-associated non-coding variants in whole-genome sequencing (WGS) data*. We analyzed WGS data for 40 simplex ASD quartets composed of both parents, an affected child and an unaffected sibling control. The families were selected from the Simons Simplex Collection on the basis of no previous *de novo* loss of function or CNV mutations in exome and microarray data and high paternal age. The samples were sequenced to greater than 30x mean coverage (mean±standard 35.7±5.8). Raw data were aligned to hg19 human reference genome using BWA-mem[56]. Duplicate reads were removed with Picard (http://broadinstitute.github.io/picard/); GATK best practices[57] were used for all downstream steps including, local realignment, base quality score recalibration, SNV and indel calling, cohort-wide joint genotyping, and variant quality score recalibration. Data were normalized within families by only analyzing bases with at least 20 unique reads in all family members. A combination of PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq/) and in-house scripts were used to identify autosomal *de novo* variants based on stringent criteria designed to maximize specificity: minimum genotype likelihood (GQ) ≥20, alternate allele frequency (AB) ≤0.05 in the parents, and 0.3-0.7 in the child, minimum map quality (MQ) ≥30 in all family members, and allelic depth for the alternate allele (AD) ≥8. Approximately 7,000 *de novo* mutations were identified at a rate of 87.0±13.5 *de novo* mutations per child. Confirmation with Sanger sequencing was attempted on 10% of these variants (700) selected at random and achieved a >95% confirmation rate across both SNVs and indels, suggesting identification of *de novo* mutations with accuracy. We used tissue-level ChIP-seq for the histone modification H3K27ac from human dlPFC (psychENCODE phase 1) to identify active enhancers. We observed an increased burden of mutations in cases compared to sibling controls (p=0.02, **Figure 6B**) within these active enhancers. This association was especially strong for insertion/deletions (indels), possibly due to the greater functional impact of disrupting multiple nucleotides (p=0.007, **Figure 6C**).

*3) Analysis of gene co-expression to identify spatiotemporal convergence of ASD-associated genes*. We considered the convergence between 9 ASD genes[15] for gene expression data from 57 neurotypical brains that spanned 15 developmental periods and 16 brain regions[47]. To identify spatiotemporal windows whilst retaining sufficient numbers of samples for co-expression analysis we used hierarchical clustering to identify four groups of brain regions and considered each of these in 13 overlapping time periods each composed of three developmental periods (**Figure 7A**). Within each of the resulting 52 (4 x 13) spatiotemporal windows we built networks around nine high confidence ASD genes by selecting the top 20 co-expressed genes. We assessed these 52 windows for spatiotemporal convergence related to ASD etiology through the degree of enrichment for 126 independent low confidence ASD genes (**Figure 7A**). We observed strong spatiotemporal convergence between ASD risk genes in the prefrontal and primary motor-somatosensory cortex during mid-fetal development (**Figure 7A**)[15]. Analysis of cell type specific marker genes within this network showed enrichment for cortical projection neurons. This result that has been replicated by three complementary techniques: WGCNA[31], cell specific enrichment analysis[58], and NETBAG+ systems analysis[59].

*4) Comparison of ASD-related gene sets and gene expression analysis of post-mortem ASD brains.* Two prior analyses have identified gene co-expression WGCNA modules that are differentially expressed in the brain in ASD cases compared with controls. The microarray analysis by Voineagu et al.[9] identified a module enriched

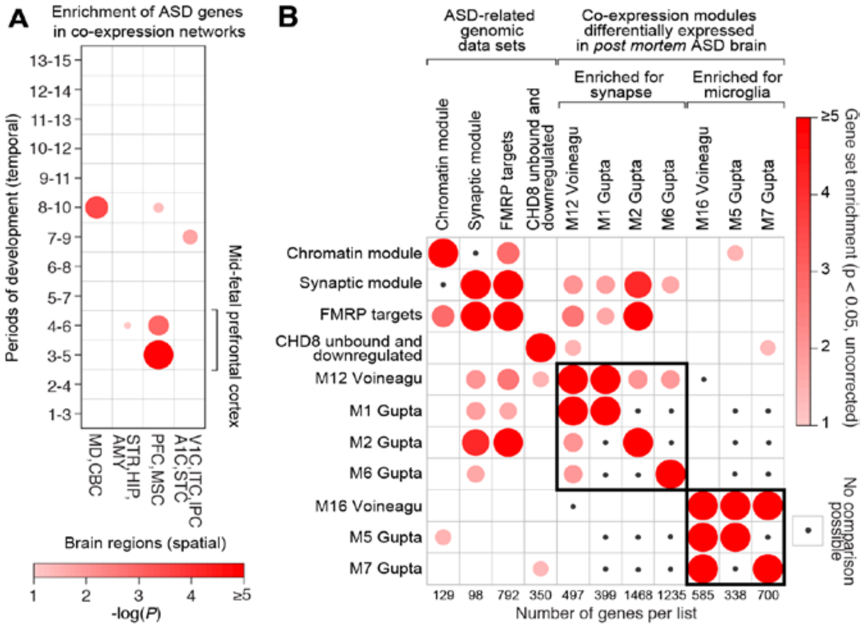


***Figure 7. Enrichment of ASD-associated genes in gene expression data. A.*** *Spatiotemporal co-expression networks were formed around nine high confidence ASD genes for 4 groups of brain regions (x-axis) and 13 overlapping developmental periods (y-axis). The –log(P) value for enrichment with 126 low confidence ASD genes is shown by the size and shade of the circle. Strong enrichment is observed in the mid-fetal PFC and primary motor-sensory cortex (PFC-MSC).* ***B.*** *Four ASD related gene sets*[9,52,53] *compared to seven WGCNA co-expression modules that are differentially expression in post mortem ASD brains (right). Fold enrichment is indicated by the size and shade of the circle. A synaptic and microglial module are indicated by the black rectangles. Small black circles show gene sets that are non-overlapping by definition (e.g. WGCNA modules in the same analysis).*

for synaptic genes (M12) that overlaps with three modules (M1, M2, and M6) identified using RNA-seq in Gupta et al[60] (**Figure 7B**). Similarly, one module enriched for microglial genes (M16) was observed in the Voineagu et al.[9] paper and this overlaps with two modules (M5 and M7) identified in the Gupta analysis (**Figure 7B**). We compared these seven WGCNA modules with four sets of ASD-related genes: the chromatin and synaptic modules from our recent analysis of exome and CNV data (**Figure 6A**)[17], RNA targets of the fragile X protein FMRP[61] that are strongly enriched within ASD-associated genes[26], and genes that are downregulated in CHD8 knockdown but not bound by CHD8 on ChIP-Seq analysis that have been described as targeting synaptic genes associated with ASD[62]. The synaptic module and FMRP targets are strongly enriched through the synaptic WGCNA modules suggesting these modules may represent causal factors that persist in the ASD brain. Further analysis is required to determine if these modules are causal or simply a consequence of ASD.

***Experimental design and methods:*** In ***subaim 3.1**, we will increase the spatiotemporal resolution of co-expression analysis of ASD neurobiology*. Our prior analysis of spatiotemporal convergence, described in detail under preliminary data[15], was based on 57 neurotypical brains, 9 high confidence ASD genes (FDR ≤0.05), and 126 low confidence ASD genes (FDR ≤0.3)[17]. These data enabled us to examine 4 groups of brain regions spanning multiple developmental periods (**Figure 7A**). The data from Aim 1 and our progress in ASD gene discovery will allow us to perform this analysis using 87 neurotypical brains, 28 high confidence ASD genes (FDR ≤0.01), and 151 low confidence ASD genes (FDR ≤0.3). As before (**Figure 7A**), the gene expression samples will be divided into spatiotemporal windows using hierarchical clustering to group related brain regions (spatial) and considering overlapping developmental windows (temporal). In each spatiotemporal window we will identify the top 20 co-expressed genes around 28 high confidence ASD genes and, following the logic that a spatiotemporal network relevant to ASD should be enriched for other ASD genes, we will assess the enrichment of the 151 low confidence ASD genes (FDR ≤0.3). The expanded number of brain samples will enable us to use small subdivisions of brain regions and developmental time regions to increase the resolution of the analysis, for example windows spanning one or two developmental periods. In addition, the larger list of high confidence ASD genes will allow us to perform the analysis by building the spatiotemporal networks around subsets of these 28 genes and improve the accuracy of the analysis through cross validation. In addition, we will divide the 28 high confidence genes by the two main functional categories observed, specifically chromatin regulators and synaptic genes, to assess the spatiotemporal dynamics of each functional category separately. The outcome of this aim will be refined gene co-expression networks that show spatial and temporal convergence among ASD risk genes. ***Pitfalls and alternatives:*** The analytical methods described here have been applied to the BrainSpan data using 9 high confidence genes resulting in the discovery of spatiotemporal convergence in the frontal cortex of the mid-fetal brain. This finding has been replicated using complementary methods[58,59]. In this aim we will be increasing the resolution through the inclusion of additional gene expression data and novel ASD-associated genes[17], therefore we do not foresee complications. An alternative 'top down' methodology such as WGCNA, in which co-expression modules are generated from the complete dataset and are then assessed for enrichment

of ASD genes, has yielded similar findings[31]. We will also apply this complementary WGCNA method across spatiotemporal windows.

In ***subaim 3.2**, we will identify ASD-associated non-coding de novo mutations in regulatory loci*. Under pre-existing funding arrangements we will have access to whole-genome sequencing (WGS) data for 5,120 individuals from 1,280 quartet families composed of two parents, an affected child, and an unaffected sibling control. We have previously reported an increased burden of *de novo* mutations between the affected and unaffected siblings[17] and we have observed this for *de novo* CNVs in microarray data and *de novo* loss of function mutations in exome data. To identify functional non-coding *de novo* mutations in regulatory loci, we will leverage the integrated RNA-Seq, ChIP-Seq, and HiC data from Aims 1 and 2 with the *de novo* mutation identification approach described in our preliminary data (**Figure 6**). To maximize our ability to discover compartments of the genome that carry risk we will assess *de novo* burden in three sets of loci: **1**) All regulatory loci identified in neurotypical brain divided by function (e.g. promoter, 3`UTR); **2**) Regulatory loci identified in neurotypical brain with a relationship to 28 high-confidence ASD genes; and **3**) Regulatory loci identified in neurotypical brain with a relationship to the points of convergence for ASD genes identified in Subaim 3.1 such as prefrontal cortex in mid-fetal development. The outcome of this aim will be non-coding mutations and regulatory loci that show association with ASD.
***Pitfalls and alternatives:*** Our methods for identifying *de novo* mutations in whole genome sequencing data are well developed and we have demonstrated a >95% confirmation rate for the mutations predicated. Additionally, our preliminary data, based on 40 families, shows evidence of ASD association for *de novo* mutations within enhancers active in human dlPFC (**Figure 6B and C**). This suggests the proposed study of 1,280 families will offer sufficient power even if the overall contribution of *de novo* mutations in the non-coding genome to ASD etiology is relatively weak. To maximize our chance of identifying ASD associated non-coding variants we will assess only the loci with the strongest evidence of functional activity, including the larger mutations, such as indels, that may carry the greatest risk. Concurrently, Dr. Sanders has an established collaboration with Mike Talkowski and the GATK CNV/SV working group to develop methods that maximize our sensitivity for detecting indels and small CNVs in whole genome sequence data.

In ***subaim 3.3**, we will identify points of spatiotemporal convergence using ASD associated non-coding mutations:* Non-coding elements such as enhancers frequently show a degree of specificity to particular developmental time points, brain regions, or cell types[63]. We will use the ASD-associated non-coding *de novo* mutations in regulatory loci and regulatory loci related to ASD associated genes to assess which integrated regulatory networks from Aim 2 show the greatest enrichment for these non-coding mutations. By considering the brain regions and developmental epochs in which these networks exist we will assess points of spatiotemporal convergence critical to ASD. The outcome of this aim will be an independent analysis of points of spatiotemporal convergence in ASD based on non-coding mutations and regulatory loci.
***Pitfalls and alternatives:*** This aim relies on the discovery of specific ASD-associated regulatory loci through the discovery of numerous *de novo* mutations in cases. Due to the small size of regulatory regions we may not see this clustering in a single regulatory element. Should this be the case we will use genomic annotation to rank the regulatory loci with a single mutation, for example considering conservation, constraint [64], and large mutations such as indels that are more likely to disrupt the element (**Figure 6C**).

In ***subaim 3.4**, we will assess regulatory networks that are observed in the post mortem ASD brain.* Aims 3.1 to 3.3 focus on neurotypical brains and their association with ASD-associated mutations. In this aim we will assess the enrichment of ASD-associated genes, non-coding mutations and regulatory networks that differ between *post mortem* ASD and neurotypical brains (**Figure 6**). Because genetic variants associated with ASD precede the onset of ASD symptoms, enrichment for these mutations will suggest that such networks are causal (**Figure 7**) to the ASD phenotype. Conversely, a lack of enrichment for these mutations in ASD-relevant networks will suggest the network is consequential to ASD. The outcome of this aim will therefore be to distinguish ASD-specific regulatory networks that are likely to be causal from those that may be consequential.
***Pitfalls and alternatives:*** Methods for assessing such enrichment are well established and we already have a large list of ASD-associated genes; we foresee no complications in performing this aim. The main challenge lies in the interpretation of a regulatory network that does not enrich for ASD-associated genes (e.g. microglia in existing *post mortem* analyses, **Figure 7B**), since this may indicate a non-causal relationship or reflect and incomplete list of ASD-associated genes. We will therefore focus on networks with positive enrichment for these genes and acknowledge the complexities of interpreting a negative result.

**TIMELINE AND MILESTONES SECTION** See Other Attachments

## BIBLIOGRAPHY

1. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature reviews. Genetics* **13**, 537-551 (2012). pmcid: 4110909.
2. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics* **45**, 984-994 (2013). pmcid: 3800159.
3. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. *Science* **349**, 1489-1494 (2015). pmcid:
4. Krystal, J. H. & State, M. W. Psychiatric disorders: diagnosis to therapy. *Cell* **157**, 201-214 (2014). pmcid: 4104191.
5. Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nature neuroscience* **17**, 764-772 (2014). pmcid: 4077789.
6. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of *de novo* mutations in the genetics of autism spectrum disorders. *Nature reviews. Genetics* **15**, 133-141 (2014). pmcid:
7. Walsh, C. A., Morrow, E. M. & Rubenstein, J. L. Autism and brain development. *Cell* **135**, 396-400 (2008). pmcid: 2701104.
8. Huguet, G., Ey, E. & Bourgeron, T. The genetic landscapes of autism spectrum disorders. *Annu Rev Genomics Hum Genet* **14**, 191-213 (2013). pmcid:
9. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384 (2011). pmcid: 3607626.
10. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature reviews. Genetics* **16**, 441-458 (2015). pmcid:
11. Abelson, J. F. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317-320 (2005). pmcid:
12. Johnson, M. B. *et al.* Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**, 494-509 (2009). pmcid: 2739738.
13. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012). pmcid: 3667984.
14. State, M. W. & Sestan, N. Neuroscience. The emerging biology of autism spectrum disorders. *Science* **337**, 1301-1303 (2012). pmcid: 3657753.
15. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013). pmcid: 3995413.
16. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat Commun* **6**, 6404 (2015). pmcid: 4355952.
17. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233 (2015). pmcid: 4624267.
18. Luo, R. *et al.* Genome-wide transcriptome profiling reveals the functional impact of rare *de novo* and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet* **91**, 38-55 (2012). pmcid: 3397271.
19. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* **14**, 1109-1120 (2015). pmcid:
20. Preuss, T. M. Human brain evolution: from gene discovery to phenotype discovery. *Proc Natl Acad Sci U S A* **109 Suppl 1**, 10709-10716 (2012). pmcid: 3386880.
21. Lui, J. H. *et al.* Radial glia require PDGFD-PDGFRbeta signalling in human but not mouse neocortex. *Nature* **515**, 264-268 (2014). pmcid: 4231536.
22. Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647 (2013). pmcid: 3922239.
23. Zeng, J. *et al.* Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* **91**, 455-465 (2012). pmcid: 3511995.
24. Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213-217 (2009). pmcid: 2778075.
25. The PsychENCODE Consortium *et al.* The PsychENCODE project. *Nature neuroscience* **Manuscript accepted** (2015). pmcid:

26. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014). pmcid:

27. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246-250 (2012). pmcid: 3350576.

28. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012). pmcid: 3613847.

29. Sanders, S. J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011). pmcid: 3939065.

30. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622 (2012). pmcid: 3528801.

31. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021 (2013). pmcid:

32. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012). pmcid: 3356448.

33. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009). pmcid: 2858594.

34. Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680 (2014). pmcid:

35. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518-529 (2013). pmcid: 3894107.

36. Network & Pathway Analysis Subgroup of Psychiatric Genomics, C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209 (2015). pmcid: 4378867.

37. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523 (2014). pmcid: 4390143.

38. Parikshak, N. N. *et al.* Global changes in patterning, splicing and primate specific lncRNAs in autism brain. *Manuscript in preprint. Please see appnedix.* (2015). pmcid:

39. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010). pmcid: 3003124.

40. Won, H. *et al.* Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development. *Manuscript in preprint. Please see appnedix.* (2015). pmcid:

41. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003 (2012). pmcid: 3816492.

42. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* **28**, 1045-1048 (2010). pmcid: 3607281.

43. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nature genetics* **45**, 723-729 (2013). pmcid: 3932982.

44. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428 (2014). pmcid: 4208299.

45. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014). pmcid:

46. McCarthy, S. E. *et al. De novo* mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658 (2014). pmcid: 4031262.

47. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011). pmcid: 3566780.

48. Sarnat, H. B., Nochlin, D. & Born, D. E. Neuronal nuclear antigen (NeuN): a marker of neuronal maturation in early human fetal nervous system. *Brain Dev* **20**, 88-94 (1998). pmcid:

49. Pletikos, M. *et al.* Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron* **81**, 321-332 (2014). pmcid: 3931000.

50. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206 (2014). pmcid: 4105188.

51. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013). pmcid: 3869051.

52. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009). pmcid: 2672628.

53. Habegger, L. *et al.* RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**, 281-283 (2011). pmcid: 3018817.

54. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015). pmcid:

55. del Rosario, R. C. *et al.* Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nature methods* **12**, 458-464 (2015). pmcid:

56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009). pmcid: 2705234.

57. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010). pmcid: 2928508.

58. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J Neurosci* **34**, 1420-1431 (2014). pmcid: 3898298.

59. Chang, J., Gilman, S. R., Chiang, A. H., Sanders, S. J. & Vitkup, D. Genotype to phenotype relationships in autism spectrum disorders. *Nature neuroscience* **18**, 191-198 (2015). pmcid: 4397214.

60. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* **5**, 5748 (2014). pmcid: 4270294.

61. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261 (2011). pmcid: 3232425.

62. Sugathan, A. *et al.* CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc Natl Acad Sci U S A* **111**, E4468-4477 (2014). pmcid: 4210312.

63. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521-1531 (2013). pmcid: 3989111.

64. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* **9**, e1003709 (2013). pmcid: 3749936.

## TIMELINE AND MILESTONES

### Timeline

**Aim 1** can start immediately (Y0) as the required control and ASD brain tissue samples are already available. Data generation will continue throughout the duration of the grant. **Aim 2** can also start at Y0 based on preliminary data generated by the Geschwind and Sestan labs (see Preliminary data). Given the richness of the dataset, and the timeline of data generation, Aim 2 will continue throughout all years of the grant. We expect continued analyses to provide continued insights into integration of these networks and differences between the neurotypical and ASD brains. **Aim 3** can also start at Y0 since the high confidence ASD genes have already been identified and can be used to increase the resolution of our spatiotemporal analysis using existing methods and gene expression data. The whole-genome sequencing data will be available at Y0 and we will leverage genomic data from psychENCODE phase 1 data and other relevant datasets such as BrainSpan, CommonMind, ENCODE, GTEx and REMC to inform our analysis of regulatory regions in which de novo mutations are enriched and the spatiotemporal dynamics of the regulatory loci identified. As with Aim 2, this aim will continue throughout the course of the grant allowing continued development of improved methodology for integrating these data and further insights into ASD etiology. **All aims:** The concurrent and parallel progress in all three aims will allow methods to be developed and refined throughout the course of the grant and for the results of each aim to inform the other aims. With analysis methods in place the analytic pipelines will enable rapid progress from the completed genomic data sets to finalized analyses.

### Milestones

**Every 6 months:** Following the psychENCODE consortium policy on milestones and data releases (see *www.psychENCODE.org* or The PsychENCODE Consortium et al., 2015), cell type specific RNA-seq, ChIP-seq and Hi-C will be released to the broader research community at six-month intervals beginning from the start of the grant (Y0). See data sharing plan for details. **End of Y1:** A preliminary integrated analysis of gene expression and chromatin state data in neurotypical samples from psychENCODE phase 1 data and newly generated data from psychENCODE phase 2 will be completed to inform the discovery of de novo non-coding mutations in Aim 3. All de novo mutations in the whole-genome sequencing data will be identified. **End of Y2:** The availability of 50% of the phase 2 data from neurotypical and ASD brains will allow a preliminary analysis of gene expression, chromatin state, and 3D chromatin structure that are enriched in ASD brains for assessment of de novo enrichment in coding and non-coding loci in Aim 3. **End of Y3:** 75% of phase 2 data completed. A preliminary spatiotemporal analysis of ASD based on non-coding ASD loci will be completed in Aim 3 using preliminary integrated networks from Aim 2. **End of Y4:** All phase 2 data generated under Aim 1. Final integrated analysis of networks in the neurotypical brains and specific to the ASD brain completed under Aim 2. Final spatiotemporal analysis based on coding and non-coding loci integrated with the full psychENCODE phase 2 dataset under Aim 3.

# ABSTRACT

Genetic and genomic investigations have yielded important findings as to the genetic contributions to major psychiatric illnesses, illustrating significant etiological heterogeneity, as well as cross-disorder overlap. It has also become clear that understanding how this genetic variation leads to alterations in brain development and function that underlies psychiatric disease pathophysiology will be greatly advanced by a roadmap of the transcriptomic and epigenetic landscape of the human cerebral cortex across key developmental windows. Here, we propose, via a highly collaborative group of investigators, each with distinct areas of expertise and research focus, to create a scaffold of genomic data for understanding ASD pathophysiology, and psychiatric disorders more broadly. The work proposed here represents an ambitious multi-PI project (Yale, UCLA, and UCSF) that brings together three principal investigators and collaborators with strong publication records and expertise in all approaches necessary to perform this work using state-of-the-art and novel methodologies. We will perform time-, region-, and cell type-specific molecular profiling of control and ASD brains (Aim 1), including RNA-seq based transcriptomics, identifying *cis*-regulatory elements via ChIP-seq, and use Hi-C to determine the 3D chromatin architecture and physical relationships that underlie transcriptional regulation in three major regions implicated in neuropsychiatric disease (frontal and temporal cortex and striatum) across five major epochs representing disease-relevant stages in human brain development. This will include complementary genomic analyses in controls and matched post mortem ASD brain to identify genetic mechanisms underlying processes altered in ASD brain. We will address cellular heterogeneity via fluorescence-activated nuclear sorting (FANS) so as to profile neurons and non-neural cells separately, which will complement the whole tissue analyses. We will analyze and integrate these datasets to identify regional, developmental, and ASD-related processes to gain insight into underlying mechanisms, harmonizing these multi-omic data with other psychENCODE studies, as well as other large scale data sets, such as BrainSpan, ENCODE, GTEx and Roadmap Epigenomics Project (Aim 2). We will perform integrated analysis of germ-line ASD variations identified in more than 1000 families from the Simons Simplex Collection to characterize causal enrichments in developmental periods, brain regions, and cell types to better characterize the mechanisms by which genetic variation in humans alters brain development and function in health and disease (Aim 3). Completion of these aims will lead to a well-integrated resource across major periods in human cortical and striatal development that will permit generation of concrete testable hypotheses of ASD mechanisms, and inform our pathophysiological understanding of other related neuropsychiatric disorders.

## APPENDIX

## Content:

- Preprint of manuscript by Parikshak, N. N. et al. Global changes in patterning, splicing and primate specific lncRNAs in autism brain.
- Preprint of manuscript by Won, H. et al. Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development.

# Global changes in patterning, splicing and primate specific lncRNAs in autism brain

Neelroop N. Parikshak[1,2,*], Vivek Swarup[1,*], T. Grant Belgard[1,2, †,*], Michael Gandal[1,2], Manuel Irimia[5,6], Virpi Leppa[1], Jennifer K. Lowe[1], Robert Johnson[7], Benjamin J. Blencowe[6], Steve Horvath[3-4], Daniel H. Geschwind[1-3]

1. Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
2. Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.
3. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
4. Department of Biostatistics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
5. EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, Barcelona 08003, Spain.
6. Donnelly Centre, University of Toronto, 160 College Street, Toronto, ON M5S 3E1, Canada; Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada.
7. NICHD Brain and Tissue Bank for Developmental Disorders, University of Maryland Medical School, Baltimore, Maryland 21201, USA.

*These authors contributed equally to this study.

[†]Current address: MRC Functional Genomics Unit, Department of Physiology, Anatomy & Genetics, University of Oxford, South Parks Road, Oxford, OX1 3PT, United Kingdom.

## Summary

We apply transcriptome-wide RNA sequencing in postmortem autism spectrum disorder (ASD) brain and controls and identify convergent alterations in the noncoding transcriptome, including primate specific lncRNA, and transcript splicing in ASD cerebral cortex, but not cerebellum. We characterize an attenuation of patterning between frontal and temporal cortex in ASD and identify *SOX5*, a transcription factor involved in cortical neuron fate specification, as a likely driver of this pattern. We further show that a genetically defined subtype of ASD, Duplication 15q Syndrome, shares the core transcriptomic signature of idiopathic ASD, indicating that observed molecular convergence in autism brain is the likely consequence of manifold genetic alterations. Using co-expression network analysis, we show that diverse forms of genetic risk for ASD affect convergent, independently replicated, biological pathways and provide an unprecedented resource for understanding the molecular alterations associated with ASD in humans.

Autism spectrum disorder (ASD) is a neurodevelopmental syndrome characterized by deficits in social communication and mental flexibility[1]. Genetic risk factors contribute substantially to ASD risk, and recent studies support the potential contribution of more than a thousand genes to ASD risk[2-4]. However, given the shared cognitive and behavioral features across the autism spectrum, one hypothesis is that diverse risk factors may converge on common molecular, cellular, and circuit level pathways to result in the shared phenotype[5,6]. Analysis of the transcriptome has been used to identify common molecular pathways in the cerebral cortex (CTX) from postmortem human brain tissue in individuals with ASD[7-11]. However, all transcriptomic studies in ASD to date have been limited to evaluating highly expressed mRNAs corresponding to protein coding genes. Moreover, most lack rigorous replication and do not assess gene expression patterns across brain regions.

We used rRNA-depleted RNA-seq (Methods) to evaluate transcriptomes from a large set of ASD and control (CTL) brain samples including neocortex (frontal and temporal) and cerebellum across 79 individuals (46 ASD, 33 CTL, 205 samples, Extended Data Fig. 1a-e, Supplementary Table 1). We first compared differential gene expression (DGE) between ASD and CTL individuals in CTX from a previously published[7] microarray study against new, independent gene expression profiles from RNA-seq to evaluate global reproducibility of DGE in ASD. We found a high degree of replication of DGE fold changes between the sample sets, despite evaluation on different gene expression platforms (fold changes at $P < 0.05$ in previously evaluated data correlate with new data with $R^2 = 0.60$, Extended Data Fig. 1f). We observed a much weaker overall signal and replication in cerebellum ($R^2 = 0.033$, Extended Data Fig. 1g). These analyses confirm the existence of a reproducible DGE signature in ASD CTX across different platforms and in independent samples.

We next combined samples from all individuals with idiopathic ASD into a covariate-matched "ASD Discovery Set" (Extended Data Fig. 1h) for CTX (106 samples, 26 ASD, 33 CTL individuals) and held out remaining samples for replication ("ASD Replication Set", Methods). For DGE analysis, we used a linear mixed effects model that accounts for biological and technical covariates (Methods) to identify 1156 genes differentially expressed in ASD CTX, 582 increased and 574 decreased (Benjamini-Hochberg FDR ≤ 0.05, Supplementary Table 2). Importantly, DGE analysis with additional covariates or different assumptions about the distribution of the data and test statistics yielded similar results (Extended Data Fig. 2a). Additionally this DGE signature clusters over two-thirds of ASD samples together and this clustering is not related to confounding factors such as cortical region, age, sex, and RNA quality (Figure 1a, Extended Data Fig. 2b). The most significantly down-regulated gene was *PVALB* (fold change = 0.53, FDR ≤ 0.05), a marker for GABAergic interneurons. *SST*, a marker for a different subpopulation of GABAergic interneurons, is also among the most downregulated (fold change = 0.61, FDR ≤ 0.05). Other down-regulated genes at FDR ≤ 0.05 include *NEUROD6*, involved in neuronal differentiation (fold change = 0.60), multiple ion channels, and *KDM5D*, a lysine demethylase (fold change = 0.66). In contrast, members of the complement cascade implicated in microglial-neuronal interactions (*C4A*, fold change = 1.94; *C1QB*, fold change = 1.65; both FDR ≤ 0.05) are upregulated in ASD CTX. Gene Ontology (GO) term enrichment analysis further supports the involvement of pathways implicated by these genes (Figure 1b), confirming previous findings[7]. Moreover, the upregulated set is enriched for astrocyte and microglia enriched genes, and the down-regulated set is enriched for synaptic genes (Extended Data Fig. 2c), consistent with previous observations[7,11].

We next sought to evaluate whether the transcriptional signature identified in the ASD Discovery Set generalizes to the ASD Replication set by assessing the 1st principal component of the DGE set, which summarizes the DGE expression pattern across all cortical samples. The ASD Discovery Set and ASD

Replication Set share this pattern, which is significantly different for both sets compared to CTL (Figure 1c). Moreover, this pattern is highly associated with ASD diagnosis, but not other biological factors, technical factors, or scores on sub-domains of an ASD diagnostic tool (Figure 1d). These analyses demonstrate that convergent differences in ASD CTX are reproducible in independent samples and are not related to confounding factors.

We also detected 2715 lncRNAs expressed in cerebral cortex (after careful filtering for high-confidence transcripts, Supplementary Information), of which 62 were significantly dysregulated between ASD and CTL (33 long intergenic RNAs, lincRNAs; 19 antisense transcripts; and 10 processed transcripts at FDR ≤ 0.05). Similar to the protein coding genes, these transcripts' expression patterns cluster ASD and CTL samples (Figure 1e). Most of these lncRNAs are developmentally regulated[12], have chromatin states indicative of transcription start sites (TSSs) near their 5′ end in brain[13], and are identified in other datasets[12,14] consistent with being valid, functional lncRNAs. Moreover, most (81%) exhibit primate-specific expression patterns in brain[15] (Supplementary Information). For example, Figure 1f depicts two lincRNAs, *LINC00693* and *LINC00689*, which are typically downregulated during development, yet are upregulated in ASD CTX relative to controls (Figure 1g), which we validated by RT-PCR (Extended Data Fig. 2d). *LINC00693* sequence is present, but poorly conserved in mouse, while *LINC00689* is primate-specific (present in macaque and other primates but not in any other species, Supplementary Information, Extended Data Fig. 3 for additional examples). These data indicate that dysregulation of lncRNAs, many of which are primate-specific and involved in brain development, is an important component of transcriptome dysregulation observed in ASD.

Previous work suggested that alterations in transcript splicing may contribute to transcriptomic changes in ASD[7,16,17] by evaluating splicing in a targeted manner and pooling samples across individuals[7,16,17]. Given the increased sequencing depth and reduced sequencing bias across transcript length in our dataset, we were able to perform an unbiased genome-wide analysis of differential alternative splicing (AS). We evaluated the percent spliced in (PSI, Extended Data Fig. 4a) for 34,025 AS events in CTX across the ASD Discovery Set, encompassing skipped exons (SE), alternative 5′ splice sites (A5SS), alternative 3′ splice sites (A3SS), and mutually exclusive exons (MXE) using the MATS pipeline[18] (Supplementary Information). We first asked whether there was a global signal, finding significant enrichment over background (Extended Data Fig. 4b). We identified 1127 events in 833 genes at FDR ≤ 0.5 in CTX (similar to the number of events at uncorrected P < 0.005). Importantly, we obtained similar results with a different splice junction mapping and quantification approach (Extended Data Fig. 4c).

We performed PCR validations with nine AS events from the differential splicing set (*ASTN2, MEF2D, ERC2, MED31, SMARCC2, SYNE1, NRCAM, GRIN1, NCAM*) and found that validated changes in splicing patterns were concordant with RNA-seq (Extended Data Fig. 4d-e), demonstrating that our approach identifies alterations in AS with high specificity. Similar to our observations with lncRNA and DGE, AS changes clustered the samples by diagnosis (Figure 2a). The most significantly different event was the inclusion of an exon in *ASTN2* (ΔPSI = 5.8 indicating a mean of 5.8% difference in inclusion in ASD vs CTL; $P = 7.8\times10^{-6}$), a gene implicated by copy number variation (CNV) in ASD and other developmental disorders[19]. GO term analysis of the genes implicated by these pathways indicates involvement of biological processes related to neuronal projection, biological adhesion, and morphogenesis (Figure 2b), pathways where alternative isoforms are critical to specifying interactions between protein products. Moreover, the 1st principal component of the cortex differential splicing signature replicates in the ASD Replication Set and is not associated with other biological or technical factors (Figures 2c-d, Extended Data Fig. 5a). Importantly, many splicing alterations occur in genes that are not differentially expressed

between ASD and CTL; removing AS events on genes exhibiting even nominal DGE ($P < 0.05$), still identified a strong difference between ASD and CTL CTX  (Extended Data Fig. 5b).

A parallel analysis in cerebellum evaluating 32,954 AS events found no differentially regulated events significant at any multiple comparison correction thresholds (Extended Data Fig. 5c, Supplementary Table 3). There was no detectable global overlap between cerebellum and CTX above chance for events significant at $P < 0.05$ in both comparisons (fold enrichment = 1.1, $P = 0.21$). This suggests that AS alterations in ASD are largely confined to CTX cell types, consistent with the stronger overall DGE patterns observed in CTX versus cerebellum.

To further explore the underlying biology of AS dysregulation, we tested whether the shared splicing signature in ASD might be a product of perturbations in AS factors known to be important to neural development or preferentially expressed in neural tissue. We found that the expression levels of *RBFOX1*, *RBFOX2*, *SRRM4*, *NOVA1*, and *PTBP1* all had high correlations ($R^2 > 0.35$, $FDR \leq 0.05$) to AS alterations in CTX (Figure 2e), but not in cerebellum (Figure 2f). Furthermore, enrichment analysis revealed that most changes in cortical AS occur in neuron-specific exons that are excluded in ASD (exons with $\Delta PSI > 50\%$ in neurons overlap with exons excluded in ASD CTX, fold enrichment = 4.1, $P = 1.8x10^{-7}$, Extended Data Fig. 5d).

To validate a regulatory relationship between splicing factors and these events, we evaluated experimental data from knockout, overexpression, and knockdown experiments for Rbfox1[20], SRRM4[21], and PTBP1[22], respectively . We found that exons regulated by each of these splicing factors were significantly enriched in the set of exons excluded in ASD (Figure 2g), while in contrast, there was no enrichment for targets of ESRP[23], a splicing factor involved in epithelial cell differentiation but not expressed in CTX. This shows that alterations in three splicing factors dysregulated in ASD regulate AS of the neuron-specific exons whose inclusion is dysregulated in ASD in CTX and not cerebellum, indicating selective alteration of neuronal splicing in ASD CTX. Remarkably, the expression patterns of these three splicing factors (and others for which appropriate validation experiments were unavailable) results in distinct clusters (Extended Data Fig. 5e), suggesting that subsets of splicing factors act in different individuals to mediate a common downstream AS alteration.

Taken together these results indicate global transcriptional alterations in ASD cerebral cortex, but not cerebellum at the level of protein coding transcripts, lncRNA and AS. Therefore, to determine how these different transcriptomic subcategories relate to each other in ASD, we compared the 1st PC for each type of transcriptomic alteration across individuals (Figure 2h).  Remarkably, the PCs are highly correlated ($R^2 > 0.8$) indicating that the transcriptomic alteration is a unitary phenomenon across protein coding, noncoding, and splicing levels, rather than distinct forms of molecular alteration.

Previous analysis with gene expression microarrays in a small cohort suggested that the typical pattern of transcriptional differences between the frontal and temporal cortex may be attenuated in ASD[7]. To further test this possibility, we evaluated DGE between CTX regions (Supplementary Information) in 16 matched frontal and temporal CTX sample pairs from ASD and CTL subjects and found 551 genes differentially expressed between regions in controls, but only 51 in ASD ($FDR \leq 0.05$; Figure 3a). We refer to the set of 523 genes with this pattern in CTL, but not ASD as the “Attenuated Cortical Patterning” set. The attenuation of patterning is evident from the global distribution of test statistics between frontal and temporal CTX in ASD and CTL and genes in this set do not show a greater difference in variability in ASD versus controls compared to other genes (Kolmogorov-Smirnov test, two-tailed $P = 0.11$, Extended Data Fig. 6a).

We complemented this analysis with a machine learning approach using all 123 cortical samples, training a regularized regression model[24] to classify frontal versus temporal CTX with independent gene expression data from BrainSpan[25] (Extended Data Fig. 6b, Supplementary Information). Multiple approaches to training the classifier with BrainSpan can differentiate between frontal and temporal CTX in both CTL and ASD (Extended Data Fig. 6c-e), demonstrating that dissection and sample quality in our samples are of high quality. Loss of classification accuracy in ASD compared to CTL was observed when restricting the model to the genes with the most attenuated patterning in ASD (Extended Data Fig. 6f), demonstrating that attenuation of patterning generalizes across all samples. The Attenuated Cortical Patterning set includes multiple genes known to be involved in cell-cell communication and cortical patterning, such as *PCDH10*, *PCDH17*, *CDH12, MET*, and *PDGFD*, which was recently shown to mediate human specific aspects of cerebral cortical development[26]. GO term enrichment analysis of the Attenuated Cortical Patterning set identified enrichment for G protein coupled signaling, Wnt receptor signaling, and calcium binding, among several developmental processes (Figure 3b), and cell type enrichment analysis did not identify a strong preference for a particular cell type (Extended Data Fig. 6g).

To identify potential drivers of the alteration in cortical patterning, we evaluated transcription factor binding site enrichment upstream of genes in the Attenuated Cortical Patterning set (Supplementary Information), and found an enrichment of *SOX5* binding motifs (upstream of 364/523 genes, Figure 3c). Remarkably, *SOX5* itself belongs to the Attenuated Cortical Patterning set: while *SOX5* is differentially expressed between frontal and temporal CTX in CTL, it is not in ASD (Figure 3d). We thus predicted that if *SOX5* regulates cortically patterned genes, its expression should correlate with target gene transcript levels. Consistent with this prediction, we found that genes in the Attenuated Cortical Patterning set are anti-correlated with *SOX5* in CTL CTX, but not in ASD CTX (Figure 3e, top left; Wilcoxon rank sum test of R values, P = 0.01), suggesting that the normal role of SOX5 as a transcriptional repressor may be disrupted in ASD. We reasoned that a true loss of SOX5-mediated cortical patterning would be specific to the predicted SOX5 targets. Consistent with this, we find a loss of correlations between *SOX5* and predicted targets, but no difference in correlations between *SOX5* and non-targets in the Attenuated Cortical Patterning set (Figure 3e). Taken together, these findings show that a loss of regional patterning downstream of the transcriptional repressor *SOX5*, which plays a crucial role in glutamatergic neuron development[27,28], contributes to the loss of regional identity in ASD.

Gene expression changes in postmortem brain may be a consequence of genetic factors, environmental factors, or both. Brain tissue from individuals with ASD that harbor known, penetrant genetic causes are very rare. However, we were able to identify postmortem brain tissue from 8 subjects with one of the more common recurrent forms of ASD, Duplication 15q Syndrome (dup15q, which is present in about 0.5-1% of ASD cases, see Extended Data Fig. 7a for characterization of duplications). We performed RNA-seq across frontal and temporal cortex and compared DGE changes in dup15q with those observed in individuals with idiopathic ASD to better understand the extent to which the observed molecular pathology overlaps. As expected, most genes in the 15q11.1-13.2 duplicated region have higher expression in dup15q CTX compared to CTL (Figure 4a), although *SNRPN* and *SNURF* were notably downregulated. Conversely, no significant upregulation of genes in this region were identified in idiopathic ASD or controls. Strikingly, when we assessed genome-wide expression changes, we observed a strong signal of DGE in dup15q that widely overlaps with that of idiopathic ASD (fold changes at FDR $\leq 0.05$ in dup15q correlate with idiopathic ASD with $R^2 = 0.79$, Figure 4b). Moreover, the slope of the best-fit line through these changes is 2.0, indicating that on average, the transcriptional changes in dup15q CTX are highly similar, but twice the magnitude of those observed in ASD CTX.

Next, we sought to evaluate AS changes in dup15q. There is only one significant splicing change in the dup15q region (Supplementary Table 3), consistent with the idea that duplication in this region duplicates all isoforms of the genes, resulting in minimal alteration of transcript structure. Similar to DGE, global AS analysis in dup15q CTX vs to CTL CTX revealed a stronger, but highly overlapping signature with idiopathic ASD CTX (fold changes at FDR ≤ 0.2 in dup15q agree correlate with idiopathic ASD with $R^2 = 0.66$) indicating that splicing changes in dup15q syndrome recapitulate those of idiopathic ASD (Figure 4c). The slope of the best-fit line through the PSI for spliced exons in dup15q CTX compared to those in ASD CTX is 2.5 similar to DGE. Notably, both gene expression and AS changes in dup15q implicated similar pathways as those found in idiopathic ASD (Extended Data Fig. 7c-d). Clustering dup15q samples and CTL samples using both the DGE set and the differential AS set showed that all dup15q samples cluster together (Figure 4d), as opposed to the more variable clustering of idiopathic ASD, supporting the hypothesis that this shared genetic abnormality leads to a more homogeneous molecular phenotype.

Next, to test whether this molecular ASD signature may be due to independent of postmortem or reactive effects (Supplementary Information), we compared our data with gene expression profiles from a iPSC-derived neurons (nIPSCs)[29] from dup15q were available, we could use these data to definitively reveal which changes in dup15q CTX are independent of postmortem or reactive effects (Supplementary Information), since such effects are not present *in vitro*. We observe that DGE in the 15q region is concordant with that seen in the nIPSCs (Figure 4e), even though the sample size is small and the analysis is likely underpowered. Upregulated changes in dup15q are also seen in nIPSCs (Figure 4f), consistent with our other statistical analyses showing limited effects of potential confounders. The very immature, fetal state of the nIPSCs[30] likely explains the absence of an enrichment signal for genes downregulated in postnatal ASD brain, which are enriched for genes involved in neurons with more mature synapses.

We next applied gene network analysis to construct an organizing framework to understand shared biological functions across idiopathic ASD and dup15q (combining the ASD Discovery Set, ASD Replication Set, and dup15q set). We utilized Weighted Gene Co-expression Network Analysis (WGCNA), which identifies groups of genes with shared expression patterns across samples (modules) from which shared biological function is inferred. Modules identified via WGCNA can than be related to a range of relevant phenotypes and potential confounders[31,32]. We applied signed co-expression analysis and used bootstrapping to ensure the network was robust, and not dependent on any subset of samples (Supplementary Information), while controlling for technical factors and RNA quality ("Adjusted FPKM" levels, Methods). WGCNA identified 16 co-expression modules (Extended Data Fig. 8a, Supplementary Table 2), which are further characterized by their association to ASD (Extended Data Fig. 8b), enrichment for cell-type specific genes (Extended Data Fig. 8c), and enrichment for GO terms (Extended Data Fig. 9). Of the downregulated modules, three are associated with ASD and dup15q (M1/10/17) and one with dup15q only (M11). Five of the upregulated modules are associated with ASD and dup15q (M4/5/6/9/12) and one is specific to dup15q (M13) (Figure 5a, top). Additionally, we identified a module strongly enriched for genes from the Attenuated Cortical Patterning set and *Wnt* signaling that contains *SOX5* (M12; fold enrichment = 3.0, P = $3x10^{-8}$), verifying the strong relationship observed between the *Wnt* pathway regulating TF *SOX5* and attenuation of cortical patterning[33].

Notably, the modules identified here significantly overlap with previous patterns identified in ASD (asdM12$_{array}$ and asdM16$_{array}$[7]; Figure 5a, middle). We found that the ASD-associated modules identified by our larger sample size and RNA-seq provide significant refinement of previous observations by identifying more discrete biological processes related to cortical development[34], the post-synaptic density[35], and

lncRNAs (Figure 5a, bottom). For example, M1 overlaps a subset of asdM12$_{array}$ (fold enrichment = 5.7) and developmental modules (devM16 fold enrichment = 3.7), and is enriched for proteins found in the PSD and genes involved in calcium signaling and gated ion channel signaling. Another subset of asdM12$_{array}$, M10 (fold enrichment = 11) overlaps more with a mid-fetal upregulated cortical development module (devM13 fold enrichment = 4.0), and genes involved in secretory pathways and intracellular signaling. A third module, M17 shows the least overlap with asdM12$_{array}$ (fold enrichment = 2.2) and is related to energy metabolism. Notably, these three modules are enriched for neuron-specific genes (Extended Data Fig. 8c), but not all neuronal modules are down regulated in ASD (M3 is not altered in ASD CTX). Taken together, specific neurobiological processes are affected in individuals with ASD related to developmentally regulated neurodevelopmental processes.

The most upregulated modules, M5 and M9, both strongly overlap (fold enrichments > 20) with previously identified upregulated co-expression module asdM16$_{array}$. M5 is enriched for microglial cell markers and immune response pathways, whereas M9 is enriched for astrocyte markers and immune-mediated signaling and immune cell activation (Extended Data Fig. 8c, Extended Data Fig. 9). This analysis clearly separates the contributions of the coordinated biological processes of microglial activation and reactive astrocytosis, which were previously not distinguishable as separate modules[7]. Thus, our analysis pinpoints more specific biological pathways in idiopathic ASD than those previously identified and reveals that similar changes occur downstream of the genetic perturbation in dup15q.

We evaluated the relationship between the five modules most strongly associated with ASD (M1/5/9/10/17, which are supported by module-trait association analysis and gene set enrichment analysis, Supplementary Information), and found that there was a remarkably high anti-correlation between the eigengene of M5 and downregulated modules, particularly M1 ($R^2 = 0.76$) (Figure 5b). M1 (Figure 5c) is downregulated in ASD and enriched for genes at the PSD and genes involved in synaptic transmission, while M5 (Figure 5d) is enriched for microglial genes and cytokine activation. This strong anti-correlation between microglial signaling and synaptic signaling in ASD and dup15q provides evidence in humans for dysregulation of microglia-mediated synaptic pruning, as previously suggested[36].

Next, to determine the role of causal genetic variation, we evaluated enrichment of both rare genetic variants, focusing on genes affected by ASD associated gene disrupting (LGD) *de novo* mutations[37], and common variants[38,39]. Genes within three modules, M1, M3, and M12, show enrichment for common variation signal for ASD (Figure 5e, Methods). Remarkably, M12 (Figure 5f), which is related to cortical patterning and Wnt signaling, also exhibit GWAS signal enrichment, providing the first evidence that risk conferred by common variation in ASD may affect regionalization of the cortex. Interestingly, M3 is significantly enriched for both schizophrenia (SCZ) and ASD common variants, is related to synaptic transmission, nervous system development, and regulation of ion channel activity (Extended Data Fig. 9), consistent with the notion that ASD and SCZ share common and rare genetic risk[1,40-43].

We only identified one module, M2 (Figure 5g), as significantly enriched in protein disrupting (nonsense, splice site, or frameshift) rare *de novo* variants previously associated with SCZ and ASD. M2 overlaps with a cortical developmental module implicated in ASD[34] (devM2 fold enrichment = 5.1). Notably, M2 is not differential between ASD and CTL in our dataset, consistent with the observation that these genes are primarily expressed during early neuronal development in fetal brain[34]. Remarkably, M2 contains an unusually large fraction of lncRNAs (15% of the genes in M2 are classified as lncRNAs, while other modules are 1-5% lncRNA). We hypothesize that, in addition to protein coding genes involved in transcriptional and chromatin regulation, rare *de novo* variants may also affect lncRNAs in ASD, a prediction that will be testable once large sets of whole genome sequences are available.

These combined transcriptomic and genetic analyses reveal that different forms of genetic variation affect biological processes involved in multiple stages of cortical development. Common genetic risk is enriched in M3, M1, and M12, which reflect early glutamatergic neurogenesis, later neuronal function, and cortical patterning, respectively. We also observe that rare *de novo* variation, which is enriched in M2, affects distinct biology related to transcriptional regulation and chromatin modification. These findings are consistent with transcriptomic analyses of early prenatal brain development and ASD risk mutations that implicate chromatin regulation and glutamatergic neuron development[34,44].

We provide the first comprehensive picture of largely unexplored aspects of transcription in ASD, lncRNA and alternative splicing, and identify a strong convergent signal in these, as well as protein coding genes[7]. These results will aid in interpreting genetic variation outside of the known exome, as whole genome sequencing supplants current methods. A role of lncRNAs has been previously explored in ASD[45], but only two individuals were evaluated with targeted microarrays. We evaluate lncRNAs in an unbiased manner across many individuals, notably identifying an enrichment of lncRNAs in M2, most of which are uncharacterized in brain and arose on the primate lineage. The involvement of lncRNAs in this early developmental program that is enriched for *de novo* mutations implicated in ASD suggests their study will be particularly relevant to understanding the emergence of primate higher cognition on the mammalian lineage, and by extension human brain evolution[15,46,47].

We also provide the first confirmation of an attenuation of genes that typically show differential expression between frontal and temporal lobe in ASD CTX and further identified *SOX5*, known to regulate cortical laminar development[50,51], as a putative regulator of this disruption. That M12, which is enriched for genes exhibiting cortical regionalization and is also enriched in ASD GWAS signal, supports the prediction that attenuation of patterning may be mediated by common genetic variation in or near the *SOX5* target genes. Disruption of cortical lamination by direct effects on glutamatergic neurogenesis and function has been predicted by independent data, including network analyses of rare ASD associated variants identified in exome sequencing studies[34,44].

These data, in conjunction with previous studies, reveal a consistent picture of the ASD's emerging postnatal and adult pathology. Specific neuronal signaling and synaptic molecules are downregulated and astrocyte and microglial genes are upregulated in over 2/3 of cases. Microglial infiltration has been observed in ASD cortex with independent methods[52], and normal microglial pruning has been shown to be necessary for brain development[36]. Our findings further suggest that aberrant microglial-neuronal interactions may be pervasive in ASD and related to the gene expression signature seen in a majority of individuals. In our comprehensive AS analysis, we identify three splicing factors upstream of the altered splicing signature observed in ASD CTX. These factors are known to be involved in coordinating sequential processes in neuronal development[17,21] and maintaining neuronal function[48,49]. It may therefore be sufficient to disrupt any one of these factors to induce a similar outcome during brain development, which would be consistent with the shared downstream perturbation observed here.

Finally, evaluation of the transcriptome in dup15q supports the enormous value of the "genotype first" approach of studying syndromic forms of ASD, with known penetrant genetic lesions[53]. It is highly unlikely that the shared transcriptional dysregulation in dup15q is due to a shared environmental insult. Thus, the most parsimonious explanation for the convergent transcriptomic pathology seen in all dup15q and over 2/3 of the cases of idiopathic ASD is that it represents an adaptive or maladaptive response to a primary genetic insult, which in most cases of ASD will be genetic[2,54]. As future investigations pursue the full range of causal genetic variation contributing to ASD risk, these analyses and data will be valuable for interpreting genetic and epigenetic studies of ASD as well as those of other neuropsychiatric disorders.

## Figure Legends

Figure 1 | Transcriptome-wide differential gene expression in ASD.  a, Average linkage hierarchical clustering of samples in the ASD Discovery Set using the top 100 upregulated and top 100 downregulated protein coding genes. b, Gene Ontology (GO) term enrichment analysis of upregulated and downregulated genes in ASD. *FDR ≤ 0.05 across all GO terms and gene sets. c, 1st principal component of the CTX DGE set (CTX DGE PC1) is able to distinguish ASD and CTL samples, including independent samples from the ASD Replication Set. d, CTX DGE PC1 is primarily associated with diagnosis, and not other factors. e, Average linkage hierarchical clustering of ASD Discovery Set using all lncRNAs in the DGE set. f, UCSC genome browser track displaying reads per million (RPM) in a representative ASD and CTL sample, superimposed over the gene models and sequence conservation for genomic regions including *LINC00693* and *LINC00689*. g, *LINC00693* and *LINC00689* are upregulated across ASD samples and downregulated during frontal cortex development. Abbreviations: FC, frontal cortex; TC, temporal cortex; RIN, RNA integrity number; ADI-R score, Autism Diagnostic Interview Revised score; FPKM, fragments per kilobase million mapped reads.

Figure 2 | Alteration of alternative splicing in ASD. a, Average linkage hierarchical clustering of ASD discovery set using top 100 differentially included and top 100 differentially excluded exons from the differential splicing (DS) set across the ASD Discovery Set. b, Gene Ontology term enrichment analysis of genes with DS in ASD. c, 1st principal component 1 of the CTX differential alternative splicing set (CTX DS PC1) is able to distinguish ASD and CTL samples using independent samples from the ASD Replication Set. d, CTX DS PC1 is primarily associated with diagnosis, and not other factors. e, Correlation between CTX DS PC1 and gene expression of neuronal splicing factors in CTX. f, Correlation between 1st principal component of cerebellum differential splicing (CB DS PC1) and gene expression of neuronal splicing factors in cerebellum. g, Overlap between DS set and splicing events regulated by splicing factors where experimental data was available. h, Scatterplots and correlations between the 1st principal component across the ASD versus CTL DGE sets for different transcriptome subcategories. Abbreviations: FC, frontal cortex; TC, temporal cortex; RIN, RNA integrity number; ADI-R score, Autism Diagnostic Interview Revised score; FPKM, fragments per kilobase million mapped reads.

Figure 3 | Attenuation of cortical patterning in ASD cortex. a, Heatmap of 551 genes exhibiting cortical patterning between frontal cortex (FC) and temporal cortex (TC) in ASD, with samples sorted by diagnostic status and brain region. b, Gene ontology term enrichment analysis of genes exhibiting attenuated cortical patterning (ACP). c, Schematic of transcription factor motif enrichment upstream of genes in the ACP set, with the *SOX5* motif sequence logo. d, The *SOX5* gene exhibits attenuated cortical patterning in ASD CTX compared to CTLs. Lines connect FC-TC pairs that are from the same individual. e, Correlation between *SOX5* gene expression and predicted targets in CTL and ASD, with all ACP genes (top left), SOX5 targets from the ACP set (top right),  SOX5 non-targets from the ACP set (bottom left), and all genes not in the ACP set (bottom right). Plots show the difference in correlation between *SOX5* and other genes in ASD and CTL (ΔR).

Figure 4 | Duplication 15q Syndrome recapitulates transcriptomic changes in idiopathic ASD. a, DGE changes across the 15q11-13.2 region for ASD and dup15q compared to CTL, error bars are +/- 95% confidence intervals for the fold changes. b, Comparison of effect sizes in dup15q vs CTL and ASD vs

CTL, with changes in dup15q at FDR ≤ 0.05 highlighted. c, Comparison of differential splicing (DS) changes in dup15q vs CTL and ASD vs CTL, highlighting 402 events at FDR ≤ 0.2 in dup15q. d, Average linkage hierarchical clustering of dup15q samples and controls using the DGE and DS gene sets. e, Plot of fold changes between induced pluripotent stem cells differentiated into neurons (nIPSCs) from dup15q vs CTL and postmortem CTX DGE from dup15q vs CTL in the 15q region. f,  Heatmap overlapping the top 1000 genes up- and down- regulated in the nIPSC comparison to the up- and down- regulated genes in dup15q and idiopathic ASD CTX.

Figure 5 | Co-expression network analysis across all ASD and CTL samples in CTX. a, Gene set enrichment analyses comparing the 16 co-expression modules with multiple gene sets from this RNA-seq study, from postmortem ASD CTX microarray, from human brain development, from the postsynaptic density and set of all brain-expressed lncRNAs. b, Comparison of five ASD-associated modules against each other by correlating module eigengenes. c, Module plot of M1 displaying the top 25 hub genes along with the module's Gene Ontology term enrichment. d, similar to c, but for M5. e, Gene set enrichment analysis with genome-wide whole-exome sequencing data (Rare *de novo* hit genes) and genome-wide association study (GWAS) results in ASD, schizophrenia (SCZ), and intellectual disability (ID). Boxes are filled if the odds ratio is greater than 0, and the enrichment $P < 0.05$. Asterisks* indicate FDR ≤ 0.05 across all comparisons in a and e. f,g, similar to c, but for M12 and M2, respectively. Abbreviations: LGD, likely gene disrupting, genes affected by nonsense, nonsynonymous, or splice-site mutations or frame-shift indels; AGRE, AGP/CHOP, and PGC refer to consortia that collect genetic data (Supplementary Information for details).

## Methods

Sample description: Brain tissue for ASD and control individuals was acquired from the Autism Tissue Program (ATP) brain bank at the Harvard Brain and Tissue Bank and the University of Maryland Brain and Tissue Bank (a Brain and Tissue Repository of the NIH NeuroBioBank). Sample acquisition protocols were followed for each brain bank, and samples were de-identified prior to acquisition. Brain sample and individual level metadata is available in Supplementary Table 1.

RNA-seq methodology: Starting with 1ug of total RNA, samples were rRNA depleted (RiboZero Gold, Illumina) and libraries were prepared using the TruSeq v2 kit (Illumina) to construct unstranded libraries with a mean fragment size of 150bp (range 100-300bp) that underwent 50bp paired end sequencing on an Illumina HiSeq 2000 or 2500 machine. Paired-end reads were mapped to hg19 using Gencode v18 annotations[55] via Tophat2[56]. Gene expression levels were quantified using union exon models with HTSeq[57]. For additional and information on sequencing and read alignment parameters, please see Supplementary Information.

Sample sets for analysis: For differential gene expression and splicing analysis, we defined an age matched set, referred to as the ASD Discovery Set (106 samples in CTX, 51 in cerebellum) of idiopathic ASD and control samples for the discovery set, and held out younger or unmatched samples as the ASD Discovery Set (17 in CTX, 8 in cerebellum). Dup15q individuals were analysed separately, utilizing the full set of controls from the ASD Discovery Set. For co-expression network analysis, we combined the discovery set, replication set, and dup15q individuals for a total of 137 CTX samples and 59 cerebellum samples.

Differential Gene Expression (DGE): DGE analysis was performed with expression levels adjusted for gene length, library size, and G+C content (referred to as "Normalized FPKM") Supplementary Information.

CTX samples (frontal and temporal) were analyzed separately from cerebellum samples. A linear mixed effects model framework was used to assess differential expression in log2(Normalized FPKM) values for each gene for cortical regions (as multiple brain regions were available from the same individuals) and a linear model was used for cerebellum (where one brain region was available in each individual, with a handful of technical replicates removed). Individual brain ID was treated as a random effect, while age, sex, brain region (except in the case of cerebellum, where there is only one region), and diagnoses were treated as fixed effects. We also used technical covariates accounting for RNA quality, library preparation, and batch effects as fixed effects into this model (Supplementary Information).

Reproducibility analyses: We assessed replication between datasets by evaluating the concordance between independent sample sets by comparing the squared correlation ($R^2$) of fold changes of genes in each sample set at a non-stringent P value threshold. This general approach has been shown to be effective for identifying reproducible gene expression patterns[58], and we modify it such that the P value threshold is set in one sample set (the *x* axis in the scatterplots), and the $R^2$ with fold changes in these genes are evaluated in an independent sample set (the *y* axis in the scatterplots).

Differential Splicing Analysis: Alternative splicing was quantified using the percent spliced in (PSI) metric using Multivariate Analysis of Transcript Splicing (MATS, v3.08)[18]. For each event, MATS reports counts supporting the inclusion (I) or exclusion (E) of a splicing event. To reduce spurious events due to low counts, we required at least 80% of samples to have I + S >= 10. For these events, the percent spliced in is calculated as PSI = I / (I + S) (Extended Data Fig. 4a). Statistical analysis for differential splicing was performed utilizing the linear mixed effects model regression framework as described above for DGE. This approach is advantageous over existing methods as it allows modeling of covariates and takes into consideration the variability in PSI across samples when assessing event significance with ASD (Supplementary Information).

Genotyping dup15q: For Dup15q samples, the type of duplication and copy number in the breakpoint 2-3 region were available for these brains[59]. To expand this to the regions between each of the recurrent breakpoint in these samples, 7/8 dup15q brains were genotyped (one was not genotyped due to limitations in tissue availability). The number of copies between each of the breakpoints is reported in Extended Data Fig. 7a.

Co-expression network analysis: The R package weighted gene co-expression network analysis (WGCNA) was used to construct co-expression networks using the technical variation normalized data[31,60] (referred to as “Adjusted FPKM”). We used the biweight midcorrelation to assess correlations between log2(Normalized FPKM) and parameters for network analysis are described in Supplementary Information. Notably, we utilized a modified version of WGCNA that involves bootstrapping the underlying dataset 100 times and constructing 100 networks. The consensus of these networks (50th percentile across all edges) was then used as the final network [32], ensuring that a handful of samples do not determine the network structure. For module-trait analyses, 1st principal component of each module (eigengene) was related to ASD diagnosis, age, sex, and brain region in a linear mixed effects framework as above, only replacing the expression values of each gene with the eigengene.

Enrichment analysis of gene sets and GWAS: Enrichment analyses were performed either with Fisher’s exact test (cell type and splicing factor enrichments) or logistic regression (all enrichment analyses in Figure 5). We used logistic regression in the latter case to control for gene length or other biases that may influence enrichment analysis (Supplementary Information). All GO term enrichment analysis was performed using GO Elite[61] with 10,000 permutations. We focused on molecular function and biological process terms for display purposes.

## Extended Data Figure Legends

Extended Data Figure 1 | Methodology, quality control, and differential expression replication analysis. a, RNA-seq workflow, including RNA extraction, library preparation, sequencing, read alignment, and quality control. b, RNA-seq quality and alignment statistics from this study, including RNA integrity number (RIN), number of aligned reads, proportion of reads mapping to different genomic features (mRNA, intronic, intergenic), and bias in coverage from the 5' to the 3' end of the top 1000 expressed transcripts (statistics compiled using PicardTools). c, Similar statistics as in b for another RNA-seq study that utilized polyA tail selection mRNA-seq to evaluate the transcriptome in ASD cortex[11] (primarily BA19, visual cortex, but also including some BA10/44 samples, frontal cortex). d, RNA-seq read coverage relative to normalized gene length across transcripts from the 5' to the 3' end in this study. e, Dependence between coverage and RIN across gene body (correlation between RIN and coverage in d across samples). f, Correlation of ASD vs CTL fold changes between previously evaluated and new ASD samples in CTX by microarray (left) and RNA-seq (right) using genes that were at $P < 0.05$ the samples from Voineagu et al., 2011. g, Correlation between effect sizes as in f, but for cerebellum (CB) samples. h,i, Correlation between covariates and ASD vs CTL status in CTX (h) and CB (i) in the ASD Discovery Set.

Extended Data Figure 2 | Transcriptome-wide differential gene expression (DGE) analysis in CTX. a, Comparison of P value rankings across different methods for DGE with Spearman's correlation. From left to right: removal of three additional principal components of sequencing statistics (Supplementary Information) related to RNA-sequencing quality, application of a permutation analysis for DGE P value computation, application of variance-weighted linear regression for DGE[62], and using surrogate variable analysis for DGE[63]. b, Average linkage hierarchical clustering heatmap using all genes DGE in the ASD Discovery Set, but including all idiopathic ASD frontal cortex (FC) and temporal cortex (TC) samples across 123 samples, combining the ASD Discovery set and the ASD Replication set. Bolded samples in the dendrogram are used for validation in d. c, Enrichment analysis of cell-type specific gene sets (5-fold enriched in the cell type compared to all other cells) with genes decreased and increased in ASD. d, RT-PCR validation of the two lincRNAs shown in Figure 1f-g, P values are computed with the Wilcoxon rank-sum test.

Extended Data Figure 3 | Gene browser tracks for selected primate-specific lncRNAs. For each lncRNA, expression for representative samples for ASD vs CTL (top) in human, macaque (middle), and mouse (bottom) are shown. The genome location for macaque and mouse displayed is syntenic to the human region, with the expected location of the lncRNA highlighted.

Extended Data Figure 4 | Splicing analyses and validation in ASD. a, Schematic describing how the percent spliced in (PSI) metric is computed. b, Distribution of *P* values for changes in the PSI between ASD and CTL in CTX for all events (left) and event subtypes (SE, spiced exon; A5SS, alternative 5' splice site; A3SS, alternative 3' splice site; MXE, mutually exclusive exons). c, Comparison of the CTX splicing analyses in when using PSI values obtained via read alignment by TopHat2[64] followed by the MATS[18] pipeline (used throughout this study) against read alignment by OLego followed by Quantas[65]. d, Comparison of ΔPSI values in nine splicing events between PCR and RNA-seq. e, PCR validation and sashimi plots for the nine splicing events delineated in d, from the samples highlighted in Extended Data Fig. 5a.

Extended Data Figure 5 | Additional splicing analyses in ASD. a, Average linkage hierarchical clustering heatmap using all differentially spiced (DS) events from the ASD Discovery Set, but including all idiopathic ASD neocortical samples (FC and TC) across 123 samples, combining the ASD Discovery set and the ASD Replication set. Bolded samples in the dendrogram were used for PCR validation in Extended Data Fig. 4. b, Top: difference between ASD and CTL in the DS set based on PC1 of the DS set at the PSI level, and PC1 of the gene expression levels of genes in the DS set. Bottom: Same comparison after differentially expressed genes ($p < 0.05$) are removed. c, Distribution of P values for changes in the PSI between ASD and CTL in cerebellum. d, Cell-type enrichment analysis of splicing events from CTX. e, Average-linkage hierarchical clustering using 1-(Pearson's correlation) to compare the gene expression patterns of the splicing factors investigated in Figure 2.

Extended Data Figure 6 | Attenuation of cortical patterning in ASD. a, Histograms of P values from paired Wilcoxon rank-sum test differential gene expression between 16 frontal cortex (FC) and 16 temporal cortex (TC) in CTL and ASD and a histogram of Bartlett's test P values for differences in gene expression variance between ASD and CTL for all genes (white) and genes in the Attenuated Cortical Patterning (ACP) set (red). c, Approach to training the elastic net model on BrainSpan and application of the model on 123 cortical samples in this study. c-e, Results of learned cortical region classifications with different starting gene sets, with the BrainSpan training set (left), CTL samples (middle), and ASD samples (right) in each panel and the Wilcoxon rank-sum test P value of FC vs TC difference for each comparison. f, Summary of results form c-e. g, Cell type enrichment analysis for genes in the ACP set. Abbreviations: A1C, primary auditory cortex; DFC, dorsolateral prefrontal cortex; MFC, medial prefrontal cortex; STC, superior temporal cortex; FC, frontal cortex; TC, temporal cortex; AUROC, area under the receiver-operator characteristic curve.

Extended Data Figure 7 | Dup15q syndrome analyses. a, Copy number between breakpoints (BP) in the 15q region. Genome-wide CNV analysis allowed evaluation of copy number in additional regions from previous studies[59,66]. b, Differential expression across the 15q region of interest in dup15q vs CTL and ASD vs CTL cerebellum, note only 3 samples were available for dup15q cerebellum so additional analyses were not pursued. c, Gene Ontology term enrichment analysis for the dup15q CTX differential expression set. d, Gene Ontology term enrichment analysis for the dup15q CTX differential splicing (DS) set. e, Hierarchical clustering of iPSC-derived neurons from dup15q, Angelman syndrome, and a control[29].

Extended Data Figure 8 | Co-expression network analysis in ASD CTX. a, Modules identified from a dendrogram constructed from a consensus of 100 bootstrapped datasets using the 137 CTX samples. Correlations for each gene to each measured factor are delineated below the dendrogram (blue = negative, red = positive correlation). b, Module-trait associations as computed by a linear mixed effects model with all factors on the x-axis used as covariates. All P values are displayed where the coefficient passed $p < 0.01$. Note that this alternative approach to module-trait association agrees with the Fisher's exact test used in Figure 5a when the fold enrichment for module overlap with DGE sets is > 2.8, and we use an intersection of both methods for the modules focused on in Figure 5b. c, Module enrichments for cell type specific gene expression patterns.

Extended Data Figure 9 | GO term enrichments for all modules. *FDR  < 0.05 across all GO enrichments across all modules.

## References

1. Geschwind, D. H. Genetics of autism spectrum disorders. *Trends Cogn. Sci. (Regul. Ed.)* 15, 409–416 (2011).
2. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* 46, 881–885 (2014).
3. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* (2015). doi:10.1016/S1474-4422(15)00044-7
4. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* 17, 782–790 (2014).
5. Chen, J. A., Peñagarikano, O., Belgard, T. G., Swarup, V. & Geschwind, D. H. The emerging picture of autism spectrum disorder: genetics and pathology. *Annu Rev Pathol* 10, 111–144 (2015).
6. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 9, 341–355 (2008).
7. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384 (2011).
8. Purcell, A. E., Jeon, O. H., Zimmerman, A. W., Blue, M. E. & Pevsner, J. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* 57, 1618–1628 (2001).
9. Garbett, K. *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiology of Disease* 30, 303–311 (2008).
10. Chow, M. L. *et al.* Age-Dependent Brain Gene Expression and Copy Number Anomalies in Autism Suggest Distinct Pathological Processes at Young Versus Mature Ages. *PLoS Genet.* 8, e1002592 (2012).
11. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Comms* 5, 5748 (2014).
12. Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Publishing Group* 18, 154–161 (2015).
13. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).
14. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* 47, 199–208 (2015).
15. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640 (2014).
16. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Reports* 6, 1139–1152 (2014).
17. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159, 1511–1523 (2014).
18. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* 40, e61–e61 (2012).
19. Lionel, A. C. *et al.* Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. *Human Molecular Genetics* 23, 2752–2768 (2014).
20. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* 20, 1434–1442 (2013).
21. Raj, B. *et al.* A Global Regulatory Mechanism for Activating an Exon Network Required for Neurogenesis. *Molecular Cell* 56, 90–103 (2014).
22. Gueroussov, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* 349, 868–873 (2015).
23. Dittmar, K. A. *et al.* Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Molecular and Cellular Biology* 32, 1468–1482 (2012).

24. Tibshirani, R., Johnstone, I., Hastie, T. & Efron, B. Least angle regression. *The Annals of Statistics* 32, 407–499 (2004).

25. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* 41, D996–D1008 (2013).

26. Lui, J. H. *et al.* Radial glia require PDGFD–PDGFRβ signalling in human but not mouse neocortex. *Nature* 515, 264–268 (2014).

27. Lai, T. *et al.* SOX5 Controls the Sequential Generation of Distinct Corticofugal Neuron Subtypes. *Neuron* 57, 232–247 (2008).

28. Kwan, K. Y. *et al.* SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16021–16026 (2008).

29. Germain, N. D. *et al.* Gene expression analysis of human induced pluripotent stem cell-derived neurons carrying copy number variants of chromosome 15q11-q13.1. *Mol Autism* 5, 44 (2014).

30. Stein, J. L. *et al.* A Quantitative Framework to Evaluate Modeling of Cortical Development by Neural Stem Cells. *Neuron* 83, 69–86 (2014).

31. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* 4, Article17 (2005).

32. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 1, 54 (2007).

33. Morales, P. L. M., Quiroga, A. C., Barbas, J. A. & Morales, A. V. SOX5 controls cell cycle progression in neural progenitors by interfering with the WNT–β-catenin pathway. *EMBO reports* 11, 466–472 (2010).

34. Parikshak, N. N. *et al.* Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* 155, 1008–1021 (2013).

35. Bayés, À. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* 14, 19–21 (2010).

36. Schafer, D. P. *et al.* Microglia Sculpt Postnatal Neural Circuits in an Activity and Complement-Dependent Manner. *Neuron* 74, 691–705 (2012).

37. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221 (2014).

38. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Human Molecular Genetics* 21, 4781–4792 (2012).

39. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. 459, 528–533 (2009).

40. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* 45, 984–994 (2013).

41. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506, 185–190 (2014).

42. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res* 25, 142–154 (2015).

43. Gilman, S. R. *et al.* Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nat. Neurosci.* 15, 1723–1728 (2012).

44. Willsey, A. J. *et al.* Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell* 155, 997–1007 (2013).

45. Ziats, M. N. & Rennert, O. M. Aberrant Expression of Long Noncoding RNAs in Autistic Brain. *J Mol Neurosci* 49, 589–593 (2012).

46. Geschwind, D. H. & Rakic, P. Cortical Evolution: Judge the Brain by Its Cover. *Neuron* 80, 633–647 (2013).

47. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated Recruitment of New Brain Development Genes into the Human Genome. *PLoS Biol* 9, e1001179 (2011).

48. Fogel, B. L. *et al.* RBFOX1 regulates both splicing and transcriptional networks in human neuronal

development. *Human Molecular Genetics* 21, 4171–4186 (2012).

49. Gehman, L. T. *et al.* The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nat Genet* 43, 706–711 (2011).

50. Greig, L. C., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat Rev Neurosci* 14, 755–769 (2013).

51. Srinivasan, K. *et al.* A network of genetic repression and derepression specifies projection fates in the developing neocortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19071–19078 (2012).

52. Morgan, J. T. *et al.* Abnormal microglial–neuronal spatial organization in the dorsolateral prefrontal cortex in autism. *Brain Research* 1456, 72–81 (2012).

53. Stessman, H. A., Bernier, R. & Eichler, E. E. A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell* 156, 872–877 (2014).

54. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet* 15, 133–141 (2014).

55. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7, S4–9 (2006).

56. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31, 46–53 (2012).

57. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).

58. Shi, L. *et al.* The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics* 9, S10 (2008).

59. Scoles, H. A., Urraca, N., Chadwick, S. W., Reiter, L. T. & LaSalle, J. M. Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. *Mol Autism* 2, 19 (2011).

60. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559 (2008).

61. Zambon, A. C. *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* 28, 2209–2210 (2012).

62. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15, R29 (2014).

63. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* 3, e161 (2007).

64. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578 (2012).

65. Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res* 41, 5149–5163 (2013).

66. Wintle, R. F. *et al.* A genotype resource for postmortem brain samples from the Autism Tissue Program. *Autism Res* 4, 89–97 (2011).

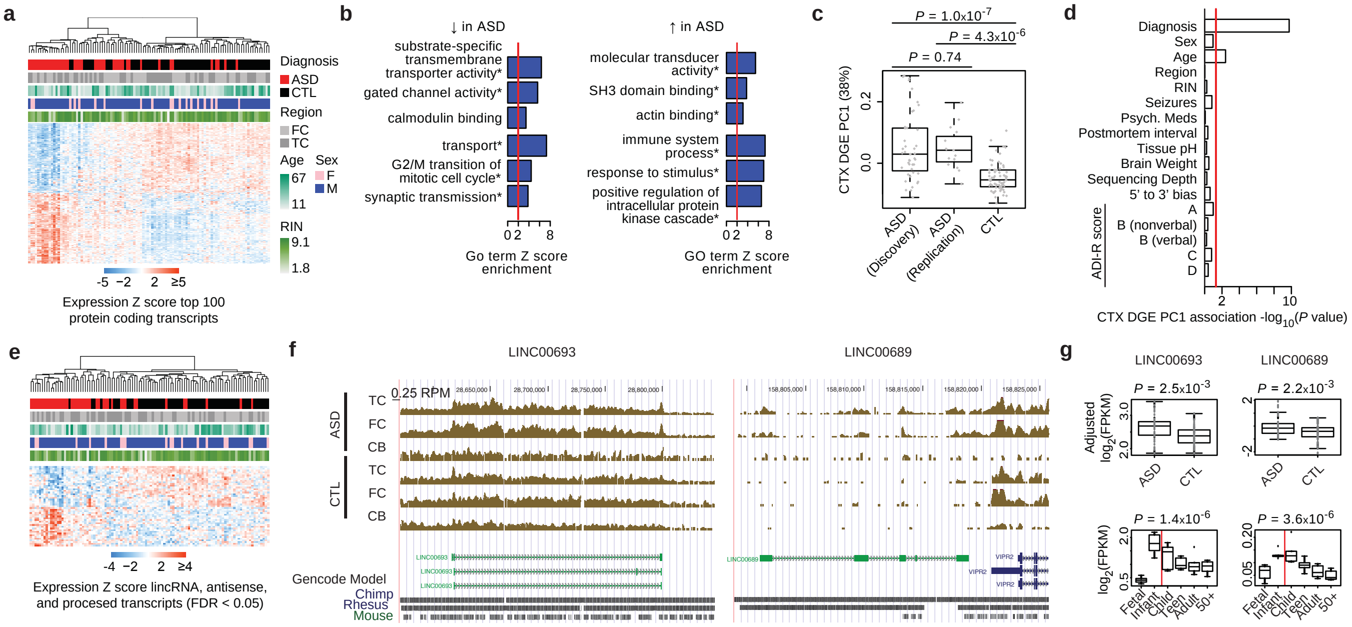
End Notes

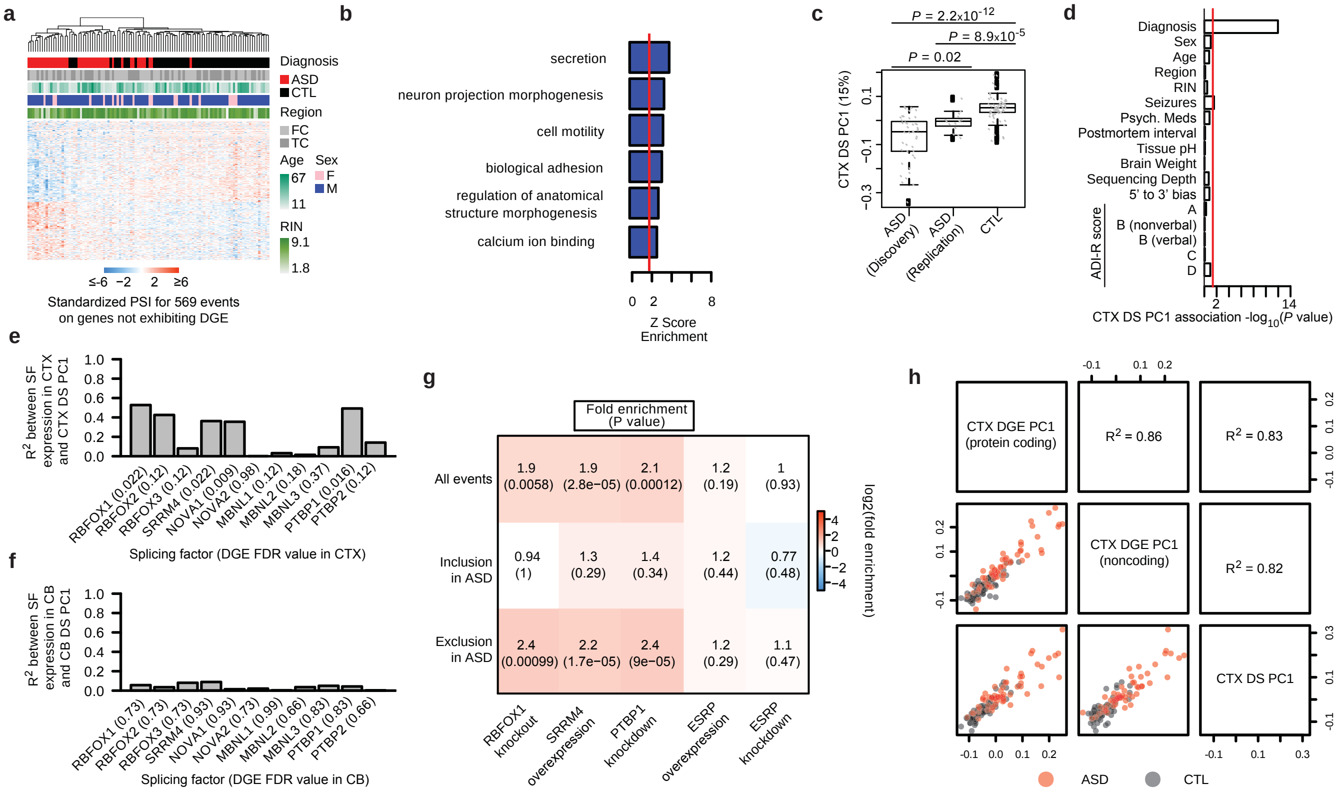
None.

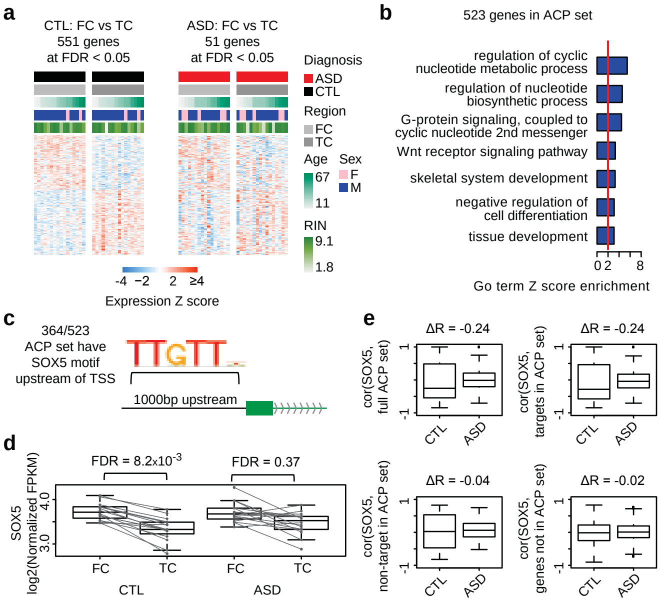
Author Contributions

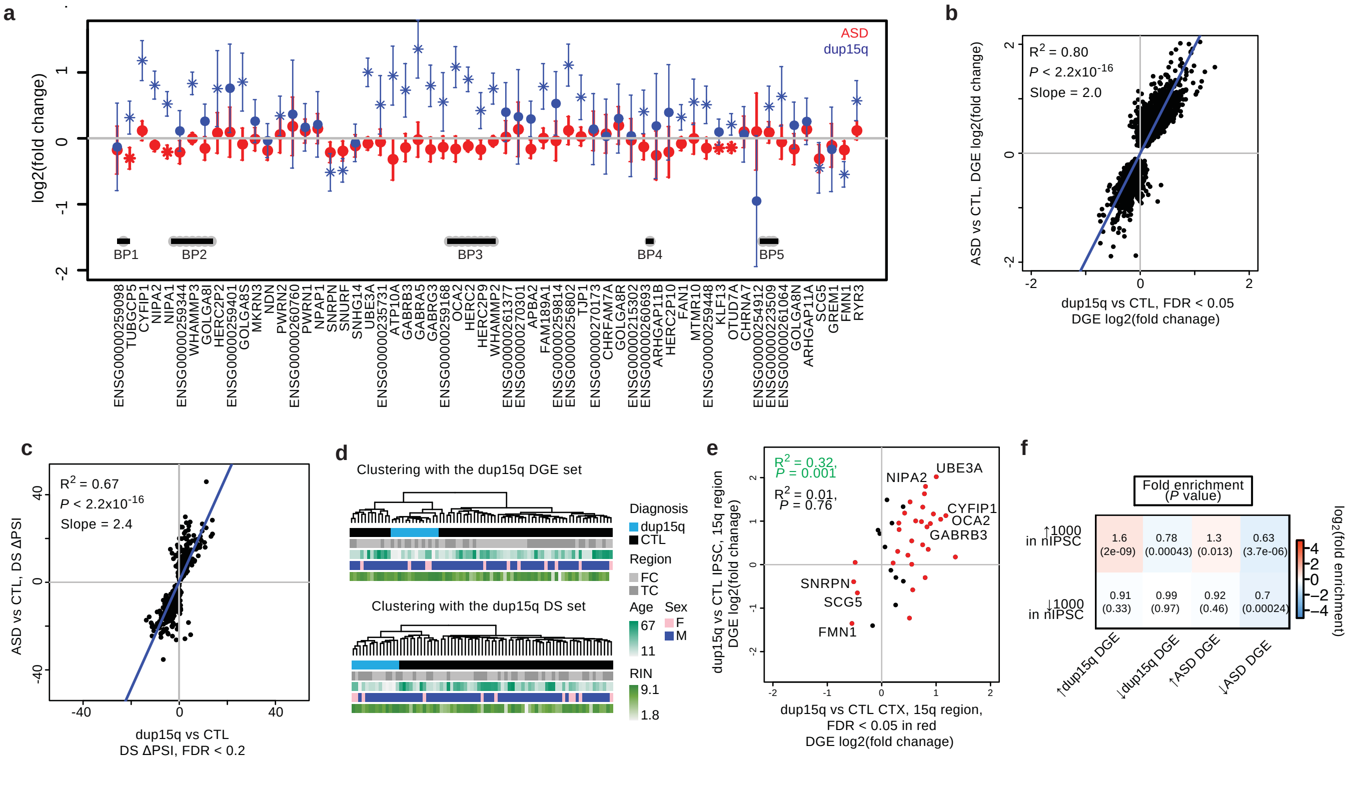
NNP, VS, and TGB performed dissections and RNA-seq analyses and differential gene expression analysis. NNP and VS performed splicing and co-expression network analysis. NNP and TGB performed analyses with Duplication 15q Syndrome individuals. NNP and MG reviewed clinical information and performed meta-analysis of ASD gene expression studies. VL and JKL performed

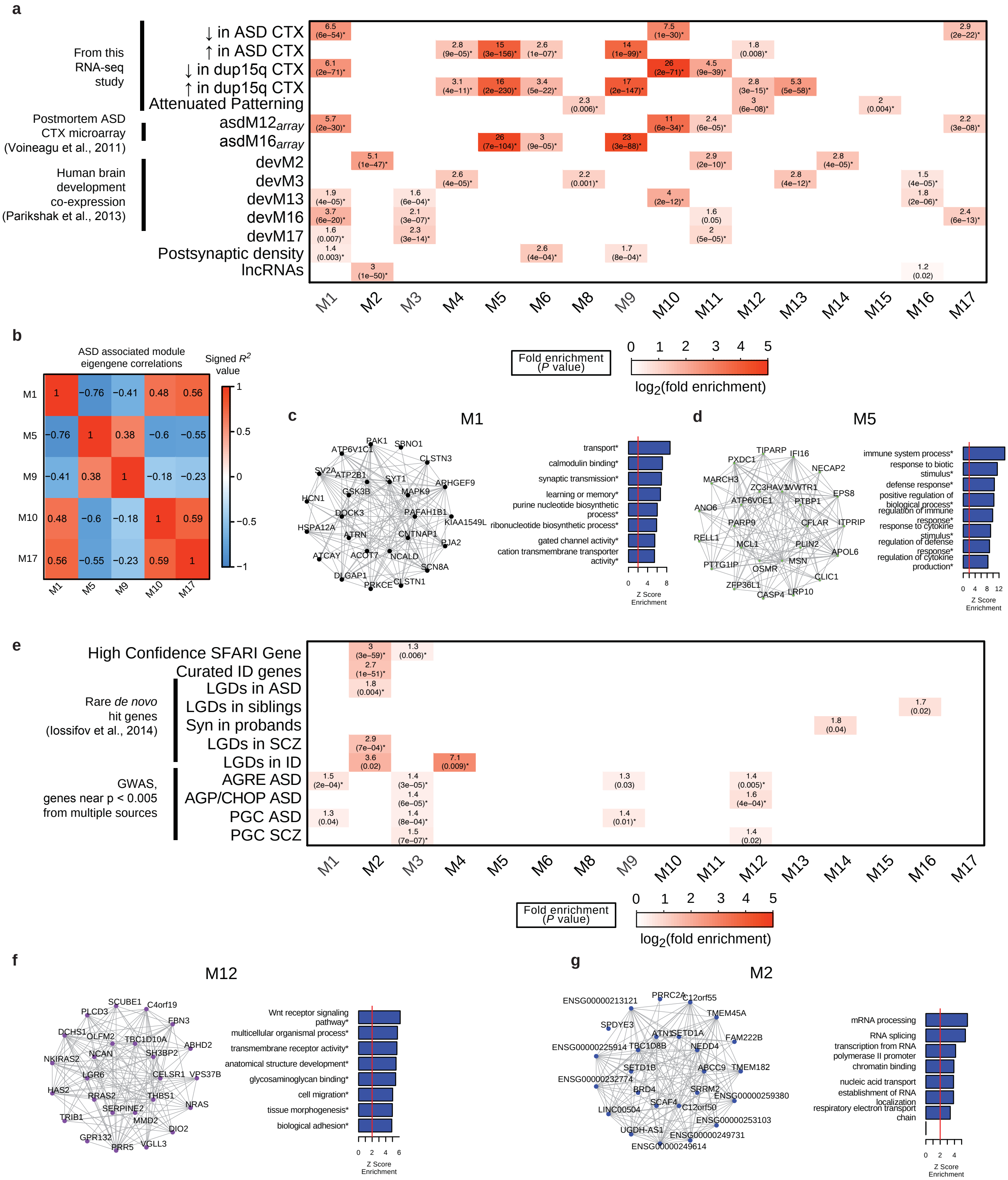
genotyping and CNV analysis on dup15q samples. VS performed validation experiments for gene and splicing level alterations in ASD. MI and BJB assisted with splicing analyses.  RJ performed dissections. SH provided guidance on differential gene expression and co-expression analyses. DHG provided guidance on all experiments and analyses. NNP and DHG wrote the manuscript. All authors contributed to revising and finalizing the manuscript.

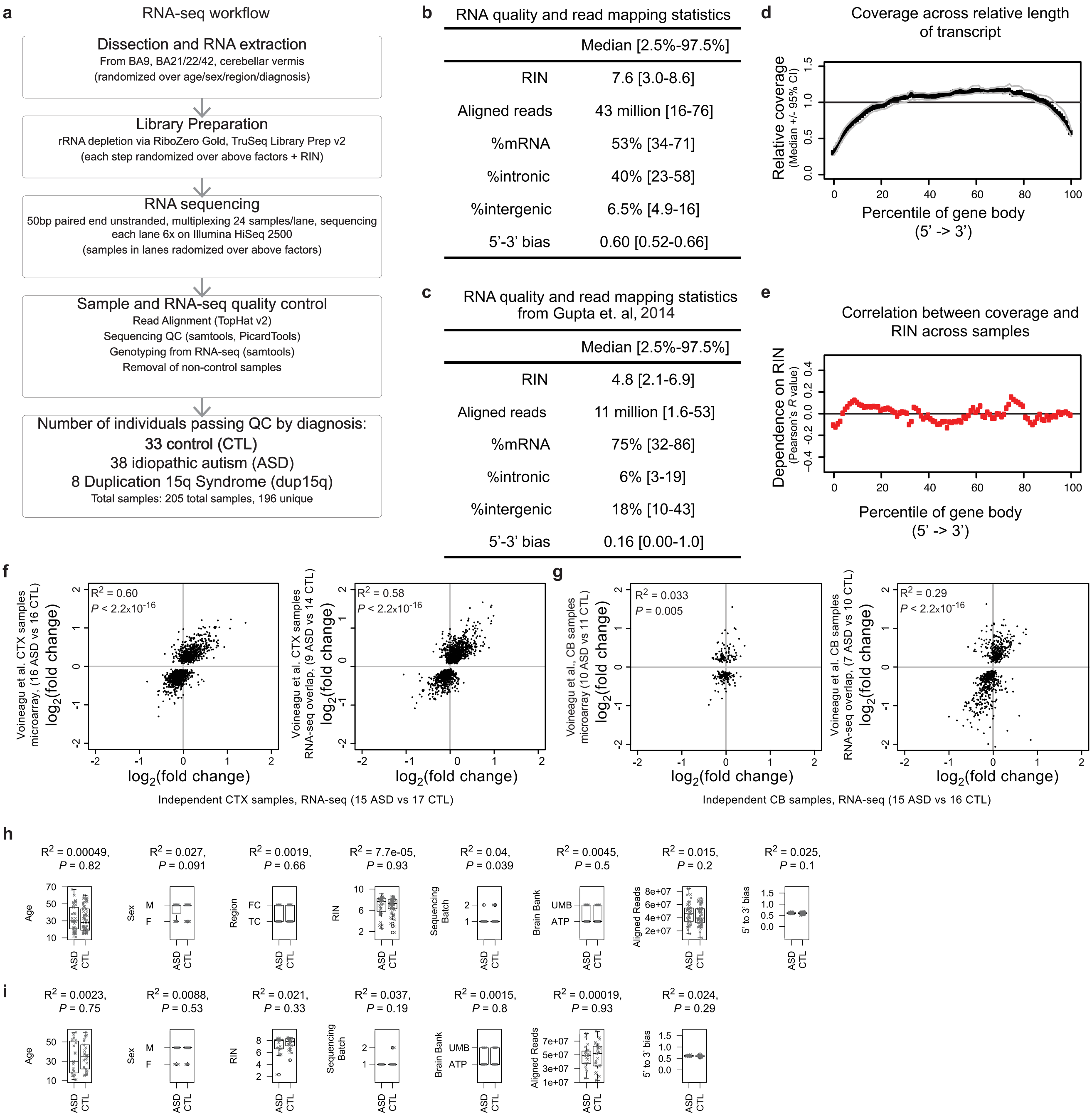

a
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
-5 −2 2 ≥5
Expression Z score top 100
protein coding transcripts
b
↓ in ASD
substrate-specific transmembrane transporter activity*
gated channel activity*
calmodulin binding
transport*
G2/M transition of mitotic cell cycle*
synaptic transmission*
0 2 8
Go term Z score enrichment
↑ in ASD
molecular transducer activity*
SH3 domain binding*
actin binding*
immune system process*
response to stimulus*
positive regulation of intracellular protein kinase cascade*
0 2 8
GO term Z score enrichment
c
P = 1.0x10-7
P = 4.3x10-6
P = 0.74
CTX DGE PC1 (38%)
0.2
0.0
ASD (Discovery)
ASD (Replication)
CTL
d
Diagnosis
Sex
Age
Region
RIN
Seizures
Psych. Meds
Postmortem interval
Tissue pH
Brain Weight
Sequencing Depth
5' to 3' bias
ADI-R score
A
B (nonverbal)
B (verbal)
C
D
2
10
CTX DGE PC1 association -log10(P value)
e
-4 −2 2 ≥4
Expression Z score lincRNA, antisense, and processed transcripts (FDR < 0.05)
f
LINC00693
LINC00689
0.25 RPM
ASD
TC
FC
CB
CTL
TC
FC
CB
Gencode Model
Chimp
Rhesus
Mouse
g
LINC00693
LINC00689
P = 2.5x10-3
P = 2.2x10-3
Adjusted log2(FPKM)
3.0
2.0
2
-2
ASD
CTL
P = 1.4x10-6
P = 3.6x10-6
log2(FPKM)
2.0
0.5
0.20
0.05
Fetal
Infant
Child
Teen
Adult
50+


a
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
≤-6 −2 2 ≥6
Standardized PSI for 569 events
on genes not exhibiting DGE
b
secretion
neuron projection morphogenesis
cell motility
biological adhesion
regulation of anatomical
structure morphogenesis
calcium ion binding
0 2 8
Z Score
Enrichment
c
P = 2.2x10-12
P = 8.9x10-5
P = 0.02
CTX DS PC1 (15%)
0.1
−0.1
−0.3
ASD (Discovery)
ASD (Replication)
CTL
d
Diagnosis
Sex
Age
Region
RIN
Seizures
Psych. Meds
Postmortem interval
Tissue pH
Brain Weight
Sequencing Depth
5' to 3' bias
ADI-R score
A
B (nonverbal)
B (verbal)
C
D
2
14
CTX DS PC1 association -log10(P value)
e
R2 between SF expression in CTX and CTX DS PC1
1.0
0.8
0.6
0.4
0.2
0.0
RBFOX1 (0.022)
RBFOX2 (0.12)
RBFOX3 (0.12)
SRRM4 (0.022)
NOVA1 (0.009)
NOVA2 (0.98)
MBNL1 (0.12)
MBNL2 (0.18)
MBNL3 (0.37)
PTBP1 (0.016)
PTBP2 (0.12)
Splicing factor (DGE FDR value in CTX)
f
R2 between SF expression in CB and CB DS PC1
1.0
0.8
0.6
0.4
0.2
0.0
RBFOX1 (0.73)
RBFOX2 (0.73)
RBFOX3 (0.73)
SRRM4 (0.93)
NOVA1 (0.93)
NOVA2 (0.73)
MBNL1 (0.99)
MBNL2 (0.66)
MBNL3 (0.83)
PTBP1 (0.83)
PTBP2 (0.66)
Splicing factor (DGE FDR value in CB)
g
Fold enrichment
(P value)
All events
Inclusion in ASD
Exclusion in ASD
1.9 (0.0058)
1.9 (2.8e−05)
2.1 (0.00012)
1.2 (0.19)
1 (0.93)
0.94 (1)
1.3 (0.29)
1.4 (0.34)
1.2 (0.44)
0.77 (0.48)
2.4 (0.00099)
2.2 (1.7e−05)
2.4 (9e−05)
1.2 (0.29)
1.1 (0.47)
RBFOX1 knockout
SRRM4 overexpression
PTBP1 knockdown
ESRP overexpression
ESRP knockdown
log2(fold enrichment)
4
2
0
−2
−4
h
CTX DGE PC1
(protein coding)
R2 = 0.86
R2 = 0.83
CTX DGE PC1
(noncoding)
R2 = 0.82
CTX DS PC1
-0.1
0.0
0.1
0.2
0.3
ASD
CTL


a
CTL: FC vs TC
551 genes
at FDR < 0.05
ASD: FC vs TC
51 genes
at FDR < 0.05
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
-4 −2 2 ≥4
Expression Z score
b
523 genes in ACP set
regulation of cyclic nucleotide metabolic process
regulation of nucleotide biosynthetic process
G-protein signaling, coupled to cyclic nucleotide 2nd messenger
Wnt receptor signaling pathway
skeletal system development
negative regulation of cell differentiation
tissue development
0 2 8
Go term Z score enrichment
c
364/523 ACP set have SOX5 motif upstream of TSS
TTGTT
1000bp upstream
d
FDR = 8.2x10^-3
FDR = 0.37
SOX5 log2(Normalized FPKM)
4.0
3.0
FC TC FC TC
CTL ASD
e
ΔR = -0.24
cor(SOX5, full ACP set)
ΔR = -0.24
cor(SOX5, targets in ACP set)
ΔR = -0.04
cor(SOX5, non-target in ACP set)
ΔR = -0.02
cor(SOX5, genes not in ACP set)
1
-1
CTL ASD


a
ASD
dup15q
log2(fold change)
1
0
-1
-2
BP1
BP2
BP3
BP4
BP5
ENSG00000259098
TUBGCP5
CYFIP1
NIPA2
NIPA1
ENSG00000259344
WHAMMP3
GOLGA8I
HERC2P2
ENSG00000259401
GOLGA8S
MKRN3
NDN
PWRN2
ENSG00000260760
PWRN1
NPAP1
SNRPN
SNURF
SNHG14
UBE3A
ENSG00000235731
ATP10A
GABRB3
GABRA5
GABRG3
ENSG00000259168
OCA2
HERC2
HERC2P9
WHAMMP2
ENSG00000261377
ENSG00000270301
APBA2
FAM189A1
ENSG00000259814
ENSG00000256802
TJP1
ENSG00000270173
CHRFAM7A
GOLGA8R
ENSG00000215302
ENSG00000260693
ARHGAP11B
HERC2P10
FAN1
MTMR10
ENSG00000259448
KLF13
OTUD7A
CHRNA7
ENSG00000254912
ENSG00000223509
ENSG00000261064
GOLGA8N
ARHGAP11A
SCG5
GREM1
FMN1
RYR3
b
R2 = 0.80
P < 2.2x10-16
Slope = 2.0
ASD vs CTL, DGE log2(fold change)
dup15q vs CTL, FDR < 0.05
DGE log2(fold chanage)
2
0
-2
c
R2 = 0.67
P < 2.2x10-16
Slope = 2.4
ASD vs CTL, DS ΔPSI
dup15q vs CTL
DS ΔPSI, FDR < 0.2
40
0
-40
d
Clustering with the dup15q DGE set
Clustering with the dup15q DS set
Diagnosis
dup15q
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
e
R2 = 0.32,
P = 0.001
R2 = 0.01,
P = 0.76
dup15q vs CTL IPSC, 15q region
DGE log2(fold chanage)
dup15q vs CTL CTX, 15q region,
FDR < 0.05 in red
DGE log2(fold chanage)
UBE3A
NIPA2
CYFIP1
OCA2
GABRB3
SNRPN
SCG5
FMN1
2
1
0
-1
-2
f
Fold enrichment
(P value)
↑1000
in nIPSC
↓1000
in nIPSC
1.6
(2e-09)
0.78
(0.00043)
1.3
(0.013)
0.63
(3.7e-06)
0.91
(0.33)
0.99
(0.97)
0.92
(0.46)
0.7
(0.00024)
↑dup15q DGE
↓dup15q DGE
↑ASD DGE
↓ASD DGE
log2(fold enrichment)
4
2
0
−2
−4


a
From this RNA-seq study
Postmortem ASD CTX microarray (Voineagu et al., 2011)
Human brain development co-expression (Parikshak et al., 2013)
↓ in ASD CTX
↑ in ASD CTX
↓ in dup15q CTX
↑ in dup15q CTX
Attenuated Patterning
asdM12array
asdM16array
devM2
devM3
devM13
devM16
devM17
Postsynaptic density
lncRNAs
6.5 (6e−54)*
7.5 (1e−30)*
2.9 (2e−22)*
2.8 (9e−05)*
15 (3e−156)*
2.6 (1e−07)*
14 (1e−99)*
1.8 (0.008)*
6.1 (2e−71)*
26 (2e−71)*
4.5 (9e−39)*
3.1 (4e−11)*
16 (2e−230)*
3.4 (5e−22)*
17 (2e−147)*
2.8 (3e−15)*
5.3 (5e−58)*
2.3 (0.006)*
3 (6e−08)*
2 (0.004)*
5.7 (2e−30)*
11 (6e−34)*
2.4 (6e−05)*
2.2 (3e−08)*
26 (7e−104)*
3 (9e−05)*
23 (3e−88)*
5.1 (1e−47)*
2.9 (2e−10)*
2.8 (4e−05)*
2.6 (4e−05)*
2.2 (0.001)*
2.8 (4e−12)*
1.5 (4e−05)*
1.9 (4e−05)*
1.6 (6e−04)*
4 (2e−12)*
1.8 (2e−06)*
3.7 (6e−20)*
2.1 (3e−07)*
1.6 (0.05)
2.4 (6e−13)*
1.6 (0.007)*
2.3 (3e−14)*
2 (5e−05)*
1.4 (0.003)*
2.6 (4e−04)*
1.7 (8e−04)*
3 (1e−50)*
1.2 (0.02)
M1 M2 M3 M4 M5 M6 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17
Fold enrichment (P value)
0 1 2 3 4 5
log2(fold enrichment)
b
ASD associated module eigengene correlations
Signed R2 value
1 0.5 0 −0.5 −1
M1 M5 M9 M10 M17
M1: 1 −0.76 −0.41 0.48 0.56
M5: −0.76 1 0.38 −0.6 −0.55
M9: −0.41 0.38 1 −0.18 −0.23
M10: 0.48 −0.6 −0.18 1 0.59
M17: 0.56 −0.55 −0.23 0.59 1
c
M1
PAK1 SBNO1 ATP6V1C1 CLSTN3 SV2A ATP2B1 SYT1 ARHGEF9 HCN1 GSK3B MAPK9 DOCK3 PAFAH1B1 KIAA1549L HSPA12A ATRN CNTNAP1 PJA2 ATCAY ACOT7 NCALD SCN8A DLGAP1 PRKCE CLSTN1
transport*
calmodulin binding*
synaptic transmission*
learning or memory*
purine nucleotide biosynthetic process*
ribonucleotide biosynthetic process*
gated channel activity*
cation transmembrane transporter activity*
0 4 8
Z Score Enrichment
d
M5
TIPARP IFI16 PXDC1 NECAP2 MARCH3 ZC3HAV1 WWTR1 EPS8 ATP6V0E1 PTBP1 ANO6 PARP9 CFLAR ITPRIP RELL1 MCL1 PLIN2 APOL6 PTTG1IP OSMR MSN CLIC1 ZFP36L1 CASP4 LRP10
immune system process*
response to biotic stimulus*
defense response*
positive regulation of biological process*
regulation of immune response*
response to cytokine stimulus*
regulation of defense response*
regulation of cytokine production*
0 4 8 12
Z Score Enrichment
e
High Confidence SFARI Gene
Curated ID genes
Rare de novo hit genes (Iossifov et al., 2014)
LGDs in ASD
LGDs in siblings
Syn in probands
LGDs in SCZ
LGDs in ID
GWAS, genes near p < 0.005 from multiple sources
AGRE ASD
AGP/CHOP ASD
PGC ASD
PGC SCZ
3 (3e−59)*
1.3 (0.006)*
2.7 (1e−51)*
1.8 (0.004)*
1.7 (0.02)
1.8 (0.04)
2.9 (7e−04)*
3.6 (0.02)
7.1 (0.009)*
1.5 (2e−04)*
1.4 (3e−05)*
1.3 (0.03)
1.4 (0.005)*
1.4 (6e−05)*
1.6 (4e−04)*
1.3 (0.04)
1.4 (8e−04)*
1.4 (0.01)*
1.5 (7e−07)*
1.4 (0.02)
M1 M2 M3 M4 M5 M6 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17
Fold enrichment (P value)
0 1 2 3 4 5
log2(fold enrichment)
f
M12
SCUBE1 C4orf19 PLCD3 FBN3 DCHS1 OLFM2 TBC1D10A ABHD2 NKIRAS2 NCAN SH3BP2 LGR6 CELSR1 VPS37B HAS2 RRAS2 THBS1 NRAS SERPINE2 TRIB1 MMD2 DIO2 GPR132 PRR5 VGLL3
Wnt receptor signaling pathway*
multicellular organismal process*
transmembrane receptor activity*
anatomical structure development*
glycosaminoglycan binding*
cell migration*
tissue morphogenesis*
biological adhesion*
0 2 4 6
Z Score Enrichment
g
M2
ENSG00000213121 PRRC2A C12orf55 TMEM45A SPDYE3 ATN1 SETD1A FAM222B ENSG00000225914 TBC1D8B NEDD4 SETD1B ABCC9 TMEM182 ENSG00000232774 BRD4 SRRM2 ENSG00000259380 SCAF4 C12orf50 LINC00504 ENSG00000253103 UGDH-AS1 ENSG00000249731 ENSG00000249614
mRNA processing
RNA splicing
transcription from RNA polymerase II promoter
chromatin binding
nucleic acid transport
establishment of RNA localization
respiratory electron transport chain
0 2 4
Z Score Enrichment



RNA-seq workflow

Dissection and RNA extraction
From BA9, BA21/22/42, cerebellar vermis
(randomized over age/sex/region/diagnosis)

Library Preparation
rRNA depletion via RiboZero Gold, TruSeq Library Prep v2
(each step randomized over above factors + RIN)

RNA sequencing
50bp paired end unstranded, multiplexing 24 samples/lane, sequencing each lane 6x on Illumina HiSeq 2500
(samples in lanes radomized over above factors)

Sample and RNA-seq quality control
Read Alignment (TopHat v2)
Sequencing QC (samtools, PicardTools)
Genotyping from RNA-seq (samtools)
Removal of non-control samples

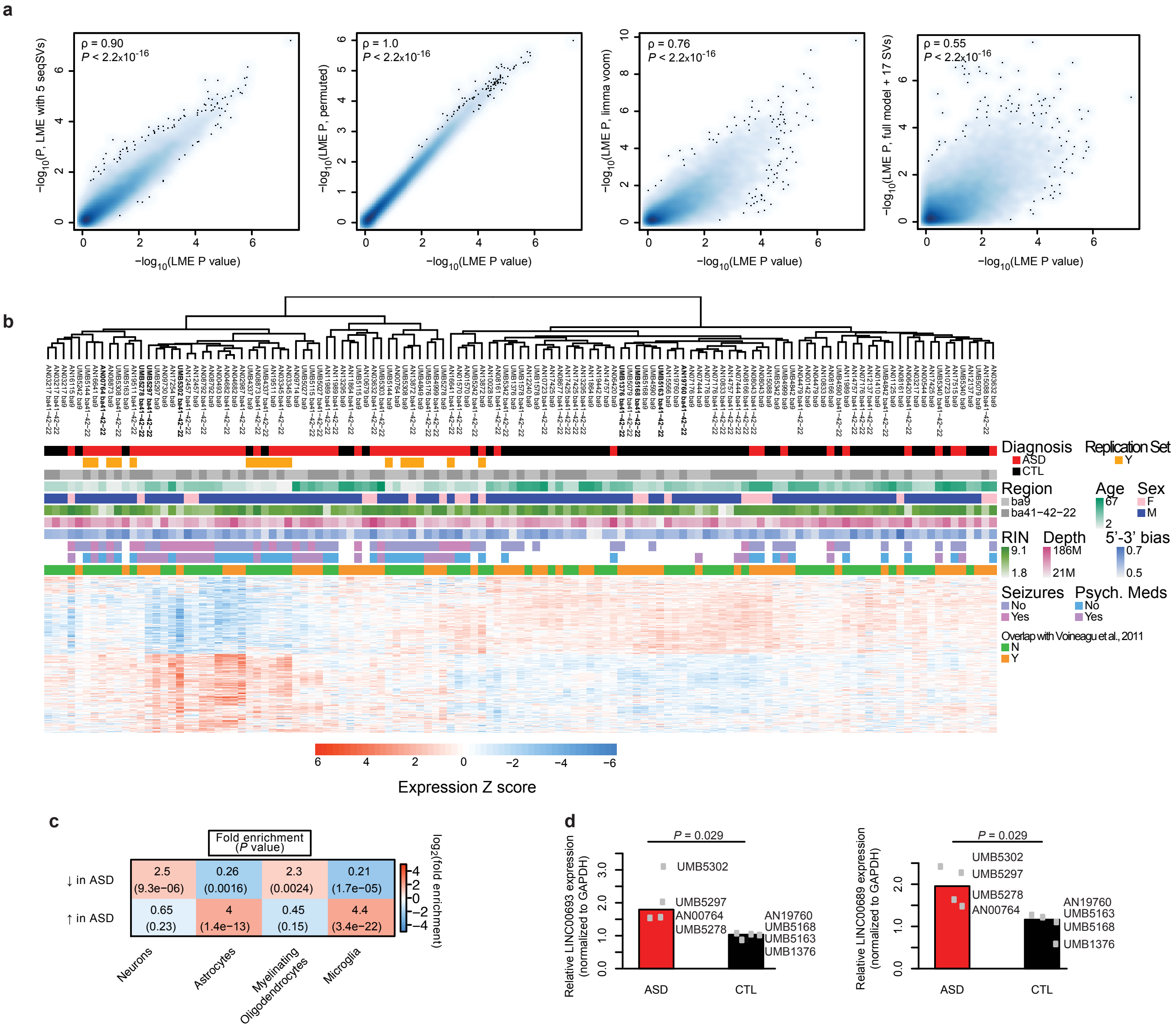
Number of individuals passing QC by diagnosis:
33 control (CTL)
38 idiopathic autism (ASD)
8 Duplication 15q Syndrome (dup15q)
Total samples: 205 total samples, 196 unique

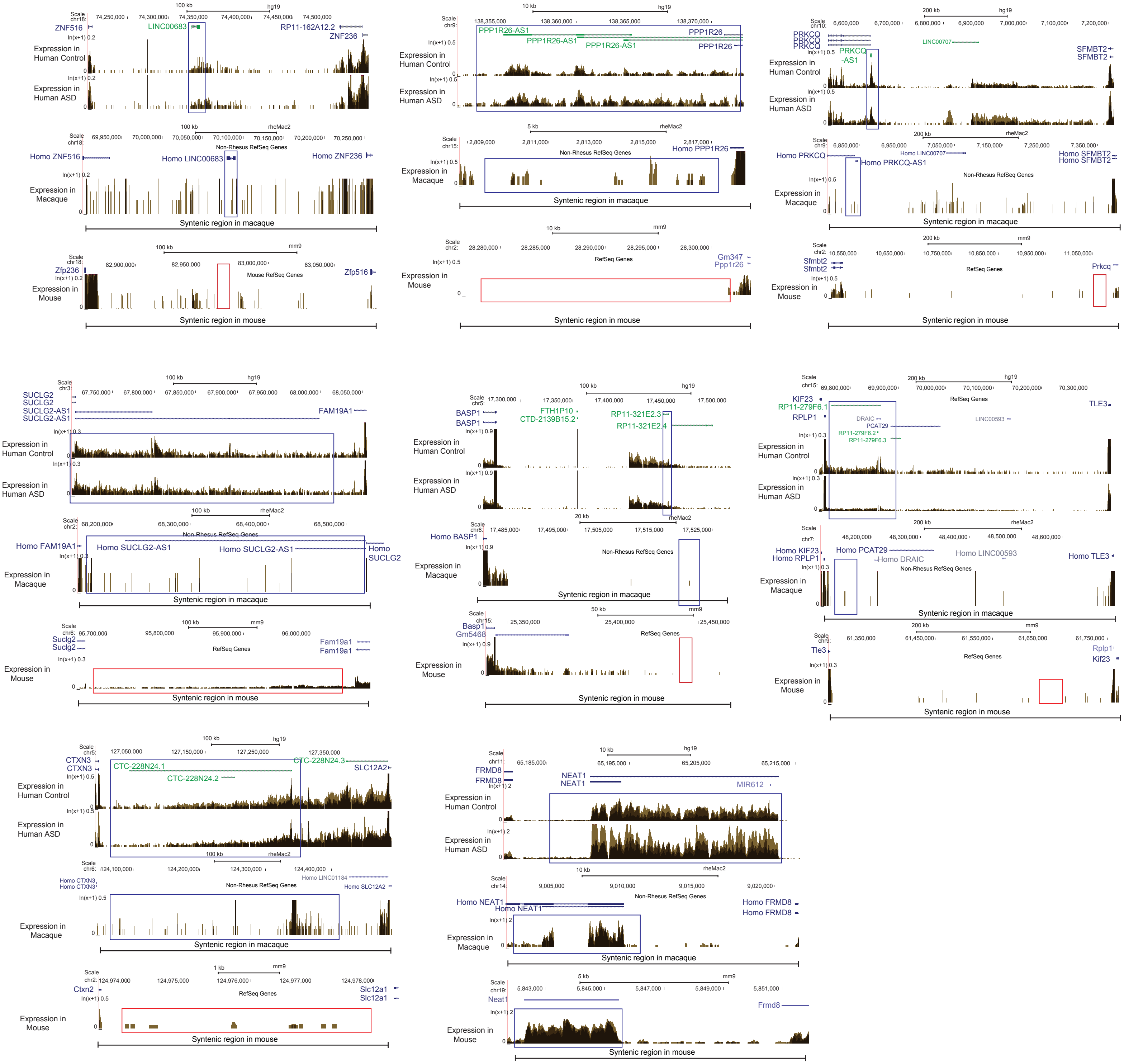
RNA quality and read mapping statistics

| | Median [2.5%-97.5%] |
|---|---|
| RIN | 7.6 [3.0-8.6] |
| Aligned reads | 43 million [16-76] |
| %mRNA | 53% [34-71] |
| %intronic | 40% [23-58] |
| %intergenic | 6.5% [4.9-16] |
| 5’-3’ bias | 0.60 [0.52-0.66] |

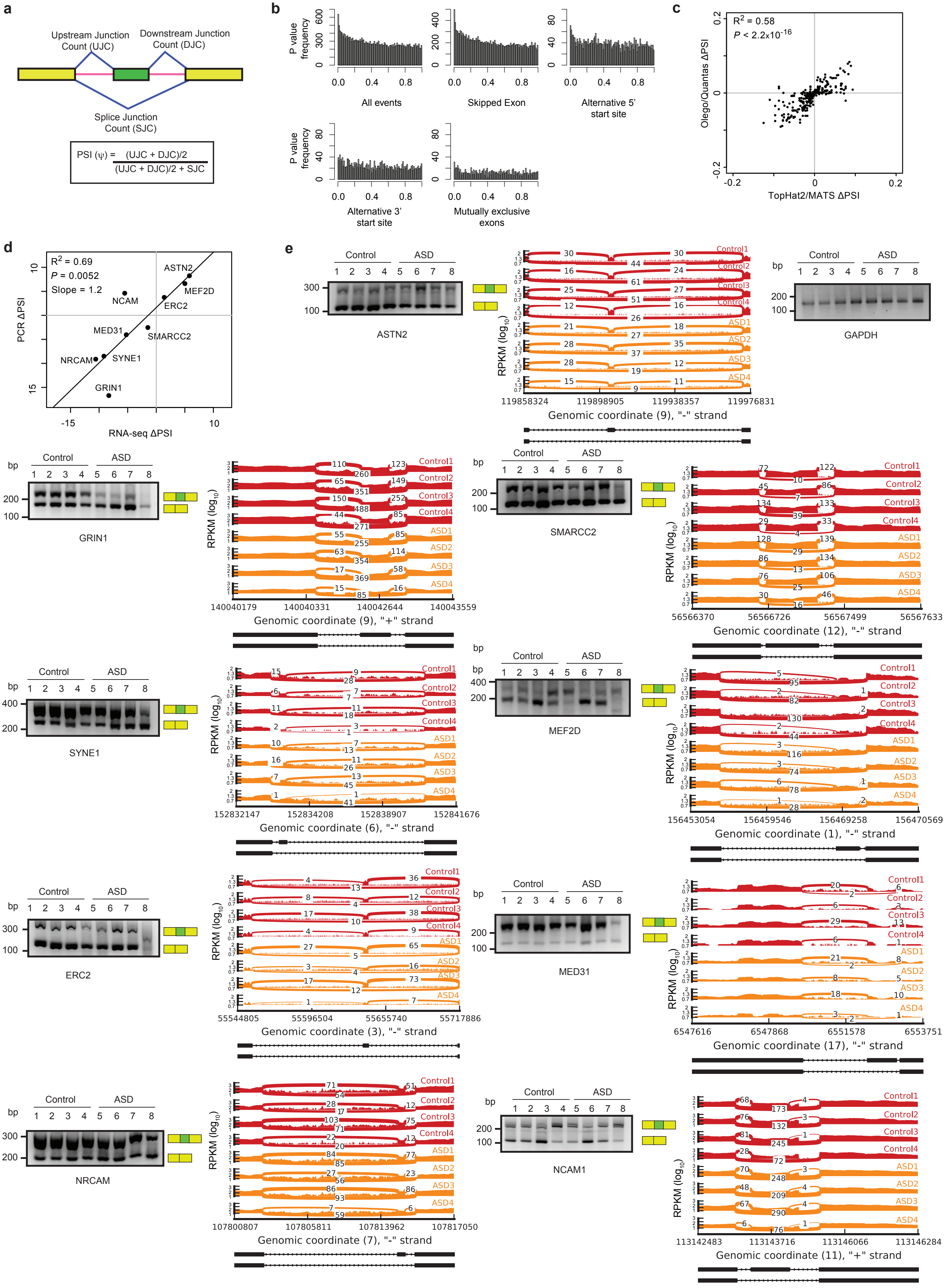
RNA quality and read mapping statistics from Gupta et. al, 2014

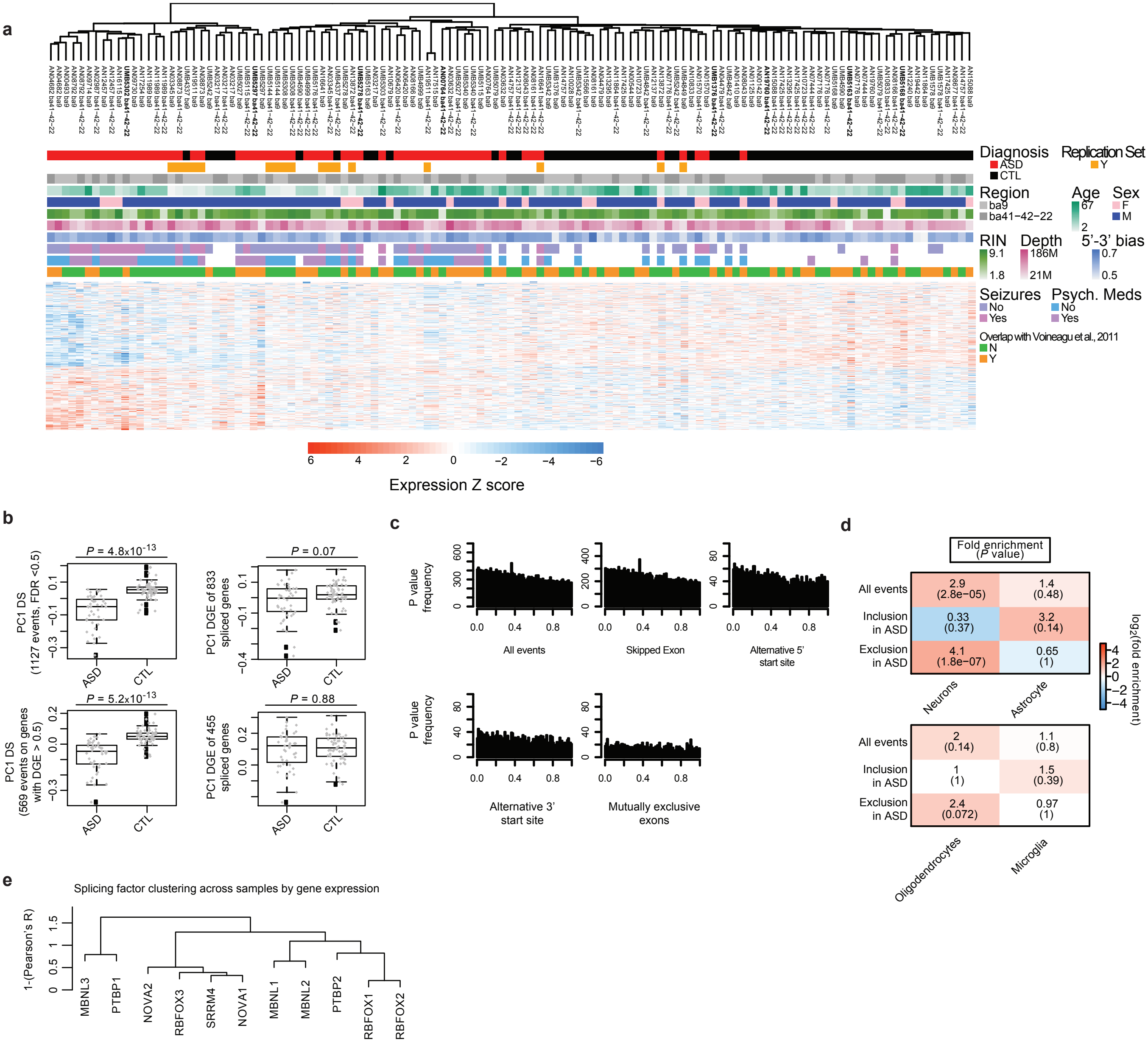
| | Median [2.5%-97.5%] |
|---|---|
| RIN | 4.8 [2.1-6.9] |
| Aligned reads | 11 million [1.6-53] |
| %mRNA | 75% [32-86] |
| %intronic | 6% [3-19] |
| %intergenic | 18% [10-43] |
| 5’-3’ bias | 0.16 [0.00-1.0] |

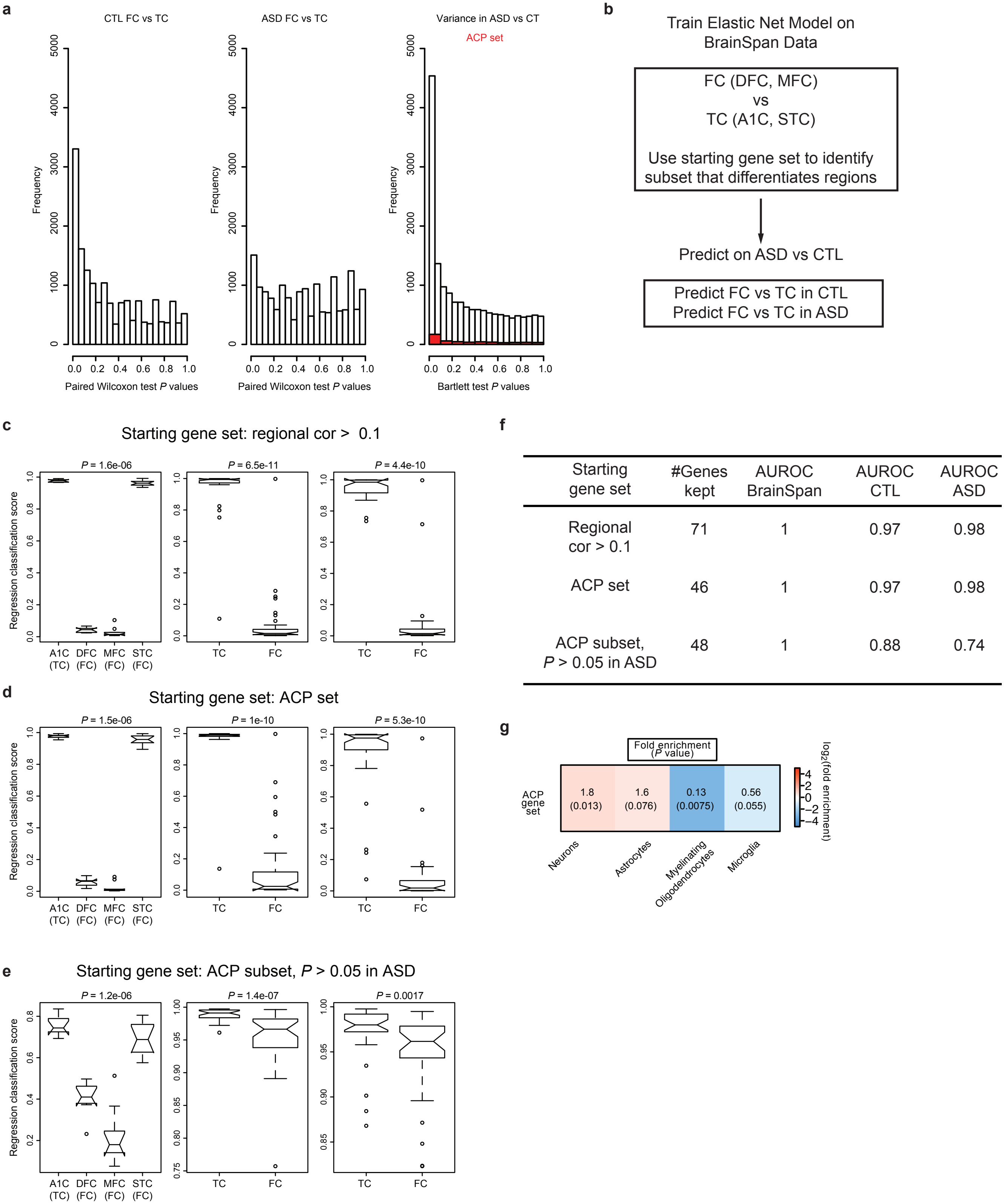

a
ρ = 0.90
P < 2.2x10^-16
−log10(P, LME with 5 seqSVs)
−log10(LME P value)
ρ = 1.0
P < 2.2x10^-16
−log10(LME P, permuted)
−log10(LME P value)
ρ = 0.76
P < 2.2x10^-16
−log10(LME P, limma voom)
−log10(LME P value)
ρ = 0.55
P < 2.2x10^-16
−log10(LME P, full model + 17 SVs)
−log10(LME P value)
b
Diagnosis
ASD
CTL
Replication Set
Y
Region
ba9
ba41−42−22
Age
67
2
Sex
F
M
RIN
9.1
1.8
Depth
186M
21M
5'-3' bias
0.7
0.5
Seizures
No
Yes
Psych. Meds
No
Yes
Overlap with Voineagu et al., 2011
N
Y
6
4
2
0
−2
−4
−6
Expression Z score
c
Fold enrichment
(P value)
↓ in ASD
2.5 (9.3e−06)
0.26 (0.0016)
2.3 (0.0024)
0.21 (1.7e−05)
↑ in ASD
0.65 (0.23)
4 (1.4e−13)
0.45 (0.15)
4.4 (3.4e−22)
Neurons
Astrocytes
Myelinating Oligodendrocytes
Microglia
log2(fold enrichment)
4
2
0
−2
−4
d
P = 0.029
Relative LINC00693 expression (normalized to GAPDH)
UMB5302
UMB5297
AN00764
UMB5278
AN19760
UMB5168
UMB5163
UMB1376
ASD
CTL
P = 0.029
Relative LINC00689 expression (normalized to GAPDH)
UMB5302
UMB5297
UMB5278
AN00764
AN19760
UMB5163
UMB5168
UMB1376
ASD
CTL



ZNF516
LINC00683
RP11-162A12.2
ZNF236
Expression in Human Control
Expression in Human ASD
Homo ZNF516
Homo LINC00683
Homo ZNF236
Non-Rhesus RefSeq Genes
Expression in Macaque
Syntenic region in macaque
Zfp236
Zfp516
Mouse RefSeq Genes
Expression in Mouse
Syntenic region in mouse
PPP1R26-AS1
PPP1R26
Homo PPP1R26
Gm347
Ppp1r26
RefSeq Genes
PRKCQ
PRKCQ-AS1
LINC00707
SFMBT2
Homo PRKCQ
Homo PRKCQ-AS1
Homo LINC00707
Homo SFMBT2
Sfmbt2
Prkcq
SUCLG2
SUCLG2-AS1
FAM19A1
Homo FAM19A1
Homo SUCLG2-AS1
Homo SUCLG2
Suclg2
Fam19a1
BASP1
FTH1P10
CTD-2139B15.2
RP11-321E2.3
RP11-321E2.4
Homo BASP1
Basp1
Gm5468
KIF23
RP11-279F6.1
RPLP1
DRAIC
PCAT29
RP11-279F6.2
RP11-279F6.3
LINC00593
TLE3
Homo KIF23
Homo RPLP1
Homo PCAT29
Homo DRAIC
Homo LINC00593
Homo TLE3
Tle3
Rplp1
Kif23
CTXN3
CTC-228N24.1
CTC-228N24.2
CTC-228N24.3
SLC12A2
Homo CTXN3
Homo LINC01184
Homo SLC12A2
Ctxn2
Slc12a1
FRMD8
NEAT1
MIR612
Homo NEAT1
Homo FRMD8
Neat1
Frmd8
hg19
rheMac2
mm9
Scale
ln(x+1)


a
Upstream Junction Count (UJC)
Downstream Junction Count (DJC)
Splice Junction Count (SJC)
PSI (ψ) = (UJC + DJC)/2 / ((UJC + DJC)/2 + SJC)
b
P value frequency
All events
Skipped Exon
Alternative 5' start site
Alternative 3' start site
Mutually exclusive exons
c
R² = 0.58
P < 2.2x10⁻¹⁶
Olego/Quantas ΔPSI
TopHat2/MATS ΔPSI
d
R² = 0.69
P = 0.0052
Slope = 1.2
PCR ΔPSI
RNA-seq ΔPSI
ASTN2
MEF2D
NCAM
ERC2
MED31
SMARCC2
NRCAM
SYNE1
GRIN1
e
Control
ASD
bp
ASTN2
GAPDH
GRIN1
SMARCC2
SYNE1
MEF2D
ERC2
MED31
NRCAM
NCAM1
RPKM (log₁₀)
Control1
Control2
Control3
Control4
ASD1
ASD2
ASD3
ASD4
Genomic coordinate (9), "-" strand
Genomic coordinate (9), "+" strand
Genomic coordinate (12), "-" strand
Genomic coordinate (6), "-" strand
Genomic coordinate (1), "-" strand
Genomic coordinate (3), "-" strand
Genomic coordinate (17), "-" strand
Genomic coordinate (7), "-" strand
Genomic coordinate (11), "+" strand
119858324
119898905
119938357
119976831
140040179
140040331
140042644
140043559
56566370
56566726
56567499
56567633
152832147
152834208
152838907
152841676
156453054
156459546
156469258
156470569
55544805
55596504
55655740
55717886
6547616
6547868
6551578
6553751
107800807
107805811
107813962
107817050
113142483
113143716
113146066
113146284


a
Diagnosis
ASD
CTL
Replication Set
Y
Region
ba9
ba41-42-22
Age
67
2
Sex
F
M
RIN
9.1
1.8
Depth
186M
21M
5'-3' bias
0.7
0.5
Seizures
No
Yes
Psych. Meds
No
Yes
Overlap with Voineagu et al., 2011
N
Y
6
4
2
0
−2
−4
−6
Expression Z score
b
P = 4.8x10^-13
PC1 DS
(1127 events, FDR <0.5)
P = 0.07
PC1 DGE of 833
spliced genes
P = 5.2x10^-13
PC1 DS
(569 events on genes
with DGE > 0.5)
P = 0.88
PC1 DGE of 455
spliced genes
ASD
CTL
c
P value
frequency
All events
Skipped Exon
Alternative 5'
start site
Alternative 3'
start site
Mutually exclusive
exons
d
Fold enrichment
(P value)
All events
Inclusion
in ASD
Exclusion
in ASD
2.9
(2.8e−05)
1.4
(0.48)
0.33
(0.37)
3.2
(0.14)
4.1
(1.8e−07)
0.65
(1)
Neurons
Astrocyte
2
(0.14)
1.1
(0.8)
1
(1)
1.5
(0.39)
2.4
(0.072)
0.97
(1)
Oligodendrocytes
Microglia
log2(fold enrichment)
4
2
0
−2
−4
e
Splicing factor clustering across samples by gene expression
1-(Pearson's R)
0
0.5
1
1.5
MBNL3
PTBP1
NOVA2
RBFOX3
SRRM4
NOVA1
MBNL1
MBNL2
PTBP2
RBFOX1
RBFOX2



Train Elastic Net Model on BrainSpan Data

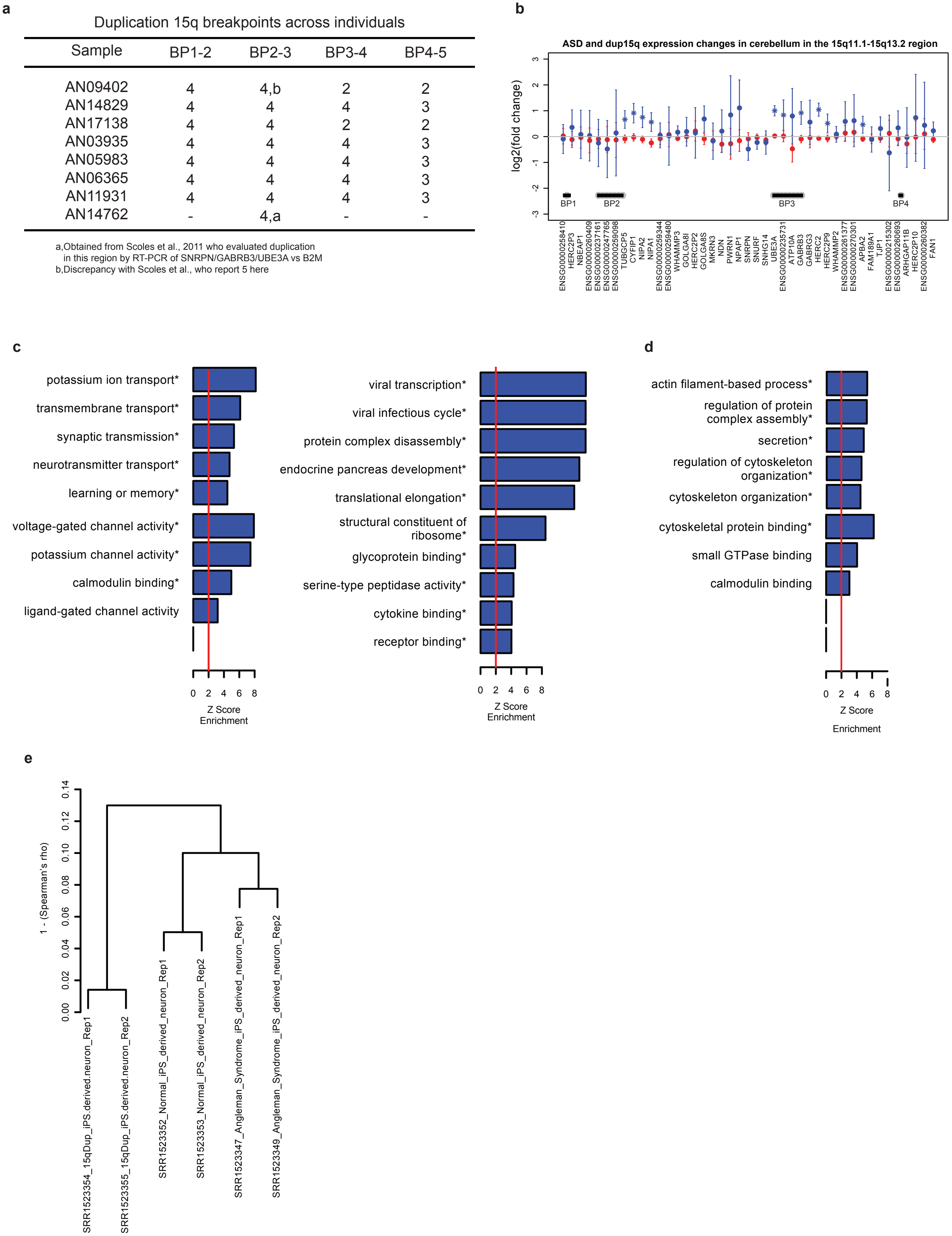
FC (DFC, MFC)
vs
TC (A1C, STC)

Use starting gene set to identify subset that differentiates regions

Predict on ASD vs CTL

Predict FC vs TC in CTL
Predict FC vs TC in ASD

| Starting gene set | #Genes kept | AUROC BrainSpan | AUROC CTL | AUROC ASD |
|---|---|---|---|---|
| Regional cor > 0.1 | 71 | 1 | 0.97 | 0.98 |
| ACP set | 46 | 1 | 0.97 | 0.98 |
| ACP subset, *P* > 0.05 in ASD | 48 | 1 | 0.88 | 0.74 |

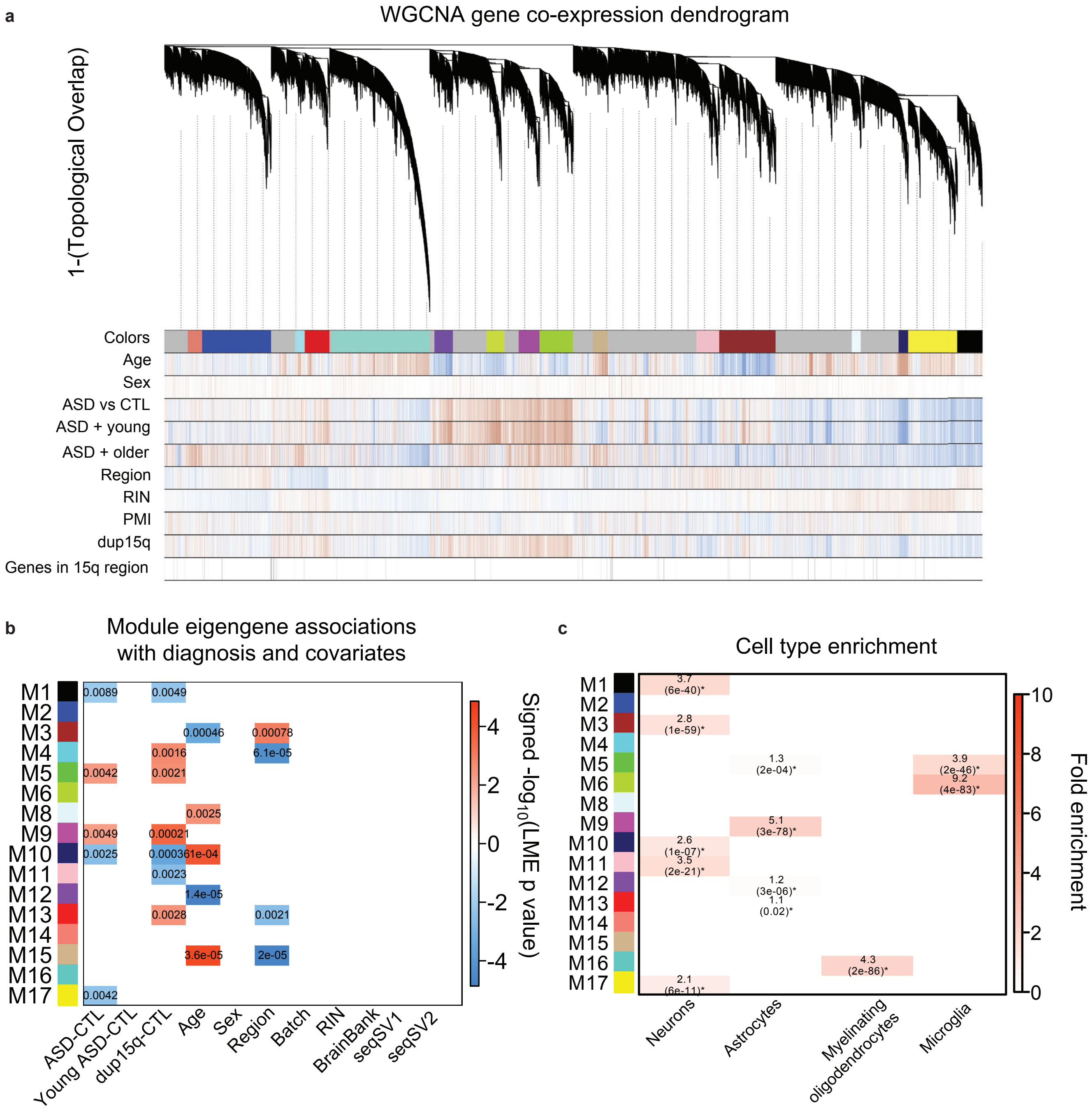


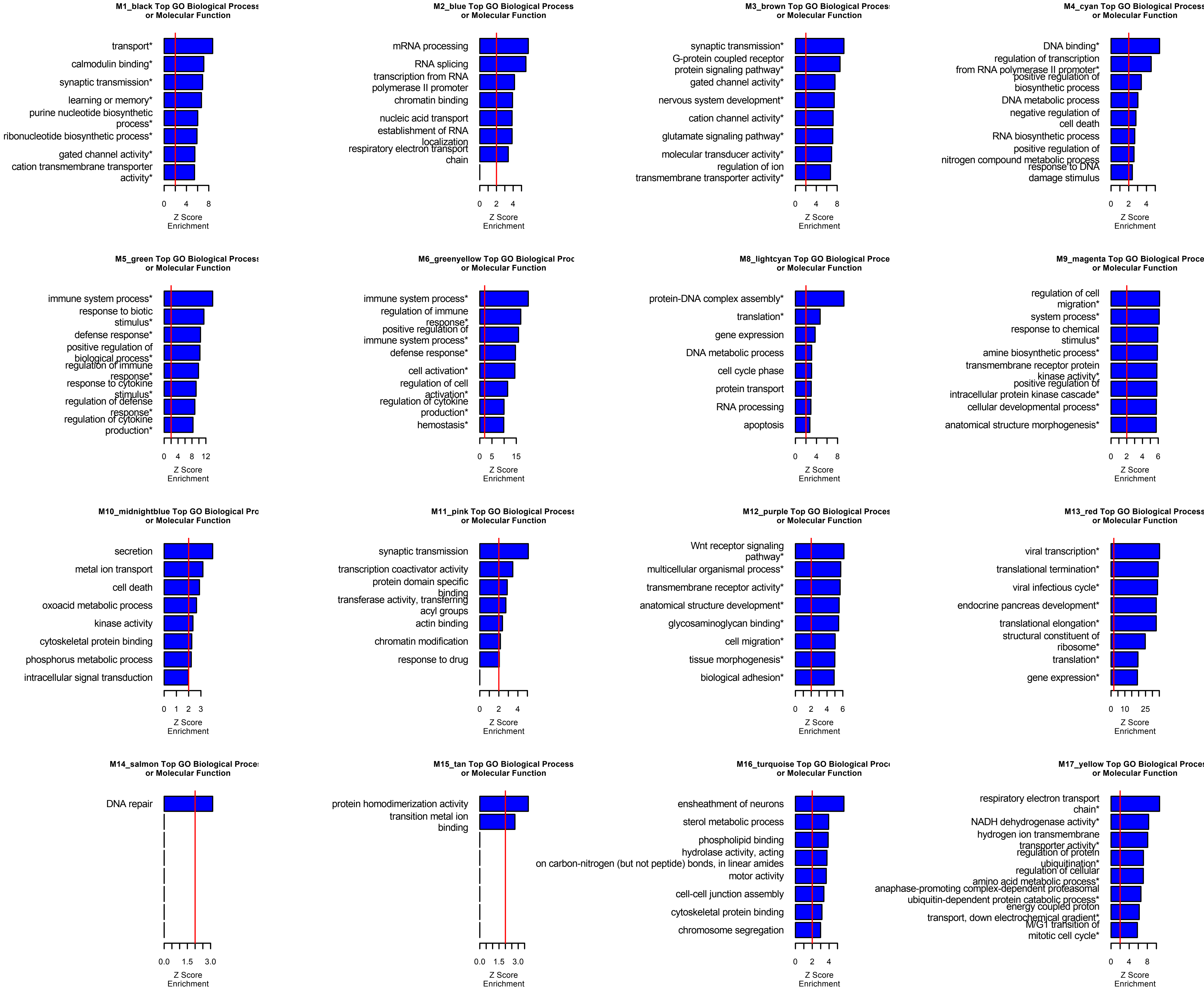
Duplication 15q breakpoints across individuals

| Sample | BP1-2 | BP2-3 | BP3-4 | BP4-5 |
|---|---|---|---|---|
| AN09402 | 4 | 4,b | 2 | 2 |
| AN14829 | 4 | 4 | 4 | 3 |
| AN17138 | 4 | 4 | 2 | 2 |
| AN03935 | 4 | 4 | 4 | 3 |
| AN05983 | 4 | 4 | 4 | 3 |
| AN06365 | 4 | 4 | 4 | 3 |
| AN11931 | 4 | 4 | 4 | 3 |
| AN14762 | - | 4,a | - | - |

a,Obtained from Scoles et al., 2011 who evaluated duplication in this region by RT-PCR of SNRPN/GABRB3/UBE3A vs B2M

b,Discrepancy with Scoles et al., who report 5 here


a
WGCNA gene co-expression dendrogram
1-(Topological Overlap)
Colors
Age
Sex
ASD vs CTL
ASD + young
ASD + older
Region
RIN
PMI
dup15q
Genes in 15q region
b
Module eigengene associations with diagnosis and covariates
M1
M2
M3
M4
M5
M6
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
0.0089
0.0049
0.00046
0.00078
0.0016
6.1e-05
0.0042
0.0021
0.0025
0.0049
0.00021
0.0025
0.00036
1e-04
0.0023
1.4e-05
0.0028
0.0021
3.6e-05
2e-05
0.0042
ASD-CTL
Young ASD-CTL
dup15q-CTL
Age
Sex
Region
Batch
RIN
BrainBank
seqSV1
seqSV2
4
2
0
-2
-4
Signed -log10(LME p value)
c
Cell type enrichment
M1
M2
M3
M4
M5
M6
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
3.7 (6e-40)*
2.8 (1e-59)*
1.3 (2e-04)*
3.9 (2e-46)*
9.2 (4e-83)*
5.1 (3e-78)*
2.6 (1e-07)*
3.5 (2e-21)*
1.2 (3e-06)*
1.1 (0.02)*
4.3 (2e-86)*
2.1 (6e-11)*
Neurons
Astrocytes
Myelinating oligodendrocytes
Microglia
10
8
6
4
2
0
Fold enrichment


M1_black Top GO Biological Process
or Molecular Function
transport*
calmodulin binding*
synaptic transmission*
learning or memory*
purine nucleotide biosynthetic process*
ribonucleotide biosynthetic process*
gated channel activity*
cation transmembrane transporter activity*
0
4
8
Z Score Enrichment
M2_blue Top GO Biological Process
or Molecular Function
mRNA processing
RNA splicing
transcription from RNA polymerase II promoter
chromatin binding
nucleic acid transport
establishment of RNA localization
respiratory electron transport chain
0
2
4
Z Score Enrichment
M3_brown Top GO Biological Proces
or Molecular Function
synaptic transmission*
G-protein coupled receptor protein signaling pathway*
gated channel activity*
nervous system development*
cation channel activity*
glutamate signaling pathway*
molecular transducer activity*
regulation of ion transmembrane transporter activity*
0
4
8
Z Score Enrichment
M4_cyan Top GO Biological Process
or Molecular Function
DNA binding*
regulation of transcription from RNA polymerase II promoter*
positive regulation of biosynthetic process
DNA metabolic process
negative regulation of cell death
RNA biosynthetic process
positive regulation of nitrogen compound metabolic process
response to DNA damage stimulus
0
2
4
Z Score Enrichment
M5_green Top GO Biological Process
or Molecular Function
immune system process*
response to biotic stimulus*
defense response*
positive regulation of biological process*
regulation of immune response*
response to cytokine stimulus*
regulation of defense response*
regulation of cytokine production*
0
4
8
12
Z Score Enrichment
M6_greenyellow Top GO Biological Proc
or Molecular Function
immune system process*
regulation of immune response*
positive regulation of immune system process*
defense response*
cell activation*
regulation of cell activation*
regulation of cytokine production*
hemostasis*
0
5
15
Z Score Enrichment
M8_lightcyan Top GO Biological Proce
or Molecular Function
protein-DNA complex assembly*
translation*
gene expression
DNA metabolic process
cell cycle phase
protein transport
RNA processing
apoptosis
0
4
8
Z Score Enrichment
M9_magenta Top GO Biological Proce
or Molecular Function
regulation of cell migration*
system process*
response to chemical stimulus*
amine biosynthetic process*
transmembrane receptor protein kinase activity*
positive regulation of intracellular protein kinase cascade*
cellular developmental process*
anatomical structure morphogenesis*
0
2
4
6
Z Score Enrichment
M10_midnightblue Top GO Biological Pro
or Molecular Function
secretion
metal ion transport
cell death
oxoacid metabolic process
kinase activity
cytoskeletal protein binding
phosphorus metabolic process
intracellular signal transduction
0
1
2
3
Z Score Enrichment
M11_pink Top GO Biological Process
or Molecular Function
synaptic transmission
transcription coactivator activity
protein domain specific binding
transferase activity, transferring acyl groups
actin binding
chromatin modification
response to drug
0
2
4
Z Score Enrichment
M12_purple Top GO Biological Proces
or Molecular Function
Wnt receptor signaling pathway*
multicellular organismal process*
transmembrane receptor activity*
anatomical structure development*
glycosaminoglycan binding*
cell migration*
tissue morphogenesis*
biological adhesion*
0
2
4
6
Z Score Enrichment
M13_red Top GO Biological Process
or Molecular Function
viral transcription*
translational termination*
viral infectious cycle*
endocrine pancreas development*
translational elongation*
structural constituent of ribosome*
translation*
gene expression*
0
10
25
Z Score Enrichment
M14_salmon Top GO Biological Proce
or Molecular Function
DNA repair
0.0
1.5
3.0
Z Score Enrichment
M15_tan Top GO Biological Process
or Molecular Function
protein homodimerization activity
transition metal ion binding
0.0
1.5
3.0
Z Score Enrichment
M16_turquoise Top GO Biological Proc
or Molecular Function
ensheathment of neurons
sterol metabolic process
phospholipid binding
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides
motor activity
cell-cell junction assembly
cytoskeletal protein binding
chromosome segregation
0
2
4
Z Score Enrichment
M17_yellow Top GO Biological Proce
or Molecular Function
respiratory electron transport chain*
NADH dehydrogenase activity*
hydrogen ion transmembrane transporter activity*
regulation of protein ubiquitination*
regulation of cellular amino acid metabolic process*
anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process*
energy coupled proton transport, down electrochemical gradient*
M/G1 transition of mitotic cell cycle*
0
4
8
Z Score Enrichment

## Title

Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development

## Authors and affiliations

Hyejung Won[1], Luis de la Torre-Ubieta[1], Jason L. Stein[1], Neelroop N. Parikshak[1], Farhad Hormozdiari[3], Changhoon Lee[1], Eleazar Eskin[3,4], Jason Ernst[2,4], Daniel H. Geschwind[1,4*]

[1] Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles

[2] Department of Biological Chemistry, David Geffen School of Medicine, University of California Los Angeles

[3] Department of Computer Science, University of California Los Angeles

[4] Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles

[5]Department of Molecular, Cell and Developmental Biology, University of California Los Angeles, Los Angeles

[*] Correspondence: dhg@mednet.ucla.edu

## Introduction

The demonstration that chromatin exhibits a complex 3 dimensional organization, whereby short and long distance physical interactions correspond to complex gene regulatory processes has opened a new window on understanding the functional organization of the human genome[1-4]. Recently, chromatin remodeling has also been causally implicated in several neurodevelopmental disorders, including autism and schizophrenia[5-7]. However, it remains unclear whether knowledge of chromosome organization in a tissue specific manner might inform our understanding of gene regulation in brain development or disease. Here we determined the genome-wide landscape of chromosome conformation during early human cortical development by performing Hi-C analysis in the mitotically active and post mitotic laminae of human fetal brain. We integrate Hi-C data with transcriptomic and epigenomic data and utilize chromosome contact information to delineate physical gene-gene regulatory interactions for non-coding regulatory elements. We show how these data permit large-scale functional annotation of non-coding variants identified in schizophrenia GWAS and of human specific enhancers[8,9]. These data provide a rubric that illustrates the power of tissue-specific annotation of non-coding regulatory elements, as well as novel insights into the pathogenic mechanisms of neurodevelopmental disorders and the evolution of higher cognition.

Recent advances in high-throughput sequencing have unveiled the epigenomic landscape of multiple human cell types, as well as 3 dimensional folding principles of chromatin[10,11]. In particular, chromosome conformation capture experiments demonstrate that chromatin is organized into hierarchical structures, which include compartments (a few megabase (Mb))[1], topological associating domains (TADs, sub-Mb)[12], and loops (ranging from few kilobase (kb) to few hundred kb)[2,4]. These structures are thought to play a role in gene regulation and biological function by defining functional genomic units and mediating the effects of *cis*-regulatory elements via both short- and long-range physical interactions (e.g. promotor-enhancer interactions), relationships that cannot simply be predicted by linear adjacency in chromosomes. Coupled with epigenomic data, such higher order chromatin interactions should facilitate systemic annotation of *cis*-regulatory elements, as well as intergenic and intronic variants, which will further expand our understanding of tissue specific developmental programs, as well as disease pathogenesis.

We constructed multiple Hi-C libraries in mid gestation fetal cerebral cortex from three individuals during the peak of neurogenesis and migration (gestation week, GW17-18). We reasoned that it would be useful to analyze mitotically active neuronal precursors involved in neurogenesis separately from post-mitotic migrating and maturing neurons, so we dissected the cortical anlage into two major structures: the cortical and subcortical plate (CP), consisting primarily of post mitotic neurons and the germinal zone (GZ), containing primarily mitotically active neural progenitors (representative heatmap in **Fig. 1a**, **Extended Data Fig. 1a-b**). For comparison with non-neuronal cell types, we also used publicly available Hi-C data on human embryonic stem (ES) cells and IMR90 cells[11,12]. To provide grounding for our data and compare global chromosome architecture between different cell types, we performed principal component analysis (PCA)[13] on the genome-wide inter-chromosomal contact matrices of CP, GZ, ES, and IMR90. As previously demonstrated, global chromosome architecture does not change dramatically between different cell types[13]. However, the first principal components (PC1s) from neuronal tissues (CP and GZ) have significantly higher correlation than the PC1s between different cell types (**Fig. 1b**), consistent with the higher similarity between tissues from brain, versus the two other cell lines.

**3D chromatin structure reflects gene regulation during neural differentiation.**

Previous studies have shown that genome-wide chromosome conformation captures multiple levels of genomic features related to biological function, ranging from GC content and gene number to marks of open chromatin, such as DNase I hypersensitivity sites (DHS)[13]. Most human-relevant Hi-C has been conducted in cell lines[1,2,4,11,12,14] and not in complex tissue, such as developing brain. As an initial first step to insure the quality and validity of our data, we analyzed the relationships between the major component of the inter-chromosomal interaction matrix with these major genomic features, finding high correlation with GC content, gene number, DHS[10], and to a lesser extent, gene expression[15] (**Fig. 1c**, **Extended Data Fig. 2a**), as has been previously observed in non-neural cell lines[13].

To further explore the biological significance of chromosome contact changes during neural differentiation, we explored whether the genes in regions of dynamic chromatin structure were related to neural differentiation by comparing the inter-chromosome contact matrices (binned to 100kb) in different cell types and selecting bins with the highest chromatin contact count changes between two cell types (**Methods**). Genes located in the regions of highest inter-chromosomal interaction changes between CP and GZ were enriched for neuronal genes, represented by the gene ontology (GO) categories of neuron recognition, axon guidance, central nervous system (CNS)

development, and synapse (**Fig. 1d**, **Extended Data Fig. 2b**; **Methods**). Genes located in regions with highest inter-chromosomal interaction changes between CP and ES cells were enriched for developmental genes involved in forebrain development and chromatin organization (**Fig. 1d**, **Extended Data Fig. 2b**), indicating that these interactions reflect tissue relevant developmental gene regulation.

To further explore how these physical chromatin interactions relate to biological function, we hypothesized that highly interacting chromatin regions would be more likely to be co-regulated. To test this, we compared the distribution of correlation patterns for genes locating in (1) the regions of highest interaction values in both CP and GZ, (2) the lowest interacting regions in both CP and GZ, and (3) the regions of differential interaction values (the regions of highest interaction values in CP and lowest interaction values in GZ and vice versa). Highly interacting regions tend to be biased toward positive correlations, while there was no bias in correlation for low and differential interacting regions (**Fig. 1e**). Interestingly, the positive correlation for high interacting regions becomes even higher when more stringent cutoffs are used, supportive of the quantitative nature of interaction-driven co-expression, whereby the relationship between physical 3D chromatin interactions and expression is mostly driven by the top percentiles of interacting regions (**Extended Data Fig. 2c**). To further elucidate the epigenetic regulatory mechanisms behind the apparent interaction-mediated co-expression, we marked bins in which epigenetic marks from two loci appear together. By comparing the epigenetic mark combination matrix with the Hi-C contact matrix, we observed that interacting regions exhibit shared epigenetic patterns at the level of both inter- and intra-chromosomal interactions (**Fig. 1f**, **Extended Data Fig. 3**; **Methods**). In particular, regions associated with positive transcriptional regulation and enhancers are more likely to physically interact with each other, consistent with their co-regulation.

One of the core functional units of general genome organization recently uncovered by chromatin capture methods across a wide variety of cell types is the compartment, a relatively large, dynamic domain[1], which is comprised of smaller, sub-Mb regions of topologically associating domains (TADs)[12]. Compartments are divided into two types, type A compartments that consist primarily of euchromatin and actively transcribed genes and type B compartments, which are heterochromatic and repressed. TADs have been previously shown to be relatively stable, whereas compartments have been shown to change during lineage specification in stem cells[11]. Consistent with this, we observed dynamic compartment switching between CP and GZ, enriched for GO categories related to neuronal genes and phosphatase activity (**Fig. 2c**), as well as compartment switching between CP and ES (**Fig. 2a,d**). Genes that change compartments from ES to CP are decreased for A to B transitions across differentiation and increased for changes from the B to A compartments (**Fig. 2b**), as expected. Compartment changes are also accompanied by epigenetic changes, so that the B to A compartment shift is associated with increased DHS and active epigenetic marks indicative of open chromatin, whereas the A to B shift is associated with decreased DHS and increased repressive marks (**Fig. 2b,e**). The same pattern was observed for GZ vs. ES and CP vs. GZ (**Fig. 2b,e**, **Extended Data Fig. 2d**), demonstrating that gene expression changes across development are tightly linked to epigenetic changes coupled with compartment switching.

TADs are thought to mediate co-transcriptional regulation primarily within their boundaries (100kb-1Mb) through physical “looping” interactions of promotors and enhancers in co-regulated genes[4,16]. Since TAD boundaries are conserved across different cell types[12], we hypothesized that changes in epigenetic marks in TADs, rather than the boundaries of TADs, would be most associated with gene expression changes

across development. To test this, we divided genes based on their fold change in expression between ES and differentiated neurons[17] (both increased and decreased), and assessed changes in epigenetic marks within the TADs where these genes reside (**Extended Data Fig. 1c-e**, **Methods**). Notably, active marks including enhancers and elements related to transcribed regions are increased in TADs that contain upregulated genes, whereas repressive marks are increased in TADs that contain downregulated genes (**Fig. 2f**). Collectively, these results indicate that our Hi-C data reflects the major elements of global chromosome architecture in fetal brains, providing a framework for exploring gene regulatory mechanism related to human neural development and function.

Next, to demonstrate how knowledge of intra-chromosomal contacts could significantly advance understanding of important gene regulatory relationships in the nervous system, we performed two integrative experiments. In the first, we used these chromatin contact data to functionally annotate specific non-coding regulatory elements in the developing brain. We leveraged recent efforts that have identified >2000 developmental enhancers gained specifically in the human cerebral cortex, providing a remarkable resource for understanding the evolution of human cognition[8]. Usually, in the absence of such tissue specific data, regulatory elements are assigned to the closest gene[18,19], a convention that we compared with our Hi-C derived interactions. We reasoned that our Hi-C data from fetal brain could be used to identify the target genes for many of these enhancers, which based on previously chromatin looping analyses in cell lines are often not the closest gene[4,16,18,19].

We derived an interaction map of human-gained enhancers, defined as significant interacting regions (at a 1% false discovery rate, FDR) compared to the null distribution generated by fitting the contact frequencies of all fetal brain enhancers identified in the same study[8] (**Extended Data Fig. 4a**, **Methods**). We defined the search space as including the 1Mb flanking regions, since most enhancer-promoter interactions are within this range[4]. Although statistically significant interactions are increased upon proximity to the enhancer, the majority of interactions are at relatively long-ranges (>100kb, **Extended Data Fig. 4b**) and are not restricted to the adjacent genes. Indeed, ~65% of the closest genes to human-gained enhancers are not identified through fetal brain Hi-C interactions, revealing that the majority of enhancers are not interacting with the most adjacent gene (**Fig. 3c**). Compared to the original study[8], which used human-gained enhancer hotspot TADs in ES cells and IMR90 cells due to the lack of Hi-C data from relevant tissue, our approach provides genes of action with higher resolution in the matching tissue (fetal cortices) from which evolutionary enhancers were identified. Human-gained enhancer-interacting regions were enriched with enhancers, promoters, and transcription start sites (TSSs) (**Fig. 3a**, **Extended Data Fig. 4c**), consistent with the previous findings that enhancers interact with promoters, as well as other enhancers[16]. The majority of interactions (>75%) were in the same TADs (**Fig. 3b**), also consistent with observations in cell lines that most enhancer-promoter interactions are in the same TAD[16,19]. Human-gained enhancer interacting genes (Hi-$C_{evol}$ genes) are involved in GTPase regulation as well as G-protein coupled receptor (GPCR) and CREB signaling, and are enriched with GO terms representing synaptic and axon guidance genes (**Fig. 3e**, representative interactions in **Fig. 3d**). One striking example is a human-gained enhancer that interacts with *ARHGAP11B*, a human-specific gene implicated in the expansion of human neocortex[20] (**Fig. 3d**).

Given the high conservation of protein-coding genes across the vertebrate lineage, comparative genomics have suggested that human-specific traits most likely result from changes in regulatory elements[8,21]. Indeed, protein-coding Hi-$C_{evol}$ genes have a lower

non-synonymous substitution (dN)/synonymous substitution (dS) ratio compared to Hi-C non-interacting protein-coding genes in multiple lineages (**Extended Data Fig. 5**). These results indicate that human-gained enhancers are interacting with protein-coding genes that undergo purifying selection, further supporting the hypothesis that non-coding elements undergo evolutionary selection to induce species-specific changes in gene expression[8,21]. We also investigated whether human-gained enhancers are interacting with lineage-specific long non-coding RNAs (lncRNAs)[22]. We observed that lineage-specific interactions with human-gained enhancers were enriched for primate-specific lncRNAs, as well as evolutionary conserved lncRNAs (**Fig. 3f**, **Extended Data Fig. 5**). Thus, while human-gained enhancers interact and possibly regulate evolutionary conserved protein-coding genes, they are more likely to interact with primate-specific lncRNAs.

Since the development of human higher cognition is dependent on the development of the human cerebral cortex via elaboration of novel gene regulatory relationships[8,23], we reasoned, as have others[8] that the genes regulated by these human specific enhancers would be associated with intellectual functioning in humans. Remarkably, we found that the Hi-$C_{evol}$ genes in fetal brain, but not the genes defined by proximity to the enhancers are significantly enriched with intellectual disability (ID) risk genes[6] (**Fig. 3g**). This result provides experimental support for the contention that human-gained enhancers are associated with the evolution of human cognitive function[8]. This enrichment was tissue-specific, as Hi-$C_{evol}$ genes defined by Hi-C interactions in ES cells did not show enrichment for ID risk genes (**Fig. 3g**). Indeed, ~56% of the Hi-$C_{evol}$ genes in neuronal tissue were not identified through chromatin contacts in ES cells, emphasizing the importance of defining tissue-relevant chromatin contacts, as well as importance of using the relevant tissue for Hi-C analysis (**Fig. 4c**).

Since most disease related common genetic variation is located in non-protein coding regions, we next assessed the ability of Hi-C data for functional annotation of common single nucleotide polymorphisms (SNPs). As a first line verification that Hi-C data could identify known functional relationships between SNPs and gene expression we used *cis*-expression quantitative trait loci (eQTL) data from adult frontal cortex[24], since such data is not yet available from fetal brain. For each significant eQTL locus, we obtained a set of significant eQTL SNPs with >95% likelihood of containing the causal SNP from association statistics and linkage disequilibrium (LD; 1000 Genomes) structure using CAVIAR[25]. We then identified genes interacting to likely causal eQTL SNPs via the chromatin contact matrix (Hi-$C_{eQTL}$ genes, **Methods**), and compared Hi-$C_{eQTL}$ genes with the known associated gene from the eQTL study, finding that Hi-$C_{eQTL}$ genes significantly overlapped with eQTL transcripts (**Extended Data Fig. 6a**). There were many Hi-$C_{eQTL}$ genes that were not identified as eQTL transcripts, likely due to a combination of factors, including low power of the eQTL sample, limited resolution of Hi-C (SNP-transcript interactions within 20kb cannot be detected), and the difference in age of tissues used for each analysis. Indeed, eQTL SNPs identified by CAVIAR were highly enriched with adult frontal cortex, but not fetal brain, enhancers (**Extended Data Fig. 6b-d**). Despite this, eQTL SNP-transcript pairs exhibit higher chromatin contact frequency than expected by chance across all distance ranges (**Extended Data Fig. 6e**), further supporting the utility of Hi-C to infer the biological function of regulatory variation.

Next, we applied a similar logic to advance our understanding of 108 genome-wide significant schizophrenia-associated loci, most of which are in relatively uncharacterized non-coding regions of the genome[9]. We obtained credible SNPs using CAVIAR, and split SNPs into those without known function and likely functional SNPs (SNPs that cause missense, frameshift, and splice variants and SNPs that fall onto gene promoters;

Methods). Credible SNPs were enriched with enhancers in fetal brain and adult frontal cortex, confirming the likely regulatory role of these SNPs in the brain (**Extended Data Fig. 7**). SNPs defined as likely functional SNPs and promoter SNPs were directly assigned to their target genes. For the remaining intergenic and intronic SNPs that were un-annotated, and therefore without clear function, we used the chromatin contact matrix to find genes with which the regions where the SNPs are located are physically interacting (diagram in **Extended Data Fig. 7**).

Combining genes annotated as functional SNPs, promoter SNPs, and by Hi-C interactions, we obtained a total of ~900 genes (Hi-$C_{SCZ}$ genes) associated with schizophrenia risk variants. Hi-C contacts identified numerous genes that were neither adjacent to index SNPs nor in LD with them (**Fig. 4a-c**, **Extended Data Fig. 9**). While almost 70-80% of the LD genes and closest genes were identified as Hi-$C_{SCZ}$ genes, only half of them were identified by chromatin contacts, indicating that many of them were identified by functional SNPs residing in the genes. Moreover, 70-90% of the Hi-$C_{SCZ}$ genes were not identified by using LD genes or the closest genes to the association signal, consistent with observations that the linear organization of genes and regulatory elements on the chromosome does not reflect regulatory interactions[4,18,19].

Hi-C analysis showed that schizophrenia-associated common variants converge into specific molecular pathways related to neuronal function, including the postsynaptic density, acetylcholine receptors, cell cycle, and chromatin remodelers (**Fig. 4d-e**, **Extended Data Fig. 7-8**). To insure that this was not an artifact of the method used for credible SNP selection, we used a different method to define the set of credible SNPs[9] (**Extended Data Fig. 9**) and found the same enrichments, demonstrating the robustness of the genes identified through the Hi-C analysis. One notable example is illustrated by credible SNPs (rs4245150, rs17602038, rs4938021, rs4936275, rs4936276) that reside upstream of the *Dopamine D2 Receptor* (*DRD2*), the target of antipsychotic drugs. Although these SNPs are close to the *DRD2* TSS, they are not within the gene, which complicates interpretation of their biological function. Hi-C analysis demonstrates for the first time that indeed these SNPs are interacting with the TSS of *DRD2* (**Fig 4e**), providing biological insights into the function of these SNPs.

Another relevant example is an index SNP (rs79212538) interacting with *GRIA1*, an ionotropic glutamate receptor subunit, although *GRIA1* is neither the closest gene nor in LD with the index SNP (**Extended Data Fig. 8**). Additionally, Hi-C shows that schizophrenia associated non-coding SNPs interact with multiple genes involved in excitatory synaptic transmission, including *CACNA1C*, *GRIN2A*, and *NLGN4X*, further supporting glutamatergic transmission defects in schizophrenia pathophysiology (**Extended Data Fig. 8**). Interestingly, Hi-$C_{SCZ}$ genes significantly overlap with ASD *de novo* likely gene-disrupting (LGD) targets (CP: OR=2.4, P=$1.6\text{x}10^{-5}$, GZ: OR=1.8, P=0.006), consistent with a shared genetic etiology between ASD and schizophrenia[26]. The fact that genes with LGD mutations in ASD are associated with regulatory variants in schizophrenia suggests that complete abrogation of these genes may cause developmental defects as in ASD, while regulatory changes in these genes may cause later-onset of neuropsychiatric symptoms as in schizophrenia. Collectively, genes annotated by chromatin contact information provide novel insights into schizophrenia pathogenesis.

In conclusion, we demonstrate how a comprehensive analysis of genome-wide chromatin configuration during human neural development informs our view of gene regulation. This chromatin contact landscape provides important biological insights on gene regulatory mechanisms, such that co-expressed genes share epigenetic co-regulation of interacting regions, and that changes in functional epigenetic marks are tightly linked to TADs and compartment switching to induce changes in gene expression.

We also annotated non-coding regulatory elements in the genome based on long-range chromatin contacts to identify enhancer-promoter interactions during human brain development, as well as genes of actions for eQTL. In turn, we show how these interactions can be used to inform our biological interpretation of risk variants for schizophrenia, which serves as a template for understanding the role of non-coding variation more broadly in neuropsychiatric disorders.

## Methods

### Fetal brain layer dissection

Human fetal cortical tissues from three individuals were collected from frontoparietal cortex at gestation week (GW) 17-18 (one sample from GW17 and two samples from GW18). In cold DMEM/F-12 (ThermoFisher, 11320-033), frontoparietal cortex was first dissected to thin (~1mm) slices to visualize layers. Under the light field microscope, cortical slice was dissected to germinal zone (GZ) and cortical plates (CP). GZ contains ventricular zone and subventricular zone, and hence comprised of proliferating neurons. CP refers to intermediate zone, cortical plate, and marginal zone, which are mainly composed of differentiated and migrating neurons. By dissecting layers from same fetal cortices, we can compare progenitors to differentiated neurons with same genotype and minimize intersample heterogeneity.

### Hi-C

Collected tissue was dissociated with trypsin and cell number was counted. Ten million cells were fixed in 1% formaldehyde for 10 min. Cross-linked DNA was digested by restriction enzyme HindIII (NEB, R0104). Digested chromatin ends were filled and marked with biotin-14-dCTP (ThermoFisher, 19518-018). Resulting blunt-end fragments were ligated under dilute concentration to minimize random intermolecular ligations. DNA purified after crosslinking was reversed by proteinase K (NEB, P8107) treatment. Biotins from unligated ends were removed by exonuclease activity of T4 DNA polymerase (ThermoFisher, 18005). DNA was sheared by sonication (Covaris, M220) and 300-600bp fragments were selected. Biotin-tagged DNA, which is intermolecular ligation products, was pulled down with streptavidin beads (Invitrogen, 65001), and ligated with Illumina paired end adapters. Resulting Hi-C library was amplified by PCR (KAPA Biosystems HiFi HotStart PCR kit, KK2502) with the minimum number of cycle (typically 12-13 cycles), and sequenced by Illumina 50bp paired-end sequencing.

### Hi-C reads mapping and pre-processing

Note that mapping and filtering of the reads, as well as normalization of experimental and intrinsic biases of Hi-C contact matrices were conducted with the following method regardless of cell types to minimize potential variance in the data obtained from different platforms. We implemented *hiclib* (https://bitbucket.org/mirnylab/hiclib) to perform initial analysis on Hi-C data from mapping to filtering and bias correction. Briefly, quality analysis was performed using a phred score, and sequenced reads were mapped to hg19 human genome by *Bowtie2* (with increased stringency, *--score-min -L 0.6,0.2--very-sensitive*) through iterative mapping. Read pairs were then allocated to HindIII restriction enzyme fragments. Self-ligated and unligated fragments, fragments from repeated regions of the genome, PCR artifacts, and genome assembly errors were removed. Filtered reads were binned at 10kb, 40kb, and 100kb resolution to build a genome-wide contact matrix at a given bin size. This contact map depicts contact frequency between any two genomic loci. Biases can be introduced to contact matrices by experimental procedures and intrinsic properties of the genome. To decompose biases from the contact matrix and yield a true contact probability map, filtered bins were subjected to iterative correction[13], the basic assumption of which is that each locus has uniform coverage. Bias correction and normalization results in a corrected heatmap of bin-level resolution. 100kb resolution bins were assessed for inter-chromosomal interactions, 40kb for TAD analysis, and 10kb for gene loop detection.

### Inter-chromosomal principal component analysis

Principal component analysis (PCA) was conducted in a genome-wide inter-chromosome contact map (100kb binned) as described previously[13]. Since intra-

chromosome conformation may drive the PCA results, *cis* contacts were iteratively replaced to random *trans* counts. After removing diagonal and poorly covered regions, we performed PCA using *hiclib* command *doEig*.

Pearson's correlations between the first principal components (PC1) from different cell types (CP, GZ, ES, and IMR90[12]) were calculated to compare similarities in inter-chromosomal interactions between different cell types.

Spearman's correlations between PC1/PC2 and biological traits (GC content, gene density, DNase I hypersensitivity (DHS), gene expression) were calculated. GC content (%) for each 100kb bin was calculated by *gcContentCalc* command from R package *Repitools*. Gene density (number of genes in 100kb bin) was obtained based on longest isoforms from GENCODE19. DHS of fetal brains from Epigenomic roadmap[10] and gene expression level of prenatal cortical layers from Miller et al.[15] were used and average values per 100kb bin were calculated.

### Gene enrichment analysis

Gene ontology (GO) enrichment was performed by GO-Elite Pathway Analysis (http://www.genmapp.org/go_elite/). All genes in the genome except the ones located in the chromosome Y and mitochondrial DNA were used as a background gene list. Because Hi-C interaction is measured in bins, sometimes we cannot dissect the individual genes when they are clustered in the genome (i.e. PCDH locus). To prevent several gene clusters overriding entire GO terms, we removed GO mainly defined by gene clusters (for 100kb or 40kb binned data) or we randomly included one gene per cluster (e.g. PCDHA1 for PCDHA1-13 cluster) prior to GO analysis (for 10kb binned data).

Gene enrichment for the curated gene lists was performed using binomial generalized linear model to regress out exome length. Autism spectrum disorder (ASD) *de novo* gene list and intellectual disability (ID) curated gene list from Iossifov et al.[27] and Parisksh ak et al.[6] were used for the enrichment test, respectively. Protein-coding genes based on biomaRt were used as a background gene list.

### Identification of the regions with largest inter-chromosomal conformation changes

Chromosome contact matrix was normalized with the total interaction counts between two cell types for comparison. Intra-chromosomal interactions were masked from the genome-wide contact matrix, and top 1000 bins with the largest interaction changes between different cell types (GZ vs. CP or ES vs. CP) were selected. As one bin is comprised of two loci that are interacting with each other, this would give ~2000 sites in the genome. Genes located in those ~2000 sites were combined to perform GO analysis.

### Co-expression of inter-chromosomal interacting regions

Using transcriptome from fetal cortical layers[28], average expression values per 100kb bin were calculated. Pearson correlation matrix was calculated from 100kb binned expression data from all layers to generate gene co-expression matrix. At this step, gene co-expression matrix has the same dimension as inter-chromosomal contact matrix.

We hypothesized that genes would be co-expressed across the layers when they are interacting in all stages (both in CP and GZ), so we selected top 2% highest interacting regions of fetal brains both at GZ and CP (high interacting regions). We also selected (1) low interacting regions: top lowest interacting regions (0 interaction from normalized Hi-C contact matrix) of fetal brains both at GZ and CP, as well as (2) variant interacting regions: top 2% highest interacting regions from one stage (e.g. GZ) that are top 2%

lowest interacting regions from the other stage (e.g. CP) for comparison. Expression correlation values of the same regions were selected from the gene co-expression matrix, and expression correlations between different states (high interacting regions vs. low interacting regions and high interacting regions vs. variant interacting regions) were compared by two-sample Kolmogorov-Smirnov test.

### Epigenetic state enrichment for inter-chromosomal interacting regions

The fetal brain epigenetic 25 state model from Epigenomic roadmap[10] was used to generate the epigenetic state combination matrix, which was generated by marking loci where two interacting chromosomal bins (defined as bins with (1) interaction counts > 75% quantile interaction count for inter-chromosome and (2) interaction counts > 0 for intra-chromosome) share epigenetic signature. For example, the epigenetic combination matrix between the active transcription start site (TssA) and active enhancers (EnhA1) was generated by marking where interacting loci have TssA on one locus and EnhA1 on the other locus. Intra- and inter-chromosomal contact frequency maps were then compared to epigenetic state matrix by Fisher's exact test to calculate enrichment of shared epigenetic combinations in interacting regions.

### Compartment analysis

Expected interaction frequency was calculated from the normalized intra-chromosomal 40kb binned contact matrix based on the distance between two bins. We summed series of submatrices of 400kb window size with 40kb step size from the normalized Hi-C maps to generate observed and expected matrices. The Pearson's correlation matrix was computed from the observed/expected matrix, and PCA was conducted on correlation matrix. PC1 from each chromosome was used to identify compartments. Eigenvalues positively correlated with the gene density were set as compartment A, while those that are negatively correlated were set as compartment B.

### Gene expression and epigenetic state change across different compartments

Genomic regions were classified into three categories according to compartments: compartment A in cell type1 that changes to compartment B in cell type2 (A to B), compartment B in cell type1 that changes to compartment B in cell type2 (B to A), regions that do not change compartment between two cell types (stable).

Genes residing in each compartment category were selected and GO enrichment was performed. Gene expression fold-change (FC) between different cell types was calculated from Miller et al.[15] (comparison for CP vs. GZ) and CORTECON[17] (comparison for ES vs. CP and ES vs. GZ). Distribution of gene expression FC for genes in different compartment categories was compared by one-way ANOVA and Tukey's posthoc test.

15 state epigenetic marks from Epigenomic Roadmap[10] in genomic regions classified based on compartments were averaged across 40kb bins. The DHS FC[10] between different cell types (ES vs. CP and ES vs. GZ) was calculated and statistically evaluated as in the gene expression comparison. Each epigenetic state counts[10] for one compartment category was normalized by total epigenetic mark number of that compartment category and compared between ES and fetal brains.

### TAD analysis

We conducted TAD-level analysis as described previously[12]. Shortly, we quantified the directionality index by calculating the degree of upstream or downstream (2Mb) interaction bias of a given bin, which was processed by a hidden Markov model (HMM) to remove hidden directionality bias.

Regions in between TADs are titled as TAD boundaries when the regions are smaller than 400kb and unorganized chromatin when the regions are larger than 400kb.

### TAD-based epigenetic changes upon differentially expressed genes

Genes were subdivided into 20 groups based on expression FC between ES and most differentiated neuronal states in CORTECON[17]: genes that are upregulated and downregulated upon differentiation were grouped into 10 quantiles, respectively, based on the FC. TADs into which genes from one subdivision reside were selected, and epigenetic state changes (from Epigenomic roadmap's 15 state epigenetic marks in ES and fetal brains[10]) in those TADs were normalized with TAD length and compared between ES and fetal brains. As different types of epigenetic marks have different absolute numbers (e.g. there are more quiescent states than enhancer states in the genome), each epigenetic state change was scaled across different quantiles to allow comparison between different states.

### Identification of Hi-C interacting regions

We identified Hi-C interacting regions and target genes for (1) human-gained enhancers[8], (2) expression quantitative trait loci (eQTL) SNPs[24], and (3) schizophrenia SNPs[9]. As the highest resolution available for the current Hi-C data was 10kb, we assigned these enhancers/SNPs to 10kb bins, obtained Hi-C interaction profile for 1Mb flanking region (1Mb upstream to 1Mb downstream) of each bin. We also made a background Hi-C interaction profile by pooling (1) 255,698 H3K27ac sites from frontal and occipital cortex at 12 PCW for human-gained enhancers[8] and (2) 9,444,230 imputed SNPs for eQTL and schizophrenia SNPs[9]. To avoid significant Hi-C interactions affecting the distribution fitting as well as parameter estimation, we used the lowest 95 percentiles of Hi-C contacts and removed zero contact values. Using these background Hi-C interaction profiles, we fit the distribution of Hi-C contacts at each distance for each chromosome using *fitdistrplus* package (**Extended Data Fig. 4a**). Significance for a given Hi-C contact was calculated as the probability of observing a stronger contact under the fitted Weibull distribution matched by chromosome and distance. P-values were adjusted by computing FDR, and Hi-C contacts with FDR<0.01 were selected as significant interactions. Significant Hi-C interacting regions were overlapped with GENCODE19 gene coordinates (including 2kb upstream to transcription start sites (TSS) to allow detection of enhancer-promoter interactions) to identify interacting genes. Same analysis was performed on Hi-C contact maps from CP, GZ, and ES[11]. To address the functional significance of target genes, GO enrichment was performed for the interacting genes.

### Protein-coding genes interacting with human-specific evolutionary enhancers

Protein-coding genes based on biomaRt (GENCODE19) were selected and non-synonymous substitution (dN)/synonymous substitution (dS) ratio was calculated for homologs in mouse, rhesus macaque, and chimpanzee for representation of mammals, primates, and great apes, respectively. Log2(dN/dS) distributions for protein-coding genes interacting vs. non-interacting to human-specific evolutionary enhancers in each lineage were then compared by two-sample Kolmogorov-Smirnov test.

### LncRNAs interacting with human-specific evolutionary enhancers

Long non-coding RNAs (lncRNAs) classified according to evolutionary lineages[22] were used to assess whether lineage-specific lncRNAs are interacting to human-specific evolutionary enhancers. We randomly selected the same number of enhancers (2,104) to the human-specific ones from the total enhancer pool (255,698), identified interacting regions based on the null distribution generated from a background enhancer interaction profile. Significant interacting regions (FDR<0.01) identified by Hi-C were intersected

with lncRNA coordinates[22] and interacting lncRNAs for each lineage were counted. This step was repeated for 3,000 times to obtain the lncRNA lineage distribution. LncRNAs interacting with human-specific evolutionary enhancers were also identified and enrichment was tested by calculating P-values as the probability of observing more interacting lncRNAs for a given lineage under the null lncRNA lineage distribution.

### Epigenetic state enrichment for Hi-C interacting regions

The functional framework for (1) eQTL SNPs, (2) schizophrenia SNPs, and (3) human-gained enhancers-interacting regions was assessed for epigenetic state enrichment. We implemented the same approach as in GREAT[29] to analyze the epigenetic state enrichment for *cis*-regulatory regions. For example, to evaluate whether schizophrenia SNPs are enriched with DHS, fraction of genome annotated with DHS (p), the number of schizophrenia SNPs (n), and number of schizophrenia SNPs overlapping with DHS (s) were calculated. Significance of the overlaps was tested by binomial probability of $P = Pr_{binom}\ (k \geq s \mid n = n,\ p = p)$[29]. Histone marks and 15-chromatin states from fetal brains, adult frontal cortex, and IMR90[10] were used for epigenetic state enrichment.

### eQTL analysis

To address whether co-localization mediates gene regulation, we compared the association between chromosome conformation with eQTL. Although fetal brain eQTL data would be optimal, since this data is currently not available, we analyzed adult frontal cortex *cis*-acting eQTL data[24]. We selected SNPs associated with gene expression (FDR<0.01) and clustered them with association $P<1\times10^{-5}$ and $r^2>0.6$ to obtain index SNPs. Using summary association statistics and linkage disequilibrium (LD) structure for each index SNP, we applied *CAVIAR*[25] to quantify the probability of each variant to be causal. Among 121,273,364 SNP-transcript pairs from frontal cortex eQTL data, this process resulted in 42,190 SNP-transcript pairs (267 transcripts and 14,882 SNPs) that are potentially credible. We refer to 14,882 credible SNPs as credible SNPs. Credible SNP interacting genes were identified as described in “identification of Hi-C interacting regions” section.

Fisher’s exact test was performed to evaluate the significance of the overlap between Hi-C interacting genes and eQTL transcripts. The background gene list for Fisher’s exact test includes genes located in 1Mb flanking regions to credible SNPs that are also tested in eQTL analysis.

For 42,190 SNP-transcript pairs, we assigned credible SNPs and genes into 10kb bins, and obtained Hi-C contacts between credible SNPs and genes from the 10kb binned Hi-C contact maps. As a gene can span across multiple 10kb bins, the highest interaction in the gene to a credible SNP was selected as Hi-C contacts as previously defined[30]. We also calculated expected interaction frequency from the normalized 10kb binned contact matrix based on the distance between two bins. Opposite interaction frequency was calculated by obtaining Hi-C contacts for the opposite site to the credible SNP with the same distance. Because interaction counts differ in different chromosomes as well as in different cell types, we normalized interaction by chromosomes and cell types. We performed one-way ANOVA and Tukey’s posthoc test for the comparison between different interaction paradigms.

### Identification of credible SNPs for schizophrenia GWAS loci

128 LD-independent SNPs with genome-wide significance ($P<5\times10^{-8}$)[9] were used as index SNPs to obtain schizophrenia credible SNPs. All SNPs that are associated with $P<1\times10^{-5}$ and in LD ($r^2>0.6$) with an index SNP were selected, and correlations among this set of SNPs (LD structure) were calculated. CAVIAR was applied to summary association statistics and LD structure for each index SNP, and potentially causal SNPs

for each index SNP were identified. Among 55,000 SNPs that are in LD with 128 index SNPs, 7,613 SNPs were selected as causal by CAVIAR. Here we refer to these CAVIAR-identified SNPs as credible SNPs. Genes interacting to credible SNPs were identified as described in "identification of Hi-C interacting regions" section for CP, GZ, and ES. A separate set of credible SNPs initially reported from the original study was also processed with the same method[9].

### Identification of schizophrenia GWAS SNP-associated genes

We classified credible SNPs based on potential functionality (flow chart in **Extended Data Fig. 7**). For credible SNPs classified as functional (stop gained variant, frameshift variant, splice donor variant, NMD transcript variant, and missense variant) from biomaRt, we selected genes in which those SNPs locate. For those that are not directly affecting the gene function, we selected SNPs that fall onto the promoter and TSS of genes (2kb upstream-1kb downstream to TSS). Remaining SNPs were tested for Hi-C interaction so that Hi-C interacting genes were identified. This pipeline gives total ~900 genes potentially associated with GWAS SNPs.

### Identification of closest genes and LD genes

Closest genes to human-gained enhancers and schizophrenia index SNPs were obtained by *closestBed* command from *bedtools*. Gene coordinates from GENCODE19 including 2kb upstream to TSS were used to identify the closest genes.

LD genes refer to all genes in the LD. Here, LD is defined as physically distinct schizophrenia-associated 108 genome-wide significant regions[9]. We overlapped gene coordinates from GENCODE19 with LD regions to find genes that reside in LD.

Closest genes and LD genes were compared with Hi-C interacting genes. Venn diagrams were generated by *Vennerable* package in R. Only protein-coding genes were included in plotting Venn diagrams.

### Calculation of distance between SNPs and genes

For LD genes and closest genes, the shortest distance between an index SNP and a target gene was selected. For credible SNPs, (1) the distance between functional credible SNPs and target genes was set as 0, because functional SNPs reside in the gene, (2) the distance between promoter credible SNPs and target genes was calculated as the distance between SNPs and TSS of a gene, (3) the distance between credible SNPs and Hi-C interacting genes was calculated based on the distance between SNPs and Hi-C interacting bins (note that this distance has a unit of 10kb). We then combined the distance distributions from the 3 categories.

## Figure Legends

**Figure 1. Chromosome conformation in fetal brains reflects genomic features. a.** Representative heatmap of the chromosome contact matrix of CP. Normalized contact frequency (contact enrichment) is color-coded according to the legend on the right. **b.** Pearson correlation of the leading principle component (PC1) of inter-chromosomal contacts at 100kb resolution between *in vivo* cortical layers and non-neuronal cell types (ES and IMR90). **c.** Spearman correlation of PC1 of chromatin interaction profile of fetal brain (GZ) with GC content (GC), gene number, DNase I hypersensitivity (DHS) of fetal brain, and gene expression level in fetal laminae. **d.** GO enrichment of genes located in the top 1000 highly interacting inter-chromosomal regions specific to CP vs. GZ (left), and CP vs. ES (right), indicating that genes located on dynamic chromosomal regions are enriched for neuronal development. **e.** The top 2% highest interacting regions of fetal brains both at GZ and CP (High) show positive correlation in gene expression, while the top 2% lowest interacting regions (Low) and top 2% highly variant regions (Variant) have no skew in distribution. P-values from Kolmogorov–Smirnov test. **f.** The epigenetic state combination in inter-chromosomal interacting regions in GZ. Inter-chromosomal contact frequency map is compared to epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1), transcriptional regulators (TxReg), and transcribed regions (Tx) interact highly to each other as marked in red. Colored bars on the left represent epigenetic marks associated with promoters and transcribed regions (orange), enhancers (red), and repressive marks (blue). Chr, chromosome. Annotation for epigenetic marks described in

http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Figure 2. Compartment and TADs provide insights into gene regulatory mechanism. a.** Leading principal component (PC1) of the intra-chromosomal contact matrix in CP, GZ, and ES, with the DNase I hypersensitivity (DHS) fold change (FC) between ES and fetal brain (FB). PC1 values indicate compartment status of a given region, where positive PC1 represents compartment A (red), and negative PC1 represents compartment B (green). **b.** Distribution of gene expression FC (left) and DHS FC (right) for genes/regions that change compartment status ("A to B" or "B to A") or that remain the same ("stable") in different cell types. P-values from one-way ANOVA. **c.** GO enrichment of genes that change compartment status from A to B (top) and B to A (bottom) in CP to GZ. **d.** Heatmap of PC1 values of the genome that change compartment status in different cell types. **e.** Percentage of epigenetic marks for genomic regions that change compartment status between ES and CP. Note that B to A shift in ES to CP is associated with increased proportion of active transcribed regions (TssA and Tx) and enhancers (Enh, top), while A to B shift in ES to CP is associated with increased proportions of repressive marks (Het and ReprPCWk, bottom). P-values from Fisher's exact test. **f.** Epigenetic changes in topological associating domains (TADs) mediate gene expression changes during neuronal differentiation. Genes were divided by expression FC between ES and differentiated neurons, and epigenetic marks in the TADs containing genes in each group were counted and compared between ES and CP. Upregulated genes in neurons locate in TADs with more active epigenetic marks in CP than in ES, while downregulated genes in neurons locate in TADs with more repressive marks in CP than in ES. Epigenetic states associated with activation and transcription of the genes were marked as a red bar, while those associated with repression were marked as blue bars on the right. Annotation for epigenetic marks

described in http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Figure 3. Genetic architecture of human-gained enhancers. a.** Fraction of epigenetic states for regions interacting to human-gained enhancers in CP and GZ. **b.** Proportions of whether human-gained enhancers and interacting regions are within the same topological associating domain (TAD) vs. outside of the TAD. **c.** Overlap between human-gained enhancer interacting genes (Hi-$C_{evol}$ genes) in CP and GZ with closest genes to human-gained enhancers (left) and Hi-$C_{evol}$ genes in ES (right). **d.** Representative interaction map of a 10kb bin, in which human-gained enhancers reside, with the corresponding 1Mb flanking regions. This interactome map provides genes of action that interact with human-gained enhancers. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; Evol, genomic coordinate for a 10kb bin in which human-gained enhancers reside; -log10(P-value), P-value for the significance of the interaction between human-gained enhancers and each 10kb bin, grey dotted line for FDR=0.01; TAD, TAD borders in CP, GZ, and ES. **e.** GO enrichment for Hi-$C_{evol}$ genes in CP (left) and GZ (right). **f.** Number of primate-specific long non-coding RNAs (lncRNAs) interacting with human-gained enhancers in CP (red vertical lines in the graph) against a background control generated from 3,000 permutations, where the number of lncRNAs interacting with the same number of enhancers pooled from all fetal brain enhancers was counted. **g.** Overrepresentation of Hi-$C_{evol}$ genes in different tissues and closest genes with a curated set of intellectual disability (ID) risk genes. *P<0.05, **P<0.01, *** P<0.001. TSS, transcription start site; OR, odds ratio; GPCR, G-protein coupled receptor; Hi-C genes: GZ, CP, ES, Hi-$C_{evol}$ genes in each tissue; Hi-C genes: FB, union of Hi-$C_{evol}$ genes in GZ and CP; Hi-C genes: ES-specific, Hi-$C_{evol}$ genes in ES but not in fetal brain (FB); Hi-C genes: FB-specific, Hi-$C_{evol}$ genes in FB (union) but not in ES; Closest genes, closest genes to human-gained enhancers.

**Figure 4. Annotation of significant chromatin interactions for schizophrenia-associated loci. a.** Overlap between closest genes to index SNPs (Closest), genes locating in linkage disequilibrium (LD), and genes identified through SNP categorization and chromatin contacts in CP and GZ (Hi-$C_{SCZ}$ genes, diagram in **Extended Data Fig. 7**). **b.** Number of closest genes and LD genes that interact to credible SNPs (Hi-C supported) and those that do not interact to credible SNPs (Hi-C non-supported, top). Number of genes that interact to credible SNPs that are closest to or in LD with index SNPs (Hi-C genes), and not closest to or in LD with index SNPs (Hi-C genes not, bottom). Note that Hi-C genes here contain physically interacting genes, but not genes identified by functional SNPs or promoter SNPs. **c.** Distance between CAVIAR/index SNPs and their target genes for closest genes to index SNPs (Closest), genes locating in linkage disequilibrium (LD), and Hi-$C_{SCZ}$ genes in CP (CP) and GZ (GZ) **d.** GO enrichment for Hi-$C_{SCZ}$ genes in CP (left) and GZ (right). **e.** Representative interaction map of a 10kb bin, in which credible SNPs reside, to the corresponding 1Mb flanking regions. This interactome provides target genes interacting to credible SNPs-containing region. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; SNP, genomic coordinate for a 10kb bin in which credible SNPs locate; -log10(P-value), P-value for the significance of the interaction between credible SNPs and each 10kb bin, grey dotted line for FDR=0.01; GWAS loci, LD region for the index SNP; TAD, topological associating domain borders in CP, GZ, and ES.

## Acknowledgements

This work was supported by NIH grants to D.H.G. (5R01MH060233:05; 5R01MH100027) and J.L.S (K99MH102357), Glenn/AFAR Postdoctoral Fellowship Program for Translational Research on Aging (20145357) and Basic Science Research Program through the National Research Foundation of Korea (NRF) (2013024227) to H.W., the California Institute for Regenerative Medicine (CIRM)-BSCRC Training Grant to L.T.U. (TG2-01169). The Hi-C library was sequenced by Broad Stem Cell Research Center (BSCRC), and fetal tissue was collected from Center for Aids Research (CFAR, 5P30 AI028697). eQTL data was kindly provided by Mina Ryten and Adaiklavan Ramasamy. We thank Suhua Feng, Michael Gandal, Yuan Tian, Vivek Swarup, Grant Belgard for helpful discussions and critical reading of the manuscript.

## Author Contributions

H.W. designed and performed experiments, interpreted results, and co-wrote the manuscript. L.T.U. performed sample collection and experiments. J.L.S., N.N.P., and F.H. analyzed data. C.L. helped establishing Hi-C protocol. J.E. and E.E. participated in the discussion of the results. D.H.G. supervised the experimental design and analysis, interpreted results, provided funding, and co-wrote the manuscript.

## Author Information

*Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles*

Hyejung Won, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Changhoon Lee, Daniel H. Geschwind

*Department of Biological Chemistry, University of California California Los Angeles*

Jason Ernst

*Department of Computer Science, University of California Los Angeles*

Farhad Hormozdiari, Eleazar Eskin

*Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles*

Daniel H. Geschwind, Eleazar Eskin, Jason Ernst

## References

1 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

2 Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

3 Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).

4 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294, doi:10.1038/nature12644 (2013).

5 Network & Pathway Analysis Subgroup of Psychiatric Genomics, C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209, doi:10.1038/nn.3922 (2015).

6 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021, doi:10.1016/j.cell.2013.10.031 (2013).

7 Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007, doi:10.1016/j.cell.2013.10.020 (2013).

8 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159, doi:10.1126/science.1260943 (2015).

9 Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

10 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

11 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

12 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

13 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).

14 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64, doi:10.1038/nature12593 (2013).

15 Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206, doi:10.1038/nature13185 (2014).

16 Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-1065, doi:10.1016/j.cell.2015.07.048 (2015).

17 van de Leemput, J. *et al.* CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. *Neuron* **83**, 51-68, doi:10.1016/j.neuron.2014.05.013 (2014).

18 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

19 Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).

20 Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465-1470, doi:10.1126/science.aaa1975 (2015).

21 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).

22 Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635-640, doi:10.1038/nature12943 (2014).

23 Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647, doi:10.1016/j.neuron.2013.10.045 (2013).

24 Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428, doi:10.1038/nn.3801 (2014).

25 Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508, doi:10.1534/genetics.114.167908 (2014).

26 McCarthy, S. E. *et al.* De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658, doi:10.1038/mp.2014.29 (2014).

27 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221, doi:10.1038/nature13908 (2014).

28 Miller, J. A., Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12698-12703, doi:10.1073/pnas.0914257107 (2010).

29 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

30 Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic acids research* **42**, 87-96, doi:10.1093/nar/gkt857 (2014).

**Extended Data Figure 1. Basic characterization of Hi-C libary. a.** Hi-C library sequencing information. Percentage for double-stranded (DS) reads indicates percentage of DS reads to all reads, and percentage for valid pairs and filtered reads indicates percentage of valid pairs and filtered reads to DS reads. **b.** Frequency distribution of Hi-C contacts in GZ (left) and CP (right) **c.** Size distribution of topological associating domains (TADs) in GZ (left) and CP (right). **d.** Size distribution of genomic regions in between TADs that are less than 400kb (TAD boundaries) in GZ (left) and CP (right). **e.** Size distribution of genomic regions in between TADs that are bigger than 400kb (unorganized chromosome) in GZ (left) and CP (right). Cis ratio, ratio of cis (intra-chromosomal) reads to the total number of reads; chr, chromosome.

**Extended Data Figure 2. Chromosome conformation is associated with various genomic features.** a. Spearman correlation of principal components (PCs) of chromatin interaction profile of CP with GC content (GC), gene number, DNase I hypersensitivity (DHS), and gene expression level of fetal brains. **b.** GO enrichment of genes located in the top 1000 regions that gain inter-chromosomal interactions in CP compared to ES (upper left), ES compared to CP (upper right), CP compared to GZ (lower left), and GZ compared to CP (lower right). **c.** Top 5% (left) and 10% (middle) highest interacting regions both in GZ and CP (High) show positive correlation in gene expression, while low interacting regions (Low) and variant interacting regions (Variant) have no skew in distribution. (Right) Mean (top) and median (bottom) values for gene expression correlation for high, low, and variant interacting regions with different cutoffs, indicating that higher the interaction, higher the correlation of gene expression. **d.** Percentage of epigenetic marks for genomic regions that change compartment status between ES and GZ. Note that B to A shift in ES to GZ is associated with increased proportion of active transcribed regions (TssA and Tx) and enhancers (Enh, top), while A to B shift in ES to GZ is associated with increased proportions of repressive marks (Het and ReprPCWk, bottom). P-values from Fisher's exact test. Annotation for epigenetic marks described in a core 15-state model from http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 3. Interacting regions share epigenetic states. a.** Epigenetic state combination in inter-chromosomal interacting regions in CP. Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1), transcriptional regulatory regions (TxReg), and transcribed regions (Tx) interact highly to each other as marked in red. **b-c.** Epigenetic state combination in intra-chromosomal interacting regions in GZ (**b**) and CP (**c**). Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1) and transcriptional regulatory regions (TxReg) interact highly to promoters (PromD1, PromD2) and transcribed regions (Tx5', Tx) as marked in red. Inter- and intra-chromosomal contact frequency map is compared to epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Colored bars on the left represent epigenetic marks associated with promoters and transcribed regions (orange), enhancers (red), and repressive marks (blue). Annotation for epigenetic marks described in a 25-state model from http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 4. Characterization of chromatin interactome of human-gained enhancers. a.** Distribution fitting of normalized chromatin interaction frequency between human-gained enhancers with 1Mb upstream (top) and 100kb upstream (bottom) regions. Weibull distribution (red line) fits Hi-C interaction frequency the best for every distance range. **b.** Distribution of the number of significant interacting loci to human-gained enhancers in GZ (top), CP (middle), and ES (bottom). **c.** Fraction of histone states (left) and epigenetic mark enrichment (right) for regions interacting with

human-gained enhancers in GZ and CP. CDF, cumulative distribution function; Annotation for epigenetic marks described in http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 5. Human-gained enhancers interact to evolutionary lineage-specific long non-coding RNAs (lncRNAs). a.** Protein-coding genes interacting with human-gained enhancers in CP (CP) and GZ (GZ) have lower non-synonymous substitutions (dN)/synonymous substitutions (dS) ratio compared to protein-coding genes non-interacting to human-gained enhancers (All) in mammals (mouse), primates (rhesus macaque), and great apes (chimpanzee), indicative of purifying selection. **b.** Number of lineage-specific lncRNAs interacting to human-gained enhancers (red vertical lines in the graph) in GZ (top) and CP (bottom). Null distribution generated from 3,000 permutations, where the number of lncRNAs interacting to the same number of enhancers pooled from all fetal brain enhancers was counted.

**Extended Data Figure 6. Association between eQTL and Hi-C interaction. a.** Overlap between eQTL transcripts and genes physically interacting to eQTL SNPs in CP and GZ. Significance of the overlap between eQTL transcripts and Hi-C interacting genes described in the upper right (Fisher's exact test). Background gene list for Fisher's exact test is all transcripts assessed in eQTL study within 1Mb from eQTL SNPs. **b-d.** Histone state enrichment for eQTL SNPs in adult frontal cortex (FCTX, **b**), fetal brain (FB, **c**), and IMR90 (**d**). **e.** Hi-C interaction frequency between eQTL SNPs and transcripts is greater than expected by chance in the relevant cell type. Lowess smooth curve plotted with actual data points. CP, chromatin contact frequency in CP; GZ, chromatin contact frequency in GZ; ES, chromatin contact frequency in ES; Exp, expected interaction frequency given the distance between two regions; Opp, opposite interaction frequency: interaction frequency of SNPs and transcripts when the position of genes was mirrored relative to the eQTL SNP. ***P<0.001, P-values from repeated measure of ANOVA.

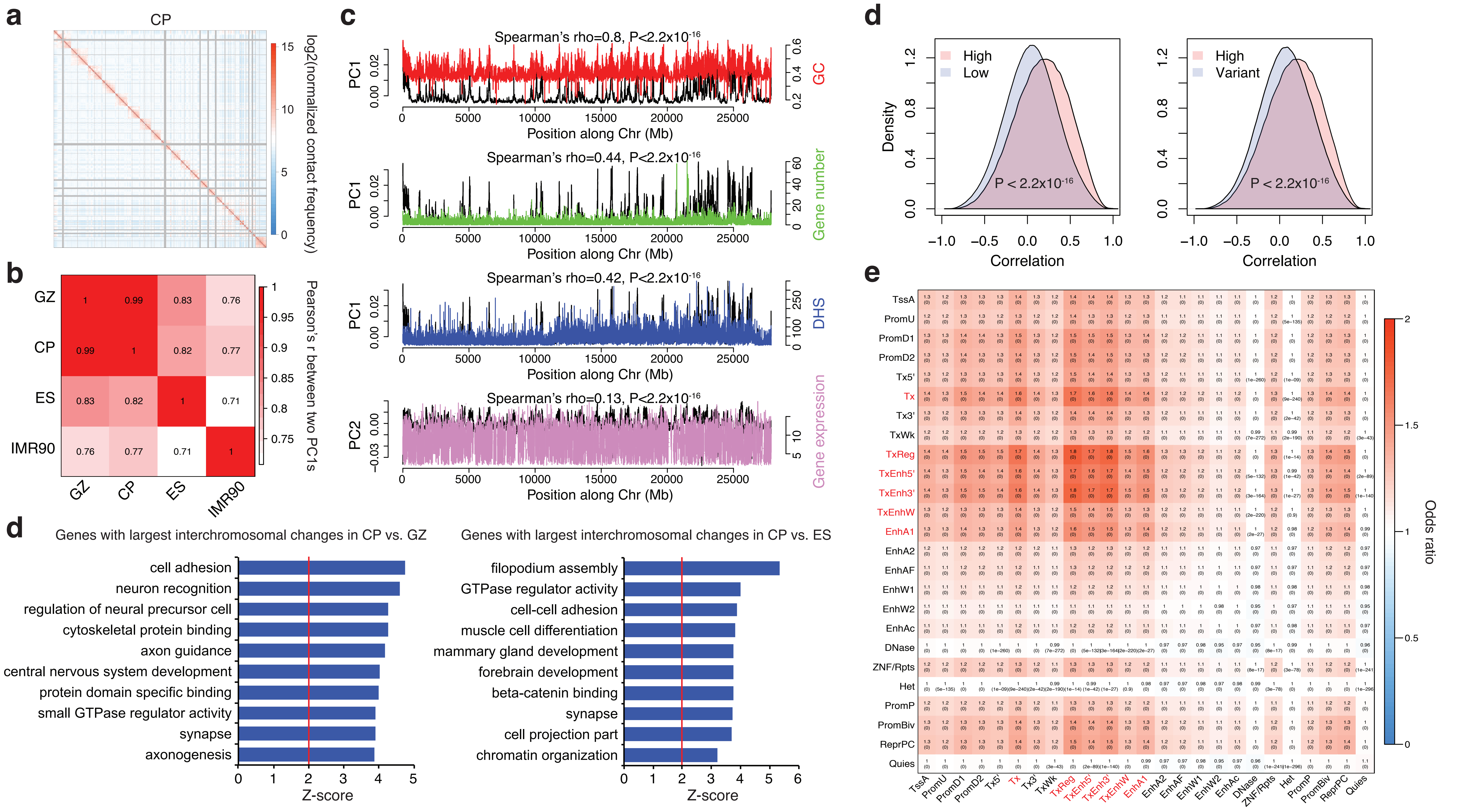
**Extended Data Figure 7. Defining schizophrenia risk genes based on functional annotation of credible SNPs.** Credible SNPs were selected using CAVIAR and categorized into functional SNPs, SNPs that fall onto gene promoters, and un-annotated SNPs. Histone state enrichment of credible SNPs was assessed in fetal brain (FB) and adult frontal cortex (FCTX). Functional SNPs and promoter SNPs were directly assigned to the target genes, while un-annotated SNPs were assigned to the target genes via Hi-C interactions in CP and GZ. GO enrichment for genes identified by each category is shown in the bottom. NMD, nonsense-mediated decay; TSS, transcription start site.

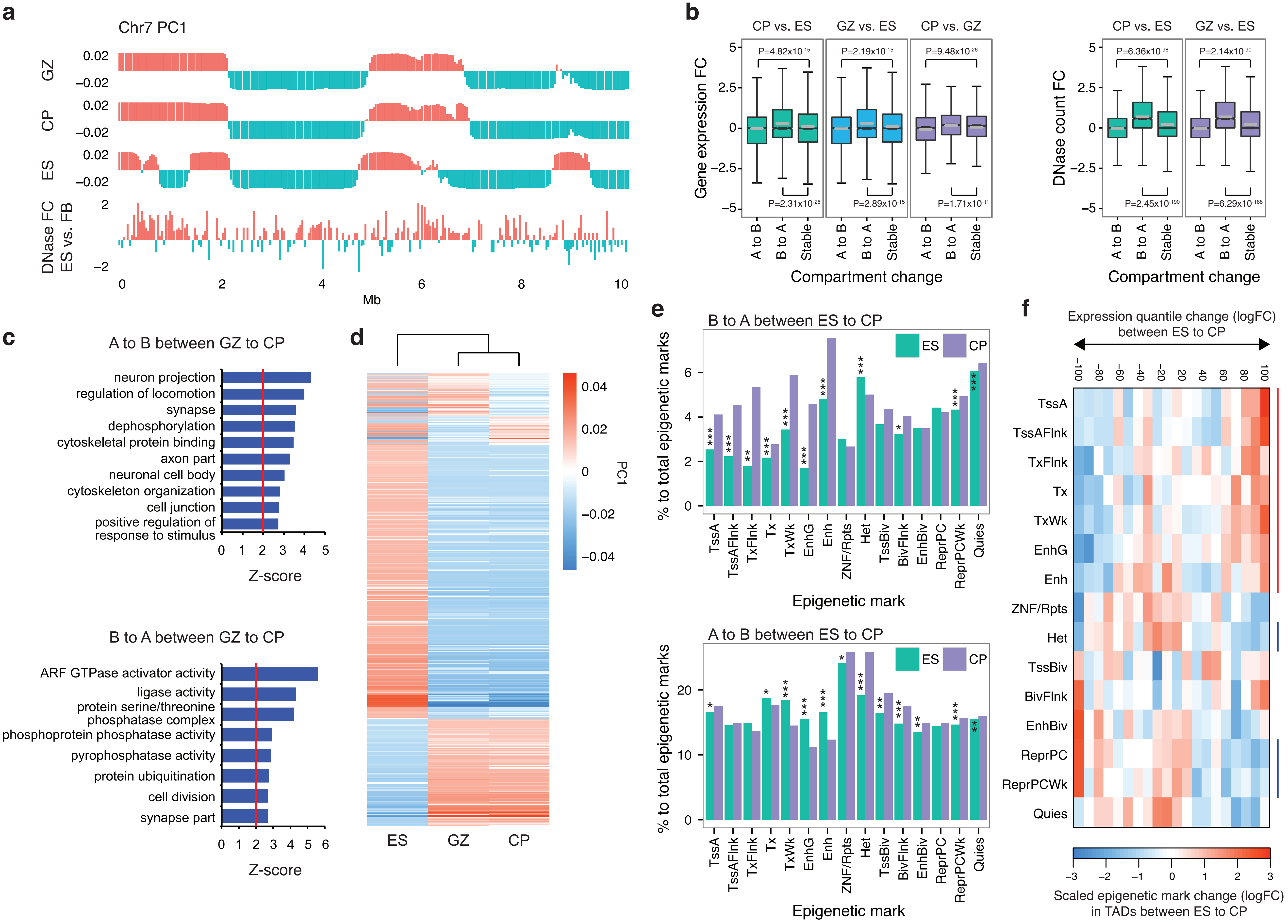
**Extended Data Figure 8. Representative interaction maps for credible SNPs to 1Mb flanking regions.** Interaction maps provide gene of actions for credible SNPs based on physical interaction. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; SNP, genomic coordinate for a 10kb bin in which credible SNPs locate; -log10(P-value), P-value for the significance of the interaction between credible SNPs and each 10kb bin, grey dashed line for FDR=0.01; GWAS loci, linkage disequilibrium (LD) region with the index SNP; TAD, TAD borders in CP, GZ, and ES.

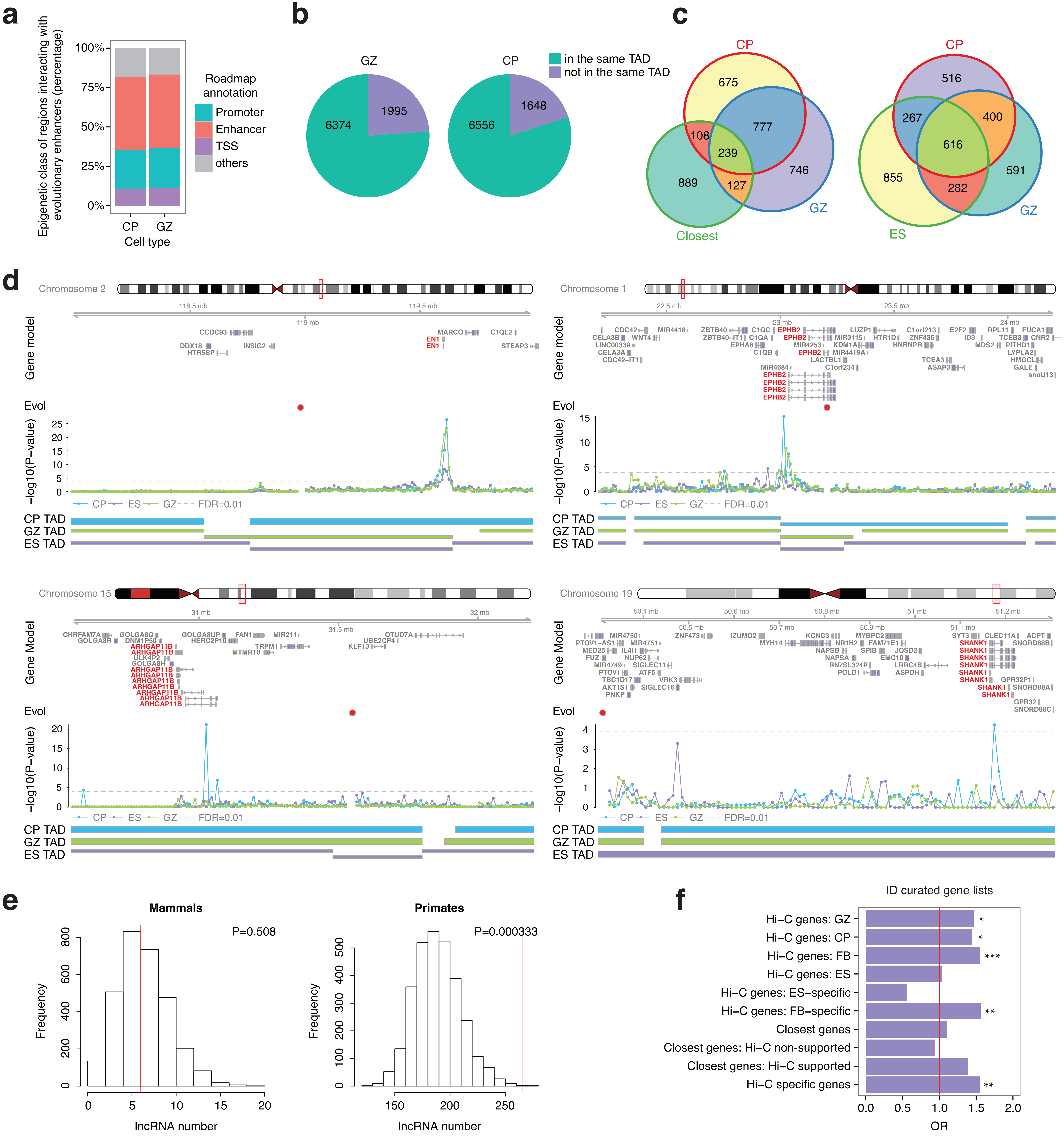
**Extended Data Figure 9. GO enrichment for schizophrenia risk genes curated by various methods. a-b.** GO enrichment for the closest genes to index SNPs (**a**) and genes in linkage disequilibrium (LD) with index SNPs (**b**) that are identified by a schizophrenia risk gene assessment pipeline in **Extended Data Fig. 7** (right) vs. not (left). **c.** GO enrichment for schizophrenia risk genes identified by a pipeline in **Extended Data Fig. 7** that are neither the closest genes nor in LD to index SNPs. Intersect and

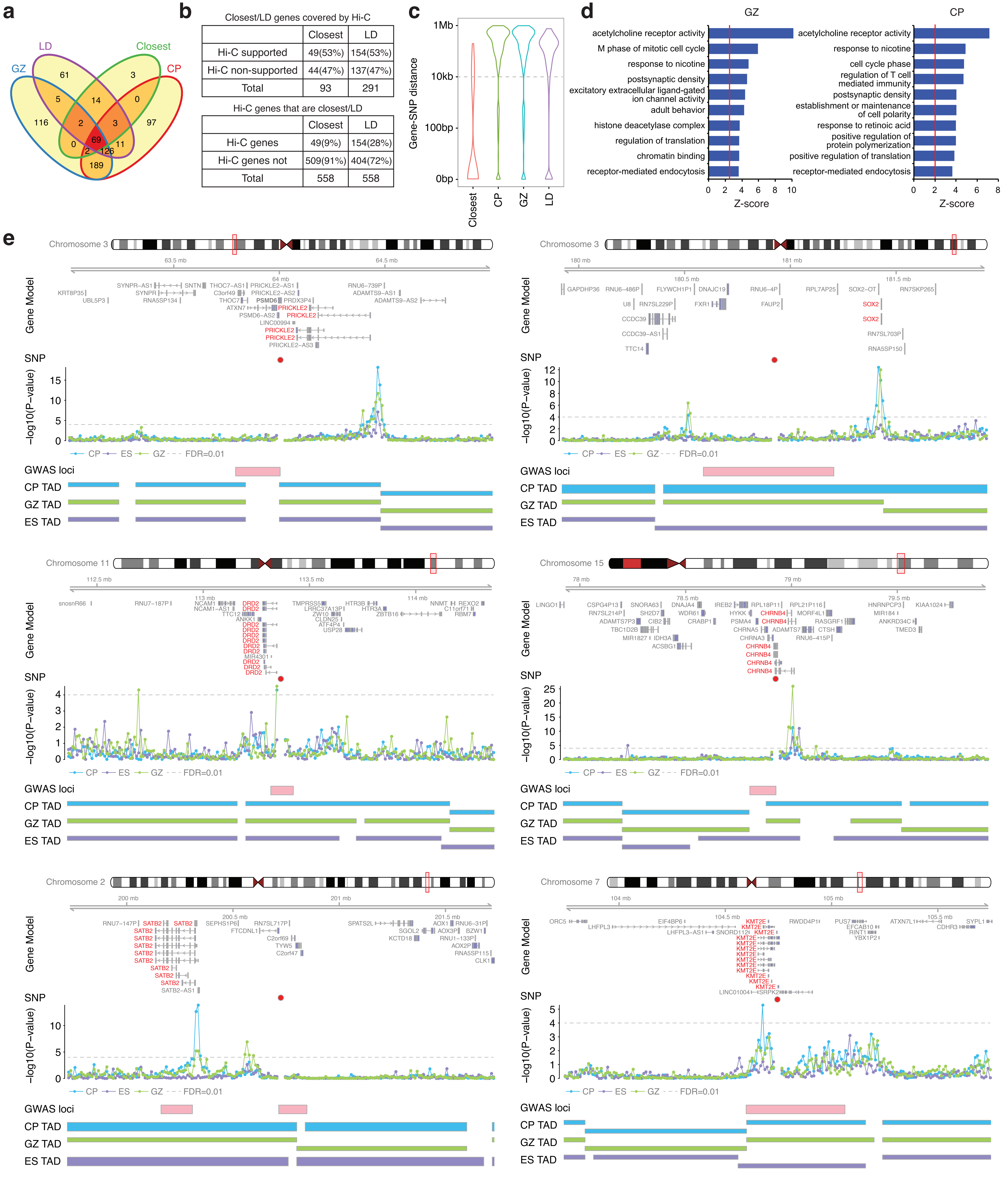
union between CP and GZ in left and right, respectively. Venn diagrams are marked in orange to depict the gene list assessed for GO enrichment.

**Extended Data Figure 10. Defining schizophrenia risk genes based on functional annotation of another set of credible SNPs.** Credible SNPs defined in the original study were categorized into functional SNPs, SNPs that fall onto gene promoters, and un-annotated SNPs. Overlap between credible SNPs identified by CAVIAR and credible SNPs originally identified indicates that two credible SNP lists overlap with each other. Histone state enrichment of credible SNPs in fetal brain (FB) and adult frontal cortex (FCTX). Functional SNPs and promoter SNPs were directly assigned to the target genes, while un-annotated SNPs were assigned to the target genes via Hi-C interactions in CP and GZ. GO enrichment for genes identified by each category and combined gene list is shown in the bottom. NMD, nonsense-mediated decay; TSS, transcription start site.


a
CP
log2(normalized contact frequency)
15
10
5
0
b
GZ
CP
ES
IMR90
1
0.99
0.83
0.76
0.99
1
0.82
0.77
0.83
0.82
1
0.71
0.76
0.77
0.71
1
Pearson's r between two PC1s
1
0.95
0.9
0.85
0.8
0.75
c
Spearman's rho=0.8, P<2.2x10^-16
PC1
GC
Spearman's rho=0.44, P<2.2x10^-16
Gene number
Spearman's rho=0.42, P<2.2x10^-16
DHS
Spearman's rho=0.13, P<2.2x10^-16
PC2
Gene expression
Position along Chr (Mb)
0
5000
10000
15000
20000
25000
d
High
Low
Variant
Density
P < 2.2x10^-16
Correlation
−1.0
−0.5
0.0
0.5
1.0
d
Genes with largest interchromosomal changes in CP vs. GZ
cell adhesion
neuron recognition
regulation of neural precursor cell
cytoskeletal protein binding
axon guidance
central nervous system development
protein domain specific binding
small GTPase regulator activity
synapse
axonogenesis
Z-score
Genes with largest interchromosomal changes in CP vs. ES
filopodium assembly
GTPase regulator activity
cell-cell adhesion
muscle cell differentiation
mammary gland development
forebrain development
beta-catenin binding
synapse
cell projection part
chromatin organization
Z-score
e
TssA
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
Odds ratio
2
1.5
1
0.5
0


a
Chr7 PC1
GZ
CP
ES
0.02
−0.02
DNase FC
ES vs. FB
2
−2
0
2
4
6
8
10
Mb
b
CP vs. ES
GZ vs. ES
CP vs. GZ
Gene expression FC
P=4.82x10-15
P=2.31x10-26
P=2.19x10-15
P=2.89x10-15
P=9.48x10-26
P=1.71x10-11
DNase count FC
P=6.36x10-98
P=2.45x10-190
P=2.14x10-90
P=6.29x10-188
5
2.5
0
−2.5
−5
A to B
B to A
Stable
Compartment change
c
A to B between GZ to CP
neuron projection
regulation of locomotion
synapse
dephosphorylation
cytoskeletal protein binding
axon part
neuronal cell body
cytoskeleton organization
cell junction
positive regulation of response to stimulus
Z-score
B to A between GZ to CP
ARF GTPase activator activity
ligase activity
protein serine/threonine phosphatase complex
phosphoprotein phosphatase activity
pyrophosphatase activity
protein ubiquitination
cell division
synapse part
Z-score
d
ES
GZ
CP
PC1
0.04
0.02
0
−0.02
−0.04
e
B to A between ES to CP
A to B between ES to CP
ES
CP
% to total epigenetic marks
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
Epigenetic mark
f
Expression quantile change (logFC) between ES to CP
−100
−80
−60
−40
−20
20
40
60
80
100
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
−3
−2
−1
0
1
2
3
Scaled epigenetic mark change (logFC) in TADs between ES to CP


a
Epigenetic class of regions interacting with evolutionary enhancers (percentage)
100%
75%
50%
25%
0%
CP
GZ
Cell type
Roadmap annotation
Promoter
Enhancer
TSS
others
b
GZ
6374
1995
CP
6556
1648
in the same TAD
not in the same TAD
c
CP
675
777
108
239
746
889
127
GZ
Closest
CP
516
267
400
616
855
282
591
ES
GZ
d
Chromosome 2
118.5 mb
119 mb
119.5 mb
CCDC93
DDX18
INSIG2
HTR5BP
MARCO
C1QL2
EN1
EN1
STEAP3
Gene model
Evol
-log10(P-value)
CP
ES
GZ
FDR=0.01
CP TAD
GZ TAD
ES TAD
Chromosome 1
22.5 mb
23 mb
23.5 mb
24 mb
EPHB2
Chromosome 15
31 mb
31.5 mb
32 mb
ARHGAP11B
Gene Model
Chromosome 19
50.4 mb
50.5 mb
50.6 mb
50.7 mb
50.8 mb
50.9 mb
51 mb
51.1 mb
51.2 mb
SHANK1
e
Mammals
P=0.508
Frequency
lncRNA number
Primates
P=0.000333
f
ID curated gene lists
Hi–C genes: GZ
Hi–C genes: CP
Hi–C genes: FB
Hi–C genes: ES
Hi–C genes: ES–specific
Hi–C genes: FB–specific
Closest genes
Closest genes: Hi–C non-supported
Closest genes: Hi–C supported
Hi–C specific genes
*
*
***
**
**
OR
0.0
0.5
1.0
1.5
2.0



b

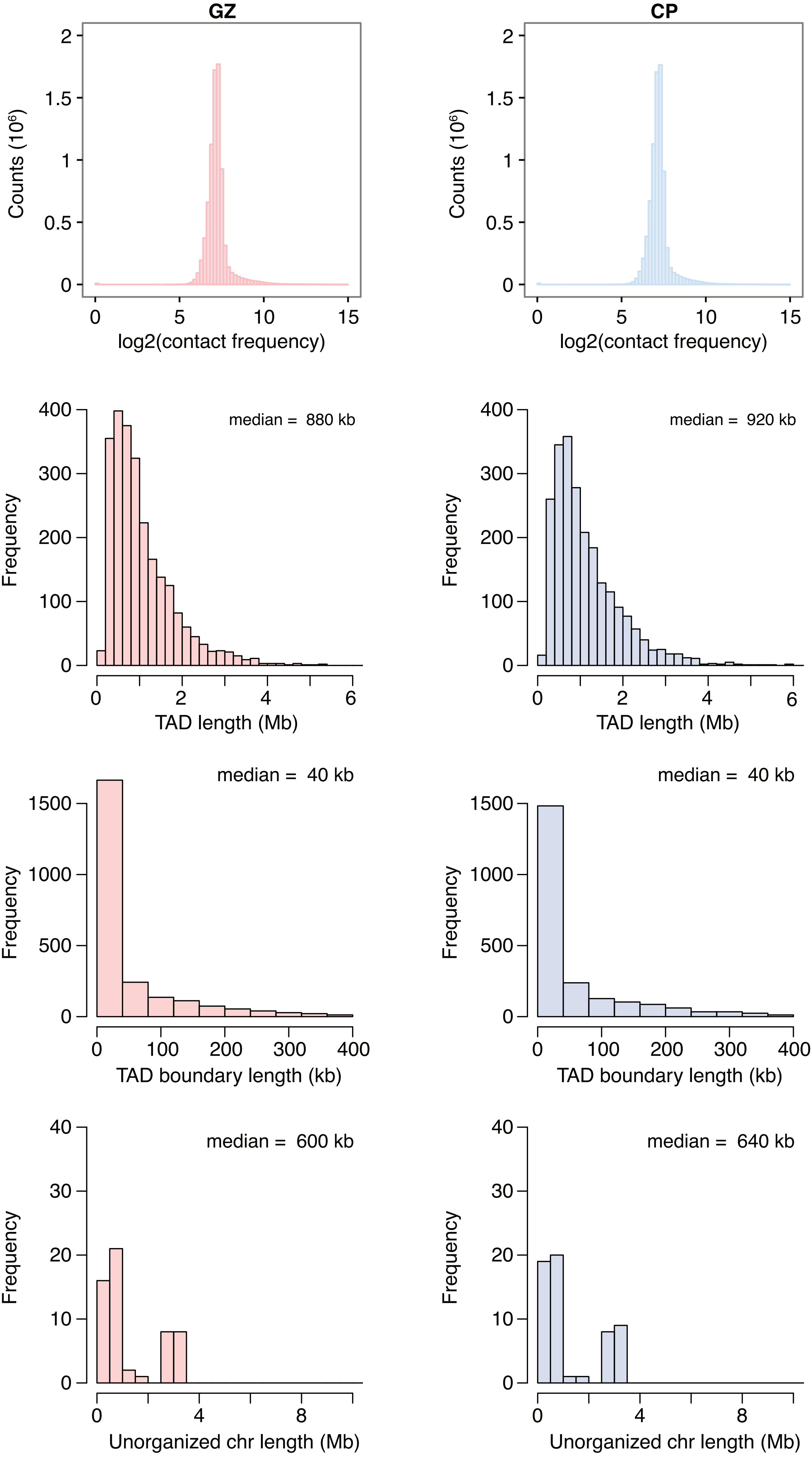
Closest/LD genes covered by Hi-C

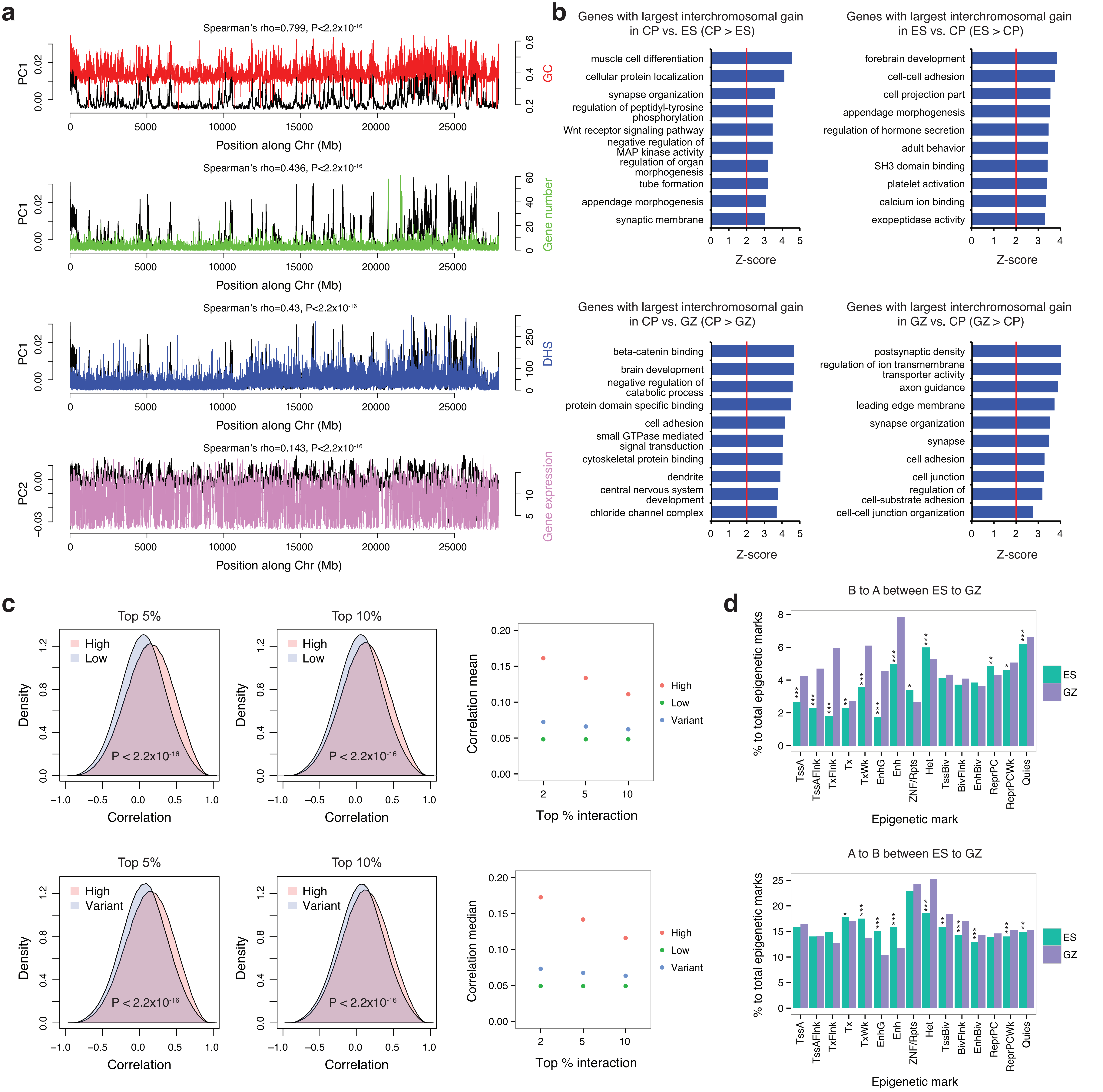
| | Closest | LD |
|---|---|---|
| Hi-C supported | 49(53%) | 154(53%) |
| Hi-C non-supported | 44(47%) | 137(47%) |
| Total | 93 | 291 |

Hi-C genes that are closest/LD

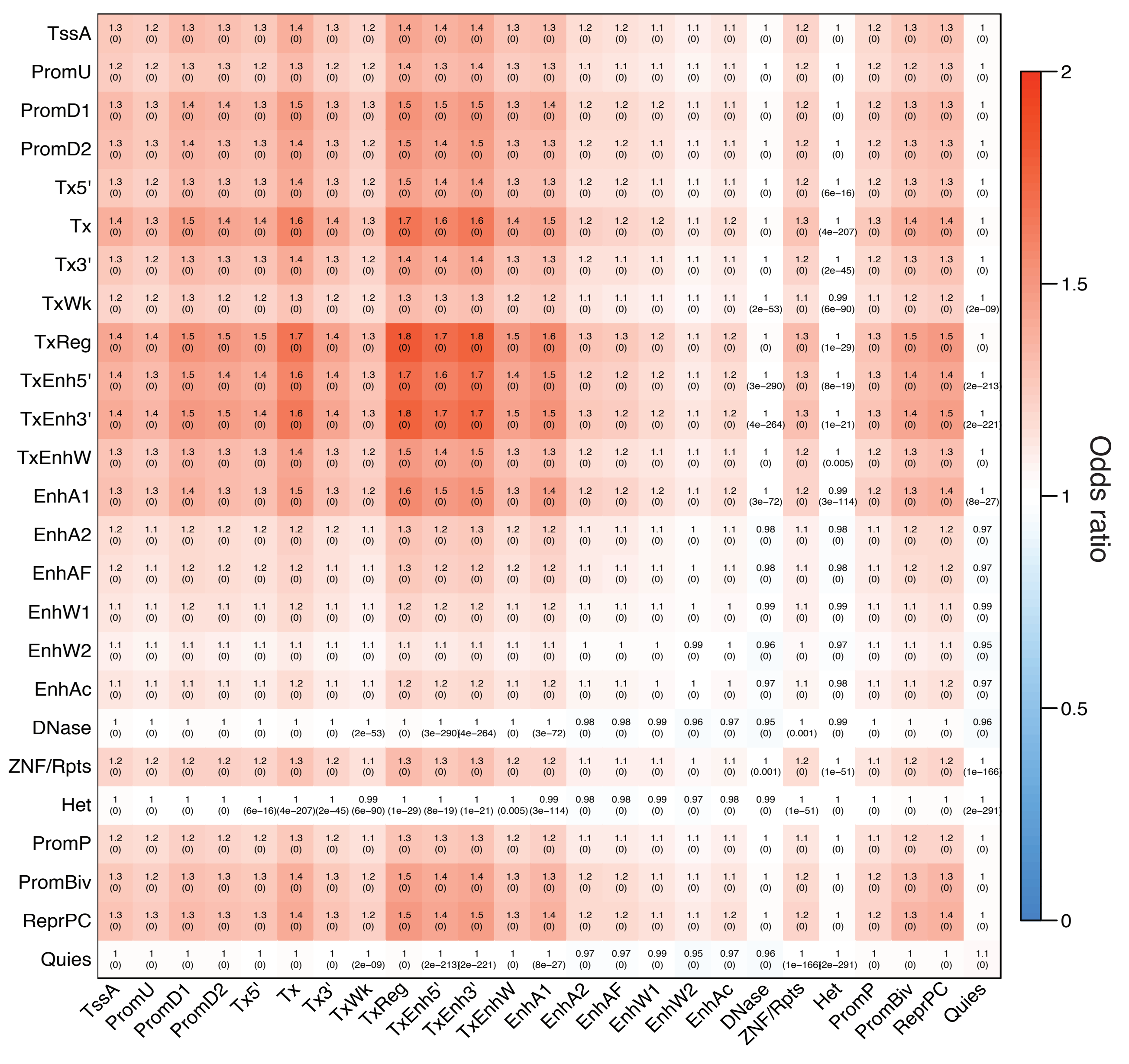
| | Closest | LD |
|---|---|---|
| Hi-C genes | 49(9%) | 154(28%) |
| Hi-C genes not | 509(91%) | 404(72%) |
| Total | 558 | 558 |

| Cell type | Cis ratio | All reads | DS mapped reads | Valid pairs | Filtered reads |
|---|---|---|---|---|---|
| GZ | 47.45% | 1,991,686,360 | 1,407,918,128<br>(70.69%) | 1,243,116,106<br>(88.29%) | 1,048,911,579<br>(74.50%) |
| CP | 46.40% | 1,958,637,304 | 1,352,951,087<br>(69.08%) | 1,225,315,488<br>(90.57%) | 1,022,593,960<br>(75.58%) |

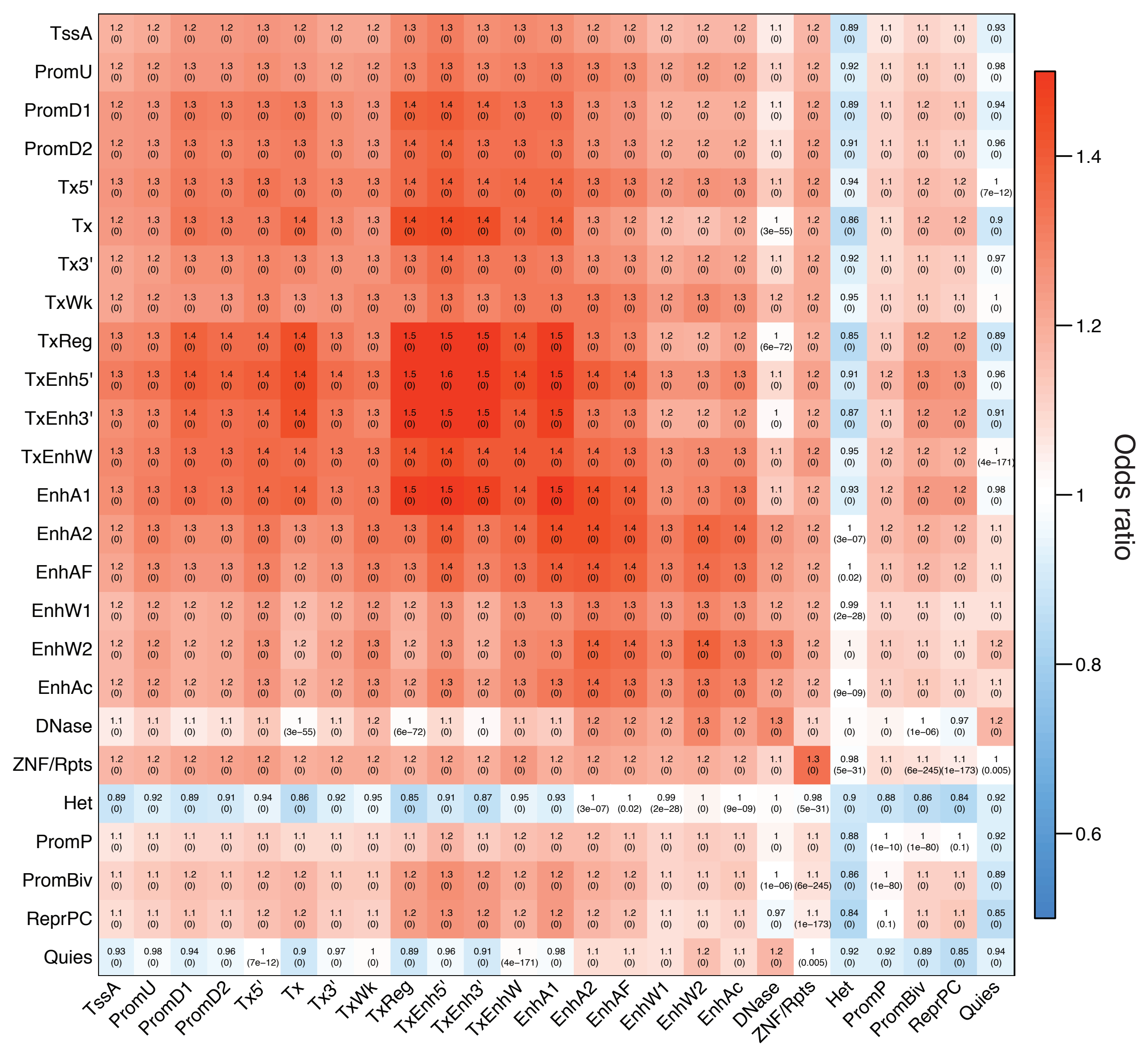



a
Spearman's rho=0.799, P<2.2x10-16
PC1
GC
Position along Chr (Mb)
Spearman's rho=0.436, P<2.2x10-16
PC1
Gene number
Position along Chr (Mb)
Spearman's rho=0.43, P<2.2x10-16
PC1
DHS
Position along Chr (Mb)
Spearman's rho=0.143, P<2.2x10-16
PC2
Gene expression
Position along Chr (Mb)
0
5000
10000
15000
20000
25000
b
Genes with largest interchromosomal gain in CP vs. ES (CP > ES)
muscle cell differentiation
cellular protein localization
synapse organization
regulation of peptidyl-tyrosine phosphorylation
Wnt receptor signaling pathway
negative regulation of MAP kinase activity
regulation of organ morphogenesis
tube formation
appendage morphogenesis
synaptic membrane
Z-score
Genes with largest interchromosomal gain in ES vs. CP (ES > CP)
forebrain development
cell-cell adhesion
cell projection part
appendage morphogenesis
regulation of hormone secretion
adult behavior
SH3 domain binding
platelet activation
calcium ion binding
exopeptidase activity
Z-score
Genes with largest interchromosomal gain in CP vs. GZ (CP > GZ)
beta-catenin binding
brain development
negative regulation of catabolic process
protein domain specific binding
cell adhesion
small GTPase mediated signal transduction
cytoskeletal protein binding
dendrite
central nervous system development
chloride channel complex
Z-score
Genes with largest interchromosomal gain in GZ vs. CP (GZ > CP)
postsynaptic density
regulation of ion transmembrane transporter activity
axon guidance
leading edge membrane
synapse organization
synapse
cell adhesion
cell junction
regulation of cell-substrate adhesion
cell-cell junction organization
Z-score
c
Top 5%
Top 10%
High
Low
Variant
Density
Correlation
P < 2.2x10-16
Correlation mean
Correlation median
Top % interaction
High
Low
Variant
d
B to A between ES to GZ
A to B between ES to GZ
% to total epigenetic marks
Epigenetic mark
ES
GZ
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies

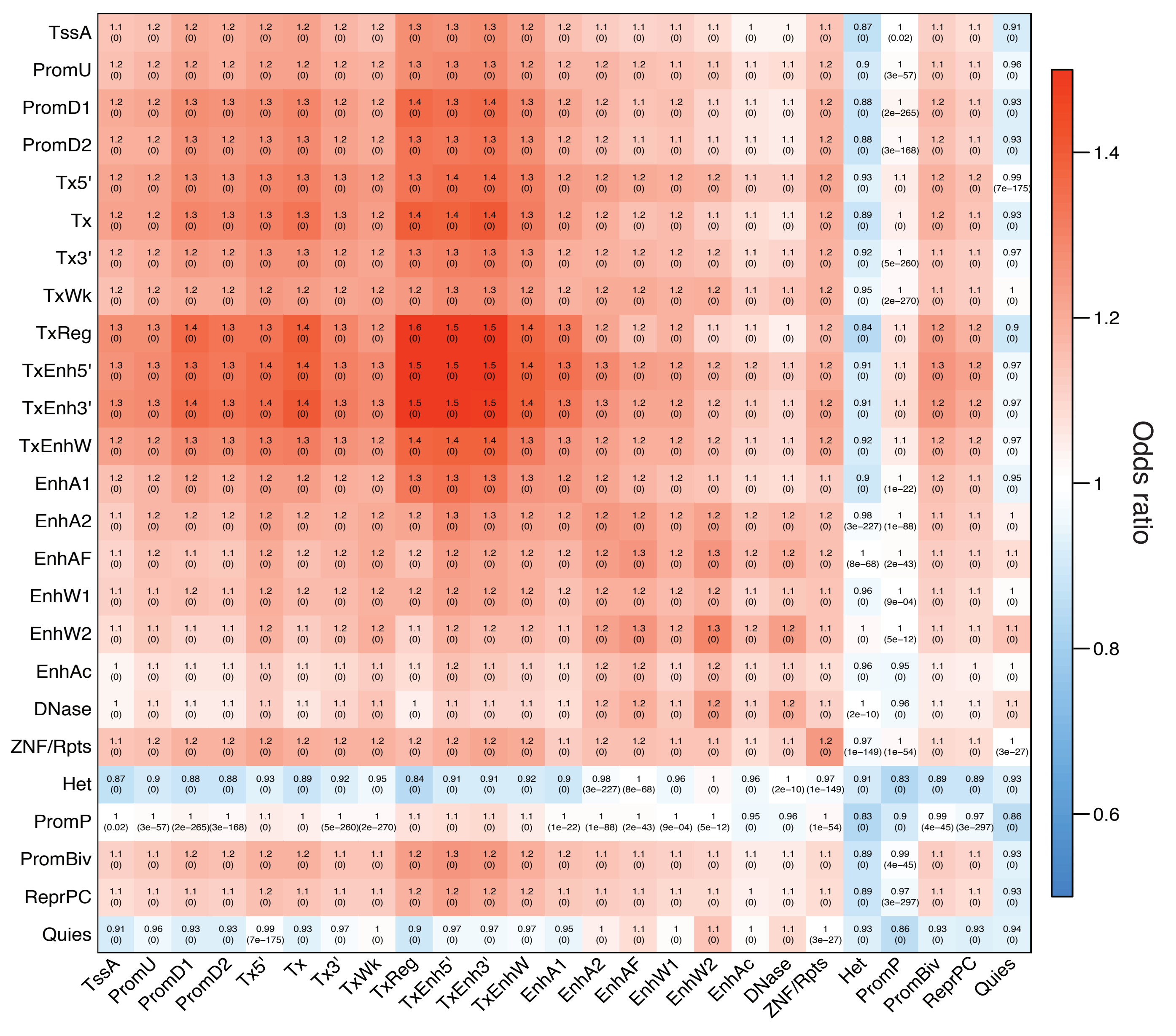
CP: interchromosomal

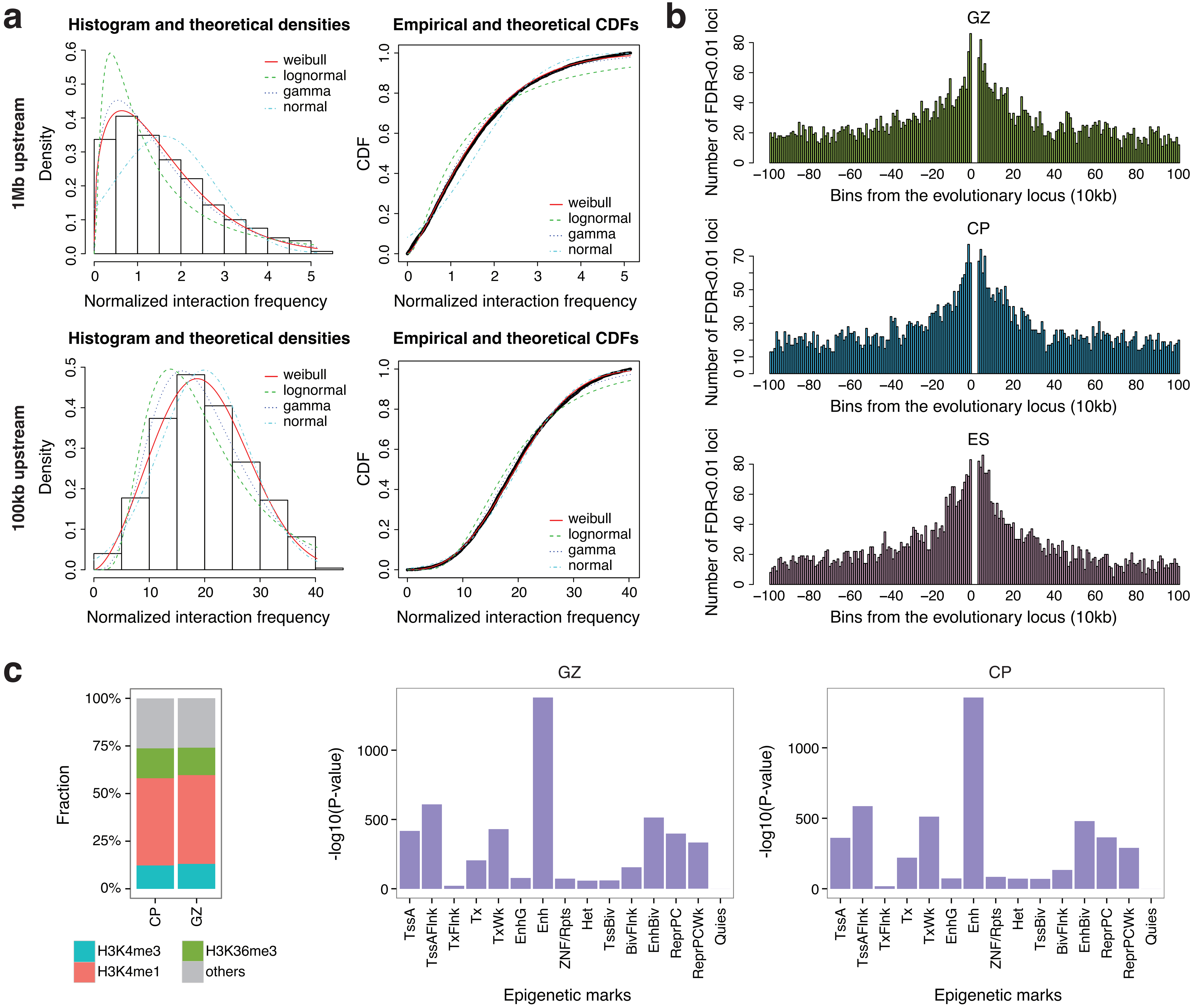


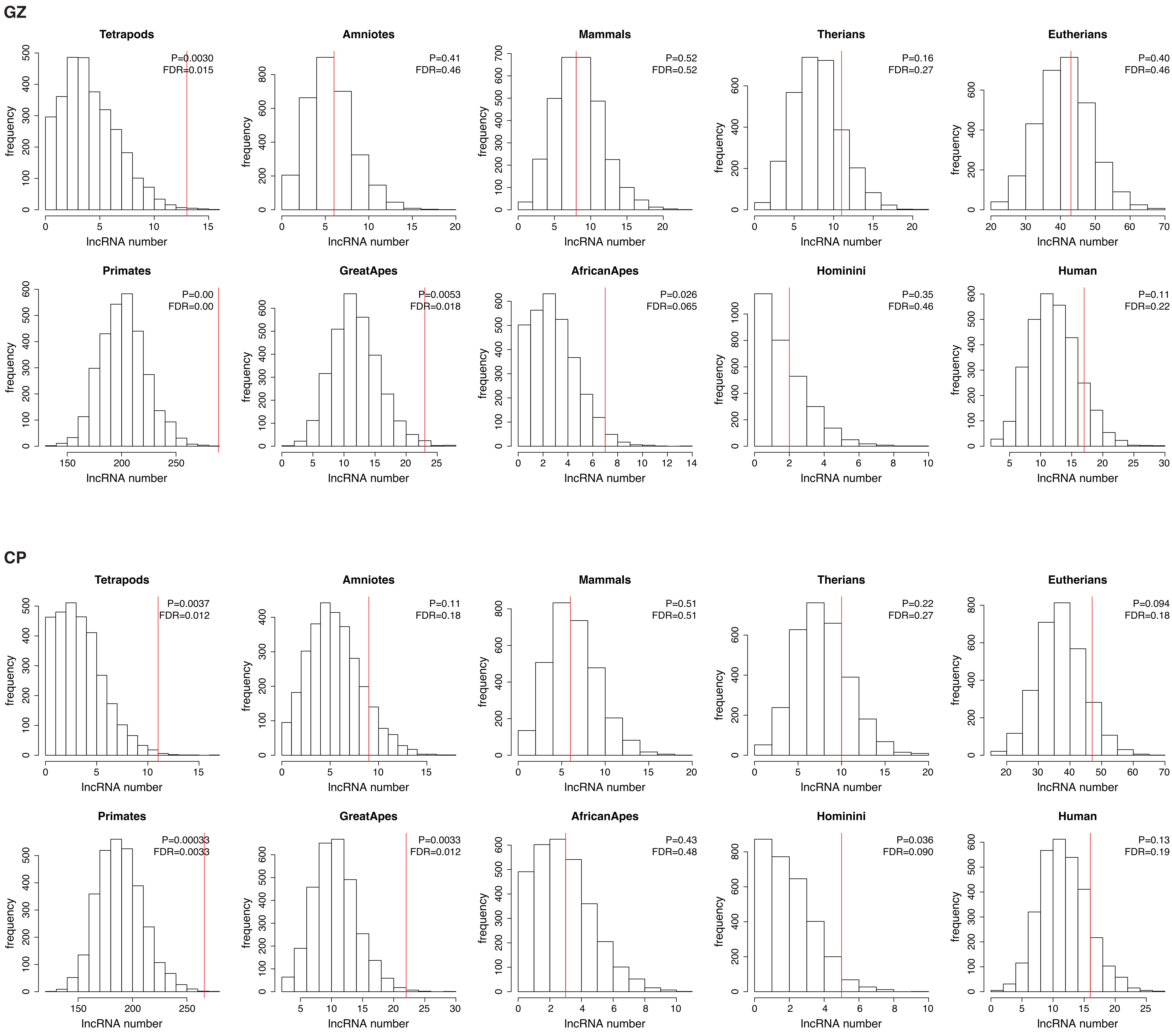
GZ: intrachromosomal

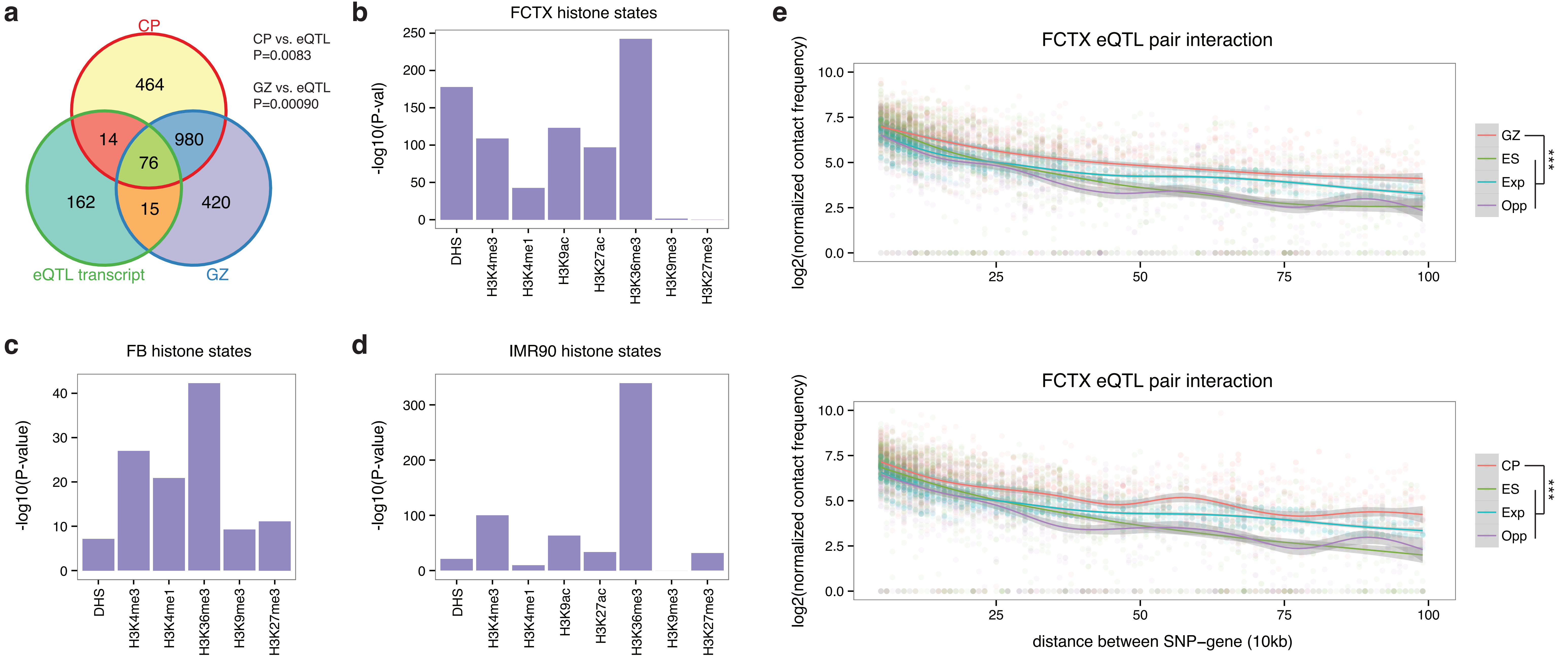


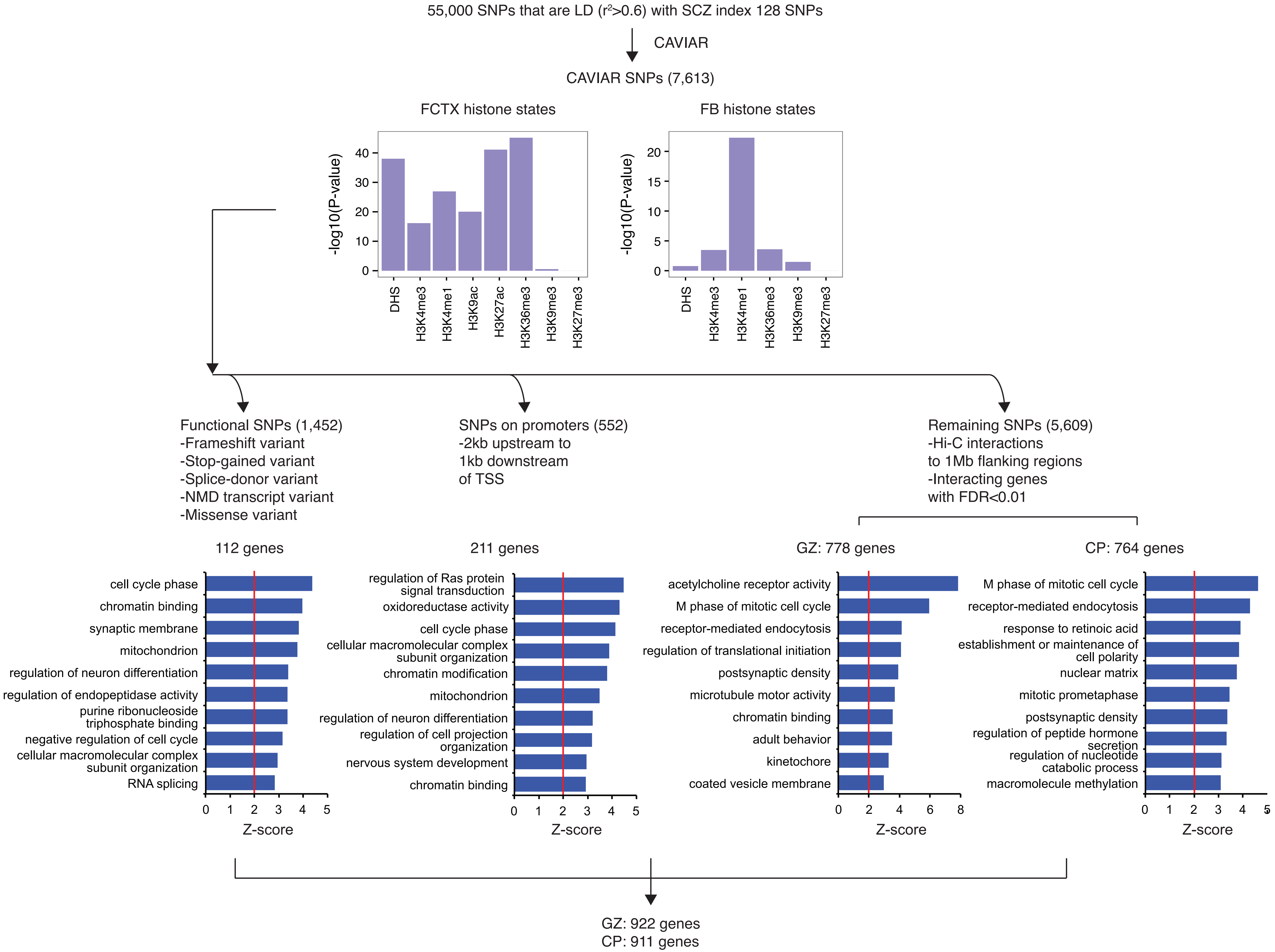
CP: intrachromosomal

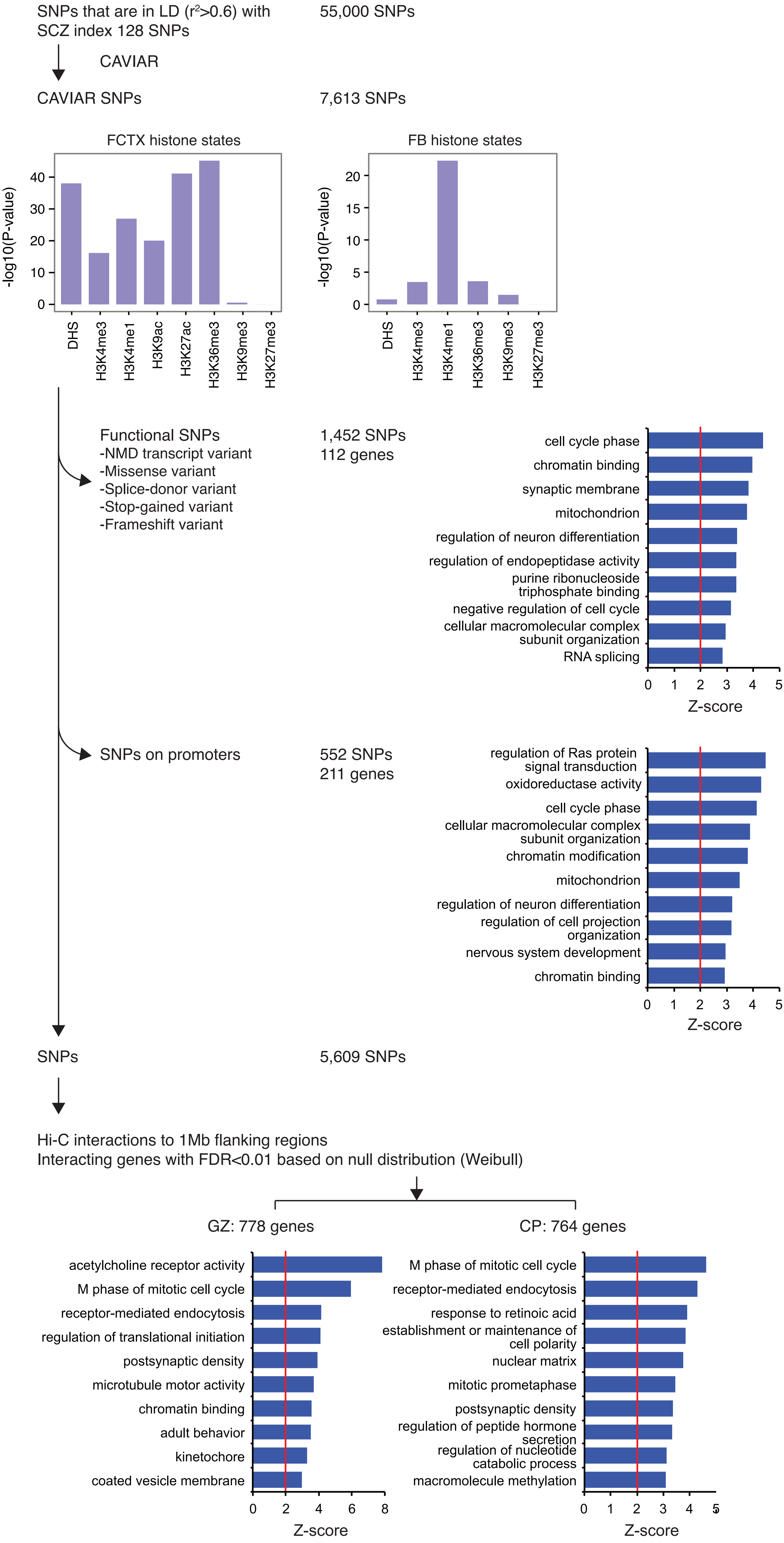


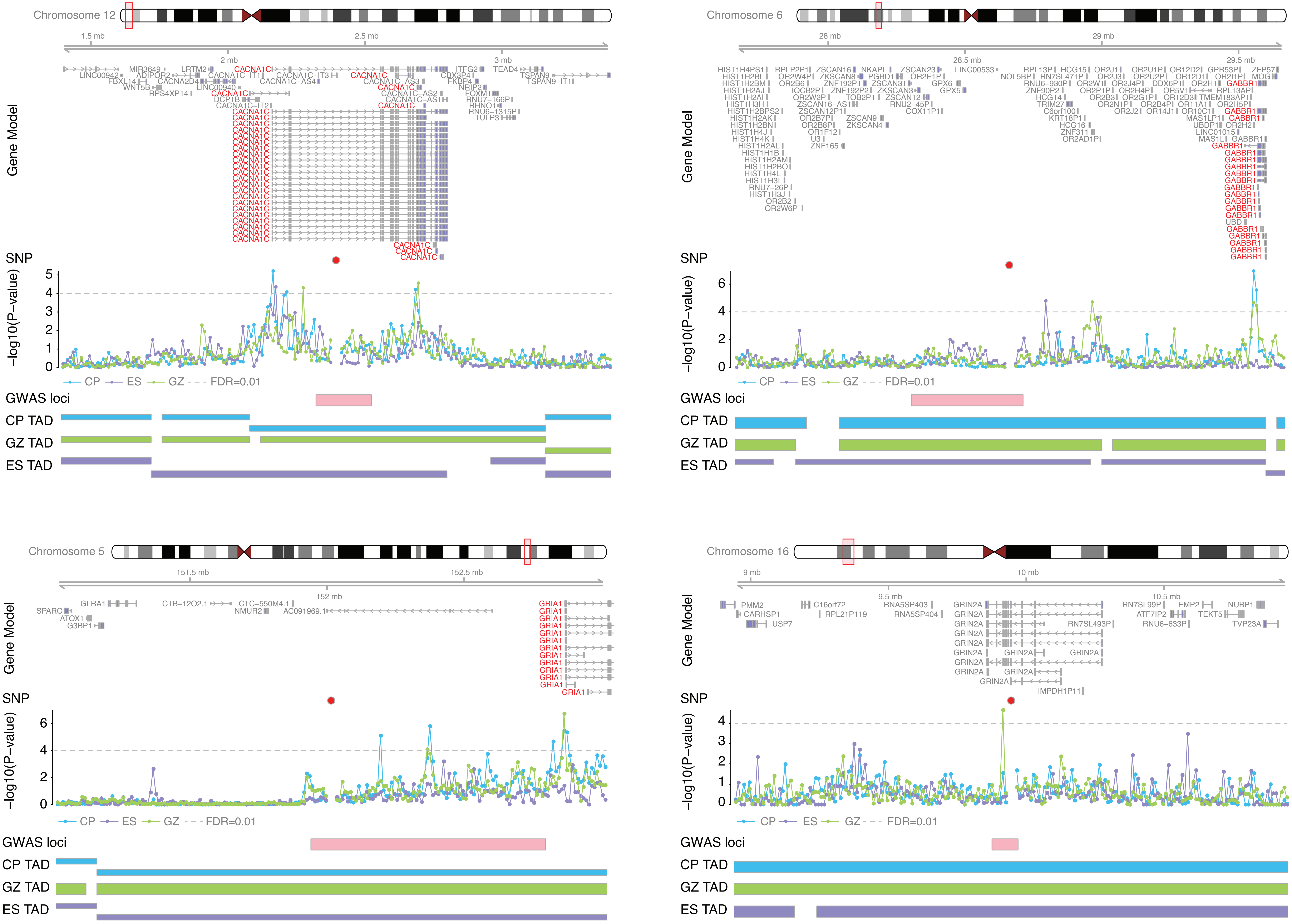

a
Histogram and theoretical densities
Empirical and theoretical CDFs
1Mb upstream
100kb upstream
Density
CDF
Normalized interaction frequency
weibull
lognormal
gamma
normal
b
GZ
CP
ES
Number of FDR<0.01 loci
Bins from the evolutionary locus (10kb)
c
Fraction
100%
75%
50%
25%
0%
CP
GZ
H3K4me3
H3K36me3
H3K4me1
others
GZ
CP
-log10(P-value)
Epigenetic marks
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies


GZ
Tetrapods
P=0.0030
FDR=0.015
Amniotes
P=0.41
FDR=0.46
Mammals
P=0.52
FDR=0.52
Therians
P=0.16
FDR=0.27
Eutherians
P=0.40
FDR=0.46
Primates
P=0.00
FDR=0.00
GreatApes
P=0.0053
FDR=0.018
AfricanApes
P=0.026
FDR=0.065
Hominini
P=0.35
FDR=0.46
Human
P=0.11
FDR=0.22
CP
Tetrapods
P=0.0037
FDR=0.012
Amniotes
P=0.11
FDR=0.18
Mammals
P=0.51
FDR=0.51
Therians
P=0.22
FDR=0.27
Eutherians
P=0.094
FDR=0.18
Primates
P=0.00033
FDR=0.0033
GreatApes
P=0.0033
FDR=0.012
AfricanApes
P=0.43
FDR=0.48
Hominini
P=0.036
FDR=0.090
Human
P=0.13
FDR=0.19
frequency
lncRNA number

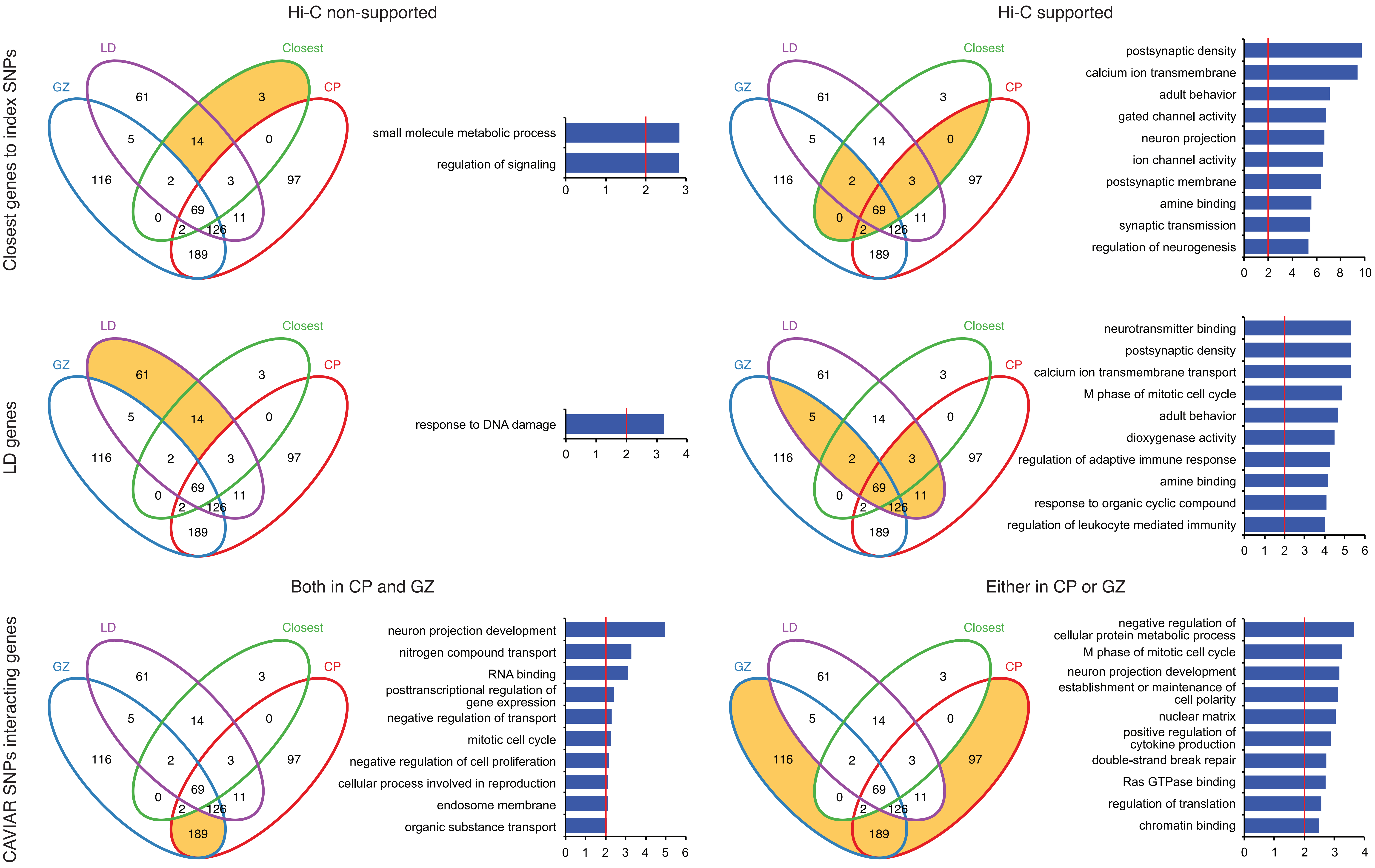

a
CP
464
14
980
76
162
15
420
eQTL transcript
GZ
CP vs. eQTL
P=0.0083
GZ vs. eQTL
P=0.00090
b
FCTX histone states
-log10(P-val)
250
200
150
100
50
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
c
FB histone states
-log10(P-value)
40
30
20
10
0
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
d
IMR90 histone states
-log10(P-value)
300
200
100
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
e
FCTX eQTL pair interaction
log2(normalized contact frequency)
10.0
7.5
5.0
2.5
0.0
25
50
75
100
GZ
ES
Exp
Opp
***
FCTX eQTL pair interaction
log2(normalized contact frequency)
10.0
7.5
5.0
2.5
0.0
25
50
75
100
distance between SNP–gene (10kb)
CP
ES
Exp
Opp
***

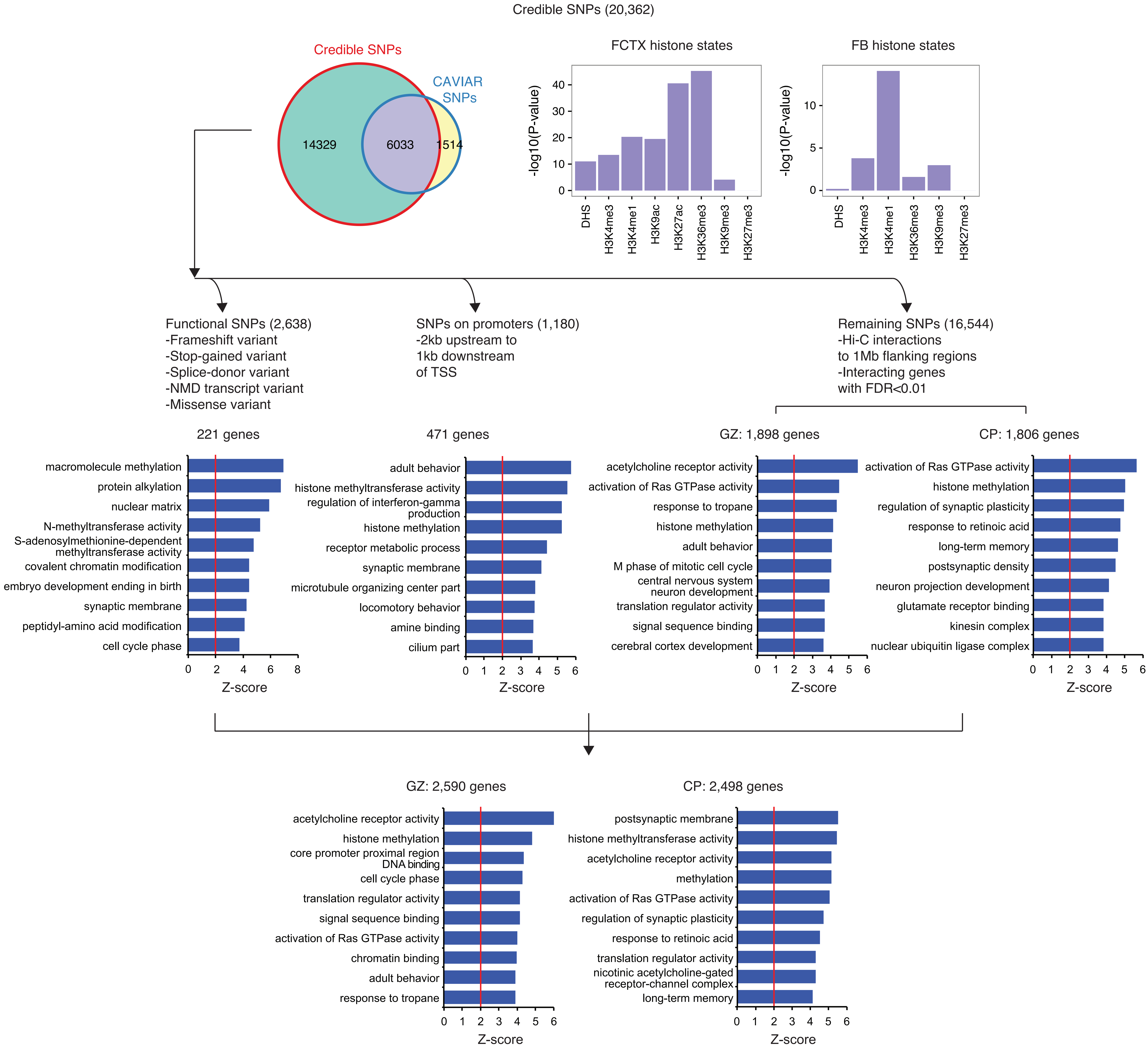

55,000 SNPs that are LD (r2>0.6) with SCZ index 128 SNPs
CAVIAR
CAVIAR SNPs (7,613)
FCTX histone states
-log10(P-value)
0
10
20
30
40
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
FB histone states
-log10(P-value)
0
5
10
15
20
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs (1,452)
-Frameshift variant
-Stop-gained variant
-Splice-donor variant
-NMD transcript variant
-Missense variant
SNPs on promoters (552)
-2kb upstream to
1kb downstream
of TSS
Remaining SNPs (5,609)
-Hi-C interactions
to 1Mb flanking regions
-Interacting genes
with FDR<0.01
112 genes
cell cycle phase
chromatin binding
synaptic membrane
mitochondrion
regulation of neuron differentiation
regulation of endopeptidase activity
purine ribonucleoside triphosphate binding
negative regulation of cell cycle
cellular macromolecular complex subunit organization
RNA splicing
0 1 2 3 4 5
Z-score
211 genes
regulation of Ras protein signal transduction
oxidoreductase activity
cell cycle phase
cellular macromolecular complex subunit organization
chromatin modification
mitochondrion
regulation of neuron differentiation
regulation of cell projection organization
nervous system development
chromatin binding
0 1 2 3 4 5
Z-score
GZ: 778 genes
acetylcholine receptor activity
M phase of mitotic cell cycle
receptor-mediated endocytosis
regulation of translational initiation
postsynaptic density
microtubule motor activity
chromatin binding
adult behavior
kinetochore
coated vesicle membrane
0 2 4 6 8
Z-score
CP: 764 genes
M phase of mitotic cell cycle
receptor-mediated endocytosis
response to retinoic acid
establishment or maintenance of cell polarity
nuclear matrix
mitotic prometaphase
postsynaptic density
regulation of peptide hormone secretion
regulation of nucleotide catabolic process
macromolecule methylation
0 1 2 3 4 5
Z-score
GZ: 922 genes
CP: 911 genes


SNPs that are in LD ($r^2$>0.6) with SCZ index 128 SNPs
55,000 SNPs
CAVIAR
CAVIAR SNPs
7,613 SNPs
FCTX histone states
-log10(P-value)
40
30
20
10
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
FB histone states
-log10(P-value)
20
15
10
5
0
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs
-NMD transcript variant
-Missense variant
-Splice-donor variant
-Stop-gained variant
-Frameshift variant
1,452 SNPs
112 genes
cell cycle phase
chromatin binding
synaptic membrane
mitochondrion
regulation of neuron differentiation
regulation of endopeptidase activity
purine ribonucleoside triphosphate binding
negative regulation of cell cycle
cellular macromolecular complex subunit organization
RNA splicing
0
1
2
3
4
5
Z-score
SNPs on promoters
552 SNPs
211 genes
regulation of Ras protein signal transduction
oxidoreductase activity
cell cycle phase
cellular macromolecular complex subunit organization
chromatin modification
mitochondrion
regulation of neuron differentiation
regulation of cell projection organization
nervous system development
chromatin binding
0
1
2
3
4
5
Z-score
SNPs
5,609 SNPs
Hi-C interactions to 1Mb flanking regions
Interacting genes with FDR<0.01 based on null distribution (Weibull)
GZ: 778 genes
acetylcholine receptor activity
M phase of mitotic cell cycle
receptor-mediated endocytosis
regulation of translational initiation
postsynaptic density
microtubule motor activity
chromatin binding
adult behavior
kinetochore
coated vesicle membrane
0
2
4
6
8
Z-score
CP: 764 genes
M phase of mitotic cell cycle
receptor-mediated endocytosis
response to retinoic acid
establishment or maintenance of cell polarity
nuclear matrix
mitotic prometaphase
postsynaptic density
regulation of peptide hormone secretion
regulation of nucleotide catabolic process
macromolecule methylation
0
1
2
3
4
5
Z-score



Chromosome 12
1.5 mb
2 mb
2.5 mb
3 mb
Gene Model
MIR3649
LRTM2
LINC00942
ADIPOR2
CACNA1C
CACNA1C-IT1
CACNA1C-IT3
CACNA1C-AS4
CACNA1C-AS3
CACNA1C-AS2
CACNA1C-AS1
CACNA1C-IT2
FBXL14
CACNA2D4
WNT5B
LINC00940
RPS4XP14
DCP1B
ITFG2
TEAD4
CBX3P4
TSPAN9
TSPAN9-IT1
FKBP4
NRIP2
FOXM1
RNU7-166P
RHNO1
RNU6-1315P
TULP3
SNP
−log10(P−value)
CP
ES
GZ
FDR=0.01
GWAS loci
CP TAD
GZ TAD
ES TAD
Chromosome 6
28 mb
28.5 mb
29 mb
29.5 mb
HIST1H4PS1
RPLP2P1
ZSCAN16
NKAPL
ZSCAN23
LINC00533
RPL13P
HCG15
OR2J1
OR2U1P
OR12D2
GPR53P
ZFP57
HIST1H2BL
OR2W4P
ZKSCAN8
PGBD1
OR2E1P
NOL5BP
RN7SL471P
OR2J3
OR2U2P
OR12D1
OR2I1P
MOG
HIST1H2BM
OR2B6
ZNF192P1
ZSCAN31
GPX6
RNU6-930P
OR2W1
OR2J4P
DDX6P1
OR2H1
GABBR1
HIST1H2AJ
IQCB2P
ZNF192P2
ZKSCAN3
GPX5
ZNF90P2
OR2P1P
OR2H4P
OR5V1
RPL13AP
HIST1H2AI
OR2W2P
TOB2P1
ZSCAN12
HCG14
OR2B3
OR2G1P
OR12D3
TMEM183AP1
HIST1H3H
ZSCAN16-AS1
RNU2-45P
TRIM27
OR2N1P
OR2B4P
OR11A1
OR2H5P
HIST1H2BPS2
ZSCAN12P1
COX11P1
C6orf100
OR2J2
OR14J1
OR10C1
HIST1H2AK
OR2B7P
ZSCAN9
KRT18P1
MAS1LP1
HIST1H2BN
OR2B8P
ZKSCAN4
HCG16
UBDP1
OR2H2
HIST1H4J
OR1F12
ZNF311
LINC01015
HIST1H4K
U3
OR2AD1P
MAS1L
HIST1H2AL
ZNF165
HIST1H1B
HIST1H2AM
HIST1H2BO
HIST1H4L
HIST1H3I
RNU7-26P
HIST1H3J
OR2B2
OR2W6P
UBD
Chromosome 5
151.5 mb
152 mb
152.5 mb
GLRA1
CTB-12O2.1
CTC-550M4.1
NMUR2
AC091969.1
SPARC
ATOX1
G3BP1
GRIA1
Chromosome 16
9 mb
9.5 mb
10 mb
10.5 mb
PMM2
C16orf72
RNA5SP403
GRIN2A
RN7SL99P
EMP2
NUBP1
CARHSP1
RPL21P119
RNA5SP404
ATF7IP2
TEKT5
USP7
RN7SL493P
RNU6-633P
TVP23A
IMPDH1P11


Hi-C non-supported
Hi-C supported
Closest genes to index SNPs
LD genes
CAVIAR SNPs interacting genes
Both in CP and GZ
Either in CP or GZ
LD
Closest
GZ
CP
61
3
5
14
0
116
2
3
97
0
69
11
2
126
189
small molecule metabolic process
regulation of signaling
0 1 2 3
postsynaptic density
calcium ion transmembrane
adult behavior
gated channel activity
neuron projection
ion channel activity
postsynaptic membrane
amine binding
synaptic transmission
regulation of neurogenesis
0 2 4 6 8 10
response to DNA damage
0 1 2 3 4
neurotransmitter binding
postsynaptic density
calcium ion transmembrane transport
M phase of mitotic cell cycle
adult behavior
dioxygenase activity
regulation of adaptive immune response
amine binding
response to organic cyclic compound
regulation of leukocyte mediated immunity
0 1 2 3 4 5 6
neuron projection development
nitrogen compound transport
RNA binding
posttranscriptional regulation of gene expression
negative regulation of transport
mitotic cell cycle
negative regulation of cell proliferation
cellular process involved in reproduction
endosome membrane
organic substance transport
0 1 2 3 4 5 6
negative regulation of cellular protein metabolic process
M phase of mitotic cell cycle
neuron projection development
establishment or maintenance of cell polarity
nuclear matrix
positive regulation of cytokine production
double-strand break repair
Ras GTPase binding
regulation of translation
chromatin binding
0 1 2 3 4


Credible SNPs (20,362)
Credible SNPs
CAVIAR SNPs
14329
6033
1514
FCTX histone states
-log10(P-value)
0
10
20
30
40
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
FB histone states
-log10(P-value)
0
5
10
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs (2,638)
-Frameshift variant
-Stop-gained variant
-Splice-donor variant
-NMD transcript variant
-Missense variant
SNPs on promoters (1,180)
-2kb upstream to
1kb downstream
of TSS
Remaining SNPs (16,544)
-Hi-C interactions
to 1Mb flanking regions
-Interacting genes
with FDR<0.01
221 genes
macromolecule methylation
protein alkylation
nuclear matrix
N-methyltransferase activity
S-adenosylmethionine-dependent methyltransferase activity
covalent chromatin modification
embryo development ending in birth
synaptic membrane
peptidyl-amino acid modification
cell cycle phase
0 2 4 6 8
Z-score
471 genes
adult behavior
histone methyltransferase activity
regulation of interferon-gamma production
histone methylation
receptor metabolic process
synaptic membrane
microtubule organizing center part
locomotory behavior
amine binding
cilium part
0 1 2 3 4 5 6
Z-score
GZ: 1,898 genes
acetylcholine receptor activity
activation of Ras GTPase activity
response to tropane
histone methylation
adult behavior
M phase of mitotic cell cycle
central nervous system neuron development
translation regulator activity
signal sequence binding
cerebral cortex development
0 1 2 3 4 5 6
Z-score
CP: 1,806 genes
activation of Ras GTPase activity
histone methylation
regulation of synaptic plasticity
response to retinoic acid
long-term memory
postsynaptic density
neuron projection development
glutamate receptor binding
kinesin complex
nuclear ubiquitin ligase complex
0 1 2 3 4 5 6
Z-score
GZ: 2,590 genes
acetylcholine receptor activity
histone methylation
core promoter proximal region DNA binding
cell cycle phase
translation regulator activity
signal sequence binding
activation of Ras GTPase activity
chromatin binding
adult behavior
response to tropane
0 1 2 3 4 5 6
Z-score
CP: 2,498 genes
postsynaptic membrane
histone methyltransferase activity
acetylcholine receptor activity
methylation
activation of Ras GTPase activity
regulation of synaptic plasticity
response to retinoic acid
translation regulator activity
nicotinic acetylcholine-gated receptor-channel complex
long-term memory
0 1 2 3 4 5 6
Z-score

## APPENDIX

## Content:

- Preprint of manuscript by Parikshak, N. N. et al. Global changes in patterning, splicing and primate specific lncRNAs in autism brain.
- Preprint of manuscript by Won, H. et al. Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development.

# Global changes in patterning, splicing and primate specific lncRNAs in autism brain

Neelroop N. Parikshak[1,2,]*, Vivek Swarup[1,]*, T. Grant Belgard[1,2, †,]*, Michael Gandal[1,2], Manuel Irimia[5,6], Virpi Leppa[1], Jennifer K. Lowe[1], Robert Johnson[7], Benjamin J. Blencowe[6], Steve Horvath[3-4], Daniel H. Geschwind[1-3]

1. Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
2. Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.
3. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
4. Department of Biostatistics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
5. EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, Barcelona 08003, Spain.
6. Donnelly Centre, University of Toronto, 160 College Street, Toronto, ON M5S 3E1, Canada; Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada.
7. NICHD Brain and Tissue Bank for Developmental Disorders, University of Maryland Medical School, Baltimore, Maryland 21201, USA.

*These authors contributed equally to this study.

†Current address: MRC Functional Genomics Unit, Department of Physiology, Anatomy & Genetics, University of Oxford, South Parks Road, Oxford, OX1 3PT, United Kingdom.

## Summary

We apply transcriptome-wide RNA sequencing in postmortem autism spectrum disorder (ASD) brain and controls and identify convergent alterations in the noncoding transcriptome, including primate specific lncRNA, and transcript splicing in ASD cerebral cortex, but not cerebellum. We characterize an attenuation of patterning between frontal and temporal cortex in ASD and identify *SOX5*, a transcription factor involved in cortical neuron fate specification, as a likely driver of this pattern. We further show that a genetically defined subtype of ASD, Duplication 15q Syndrome, shares the core transcriptomic signature of idiopathic ASD, indicating that observed molecular convergence in autism brain is the likely consequence of manifold genetic alterations. Using co-expression network analysis, we show that diverse forms of genetic risk for ASD affect convergent, independently replicated, biological pathways and provide an unprecedented resource for understanding the molecular alterations associated with ASD in humans.

Autism spectrum disorder (ASD) is a neurodevelopmental syndrome characterized by deficits in social communication and mental flexibility[1]. Genetic risk factors contribute substantially to ASD risk, and recent studies support the potential contribution of more than a thousand genes to ASD risk[2-4]. However, given the shared cognitive and behavioral features across the autism spectrum, one hypothesis is that diverse risk factors may converge on common molecular, cellular, and circuit level pathways to result in the shared phenotype[5,6]. Analysis of the transcriptome has been used to identify common molecular pathways in the cerebral cortex (CTX) from postmortem human brain tissue in individuals with ASD[7-11]. However, all transcriptomic studies in ASD to date have been limited to evaluating highly expressed mRNAs corresponding to protein coding genes. Moreover, most lack rigorous replication and do not assess gene expression patterns across brain regions.

We used rRNA-depleted RNA-seq (Methods) to evaluate transcriptomes from a large set of ASD and control (CTL) brain samples including neocortex (frontal and temporal) and cerebellum across 79 individuals (46 ASD, 33 CTL, 205 samples, Extended Data Fig. 1a-e, Supplementary Table 1). We first compared differential gene expression (DGE) between ASD and CTL individuals in CTX from a previously published[7] microarray study against new, independent gene expression profiles from RNA-seq to evaluate global reproducibility of DGE in ASD. We found a high degree of replication of DGE fold changes between the sample sets, despite evaluation on different gene expression platforms (fold changes at $P < 0.05$ in previously evaluated data correlate with new data with $R^2 = 0.60$, Extended Data Fig. 1f). We observed a much weaker overall signal and replication in cerebellum ($R^2 = 0.033$, Extended Data Fig. 1g). These analyses confirm the existence of a reproducible DGE signature in ASD CTX across different platforms and in independent samples.

We next combined samples from all individuals with idiopathic ASD into a covariate-matched "ASD Discovery Set" (Extended Data Fig. 1h) for CTX (106 samples, 26 ASD, 33 CTL individuals) and held out remaining samples for replication ("ASD Replication Set", Methods). For DGE analysis, we used a linear mixed effects model that accounts for biological and technical covariates (Methods) to identify 1156 genes differentially expressed in ASD CTX, 582 increased and 574 decreased (Benjamini-Hochberg FDR ≤ 0.05, Supplementary Table 2). Importantly, DGE analysis with additional covariates or different assumptions about the distribution of the data and test statistics yielded similar results (Extended Data Fig. 2a). Additionally this DGE signature clusters over two-thirds of ASD samples together and this clustering is not related to confounding factors such as cortical region, age, sex, and RNA quality (Figure 1a, Extended Data Fig. 2b). The most significantly down-regulated gene was *PVALB* (fold change = 0.53, FDR ≤ 0.05), a marker for GABAergic interneurons. *SST*, a marker for a different subpopulation of GABAergic interneurons, is also among the most downregulated (fold change = 0.61, FDR ≤ 0.05). Other down-regulated genes at FDR ≤ 0.05 include *NEUROD6*, involved in neuronal differentiation (fold change = 0.60), multiple ion channels, and *KDM5D*, a lysine demethylase (fold change = 0.66). In contrast, members of the complement cascade implicated in microglial-neuronal interactions (*C4A*, fold change = 1.94; *C1QB*, fold change = 1.65; both FDR ≤ 0.05) are upregulated in ASD CTX. Gene Ontology (GO) term enrichment analysis further supports the involvement of pathways implicated by these genes (Figure 1b), confirming previous findings[7]. Moreover, the upregulated set is enriched for astrocyte and microglia enriched genes, and the down-regulated set is enriched for synaptic genes (Extended Data Fig. 2c), consistent with previous observations[7,11].

We next sought to evaluate whether the transcriptional signature identified in the ASD Discovery Set generalizes to the ASD Replication set by assessing the 1st principal component of the DGE set, which summarizes the DGE expression pattern across all cortical samples. The ASD Discovery Set and ASD

Replication Set share this pattern, which is significantly different for both sets compared to CTL (Figure 1c). Moreover, this pattern is highly associated with ASD diagnosis, but not other biological factors, technical factors, or scores on sub-domains of an ASD diagnostic tool (Figure 1d). These analyses demonstrate that convergent differences in ASD CTX are reproducible in independent samples and are not related to confounding factors.

We also detected 2715 lncRNAs expressed in cerebral cortex (after careful filtering for high-confidence transcripts, Supplementary Information), of which 62 were significantly dysregulated between ASD and CTL (33 long intergenic RNAs, lincRNAs; 19 antisense transcripts; and 10 processed transcripts at FDR ≤ 0.05). Similar to the protein coding genes, these transcripts' expression patterns cluster ASD and CTL samples (Figure 1e). Most of these lncRNAs are developmentally regulated[12], have chromatin states indicative of transcription start sites (TSSs) near their 5′ end in brain[13], and are identified in other datasets[12,14] consistent with being valid, functional lncRNAs. Moreover, most (81%) exhibit primate-specific expression patterns in brain[15] (Supplementary Information). For example, Figure 1f depicts two lincRNAs, *LINC00693* and *LINC00689*, which are typically downregulated during development, yet are upregulated in ASD CTX relative to controls (Figure 1g), which we validated by RT-PCR (Extended Data Fig. 2d). *LINC00693* sequence is present, but poorly conserved in mouse, while *LINC00689* is primate-specific (present in macaque and other primates but not in any other species, Supplementary Information, Extended Data Fig. 3 for additional examples). These data indicate that dysregulation of lncRNAs, many of which are primate-specific and involved in brain development, is an important component of transcriptome dysregulation observed in ASD.

Previous work suggested that alterations in transcript splicing may contribute to transcriptomic changes in ASD[7,16,17] by evaluating splicing in a targeted manner and pooling samples across individuals[7,16,17]. Given the increased sequencing depth and reduced sequencing bias across transcript length in our dataset, we were able to perform an unbiased genome-wide analysis of differential alternative splicing (AS). We evaluated the percent spliced in (PSI, Extended Data Fig. 4a) for 34,025 AS events in CTX across the ASD Discovery Set, encompassing skipped exons (SE), alternative 5′ splice sites (A5SS), alternative 3′ splice sites (A3SS), and mutually exclusive exons (MXE) using the MATS pipeline[18] (Supplementary Information). We first asked whether there was a global signal, finding significant enrichment over background (Extended Data Fig. 4b). We identified 1127 events in 833 genes at FDR ≤ 0.5 in CTX (similar to the number of events at uncorrected P < 0.005). Importantly, we obtained similar results with a different splice junction mapping and quantification approach (Extended Data Fig. 4c).

We performed PCR validations with nine AS events from the differential splicing set (*ASTN2, MEF2D, ERC2, MED31, SMARCC2, SYNE1, NRCAM, GRIN1, NCAM*) and found that validated changes in splicing patterns were concordant with RNA-seq (Extended Data Fig. 4d-e), demonstrating that our approach identifies alterations in AS with high specificity. Similar to our observations with lncRNA and DGE, AS changes clustered the samples by diagnosis (Figure 2a). The most significantly different event was the inclusion of an exon in *ASTN2* (ΔPSI = 5.8 indicating a mean of 5.8% difference in inclusion in ASD vs CTL; $P = 7.8\text{x}10^{-6}$), a gene implicated by copy number variation (CNV) in ASD and other developmental disorders[19]. GO term analysis of the genes implicated by these pathways indicates involvement of biological processes related to neuronal projection, biological adhesion, and morphogenesis (Figure 2b), pathways where alternative isoforms are critical to specifying interactions between protein products. Moreover, the 1st principal component of the cortex differential splicing signature replicates in the ASD Replication Set and is not associated with other biological or technical factors (Figures 2c-d, Extended Data Fig. 5a). Importantly, many splicing alterations occur in genes that are not differentially expressed

between ASD and CTL; removing AS events on genes exhibiting even nominal DGE ($P < 0.05$), still identified a strong difference between ASD and CTL CTX  (Extended Data Fig. 5b).

A parallel analysis in cerebellum evaluating 32,954 AS events found no differentially regulated events significant at any multiple comparison correction thresholds (Extended Data Fig. 5c, Supplementary Table 3). There was no detectable global overlap between cerebellum and CTX above chance for events significant at $P < 0.05$ in both comparisons (fold enrichment = 1.1, P = 0.21). This suggests that AS alterations in ASD are largely confined to CTX cell types, consistent with the stronger overall DGE patterns observed in CTX versus cerebellum.

To further explore the underlying biology of AS dysregulation, we tested whether the shared splicing signature in ASD might be a product of perturbations in AS factors known to be important to neural development or preferentially expressed in neural tissue. We found that the expression levels of *RBFOX1*, *RBFOX2*, *SRRM4*, *NOVA1*, and *PTBP1* all had high correlations ($R^2 > 0.35$, FDR $\leq 0.05$) to AS alterations in CTX (Figure 2e), but not in cerebellum (Figure 2f). Furthermore, enrichment analysis revealed that most changes in cortical AS occur in neuron-specific exons that are excluded in ASD (exons with $\Delta$PSI > 50% in neurons overlap with exons excluded in ASD CTX, fold enrichment = 4.1, $P = 1.8x10^{-7}$, Extended Data Fig. 5d).

To validate a regulatory relationship between splicing factors and these events, we evaluated experimental data from knockout, overexpression, and knockdown experiments for Rbfox1[20], SRRM4[21], and PTBP1[22], respectively . We found that exons regulated by each of these splicing factors were significantly enriched in the set of exons excluded in ASD (Figure 2g), while in contrast, there was no enrichment for targets of ESRP[23], a splicing factor involved in epithelial cell differentiation but not expressed in CTX. This shows that alterations in three splicing factors dysregulated in ASD regulate AS of the neuron-specific exons whose inclusion is dysregulated in ASD in CTX and not cerebellum, indicating selective alteration of neuronal splicing in ASD CTX. Remarkably, the expression patterns of these three splicing factors (and others for which appropriate validation experiments were unavailable) results in distinct clusters (Extended Data Fig. 5e), suggesting that subsets of splicing factors act in different individuals to mediate a common downstream AS alteration.

Taken together these results indicate global transcriptional alterations in ASD cerebral cortex, but not cerebellum at the level of protein coding transcripts, lncRNA and AS. Therefore, to determine how these different transcriptomic subcategories relate to each other in ASD, we compared the 1st PC for each type of transcriptomic alteration across individuals (Figure 2h).  Remarkably, the PCs are highly correlated ($R^2 > 0.8$) indicating that the transcriptomic alteration is a unitary phenomenon across protein coding, noncoding, and splicing levels, rather than distinct forms of molecular alteration.

Previous analysis with gene expression microarrays in a small cohort suggested that the typical pattern of transcriptional differences between the frontal and temporal cortex may be attenuated in ASD[7]. To further test this possibility, we evaluated DGE between CTX regions (Supplementary Information) in 16 matched frontal and temporal CTX sample pairs from ASD and CTL subjects and found 551 genes differentially expressed between regions in controls, but only 51 in ASD (FDR $\leq 0.05$; Figure 3a). We refer to the set of 523 genes with this pattern in CTL, but not ASD as the “Attenuated Cortical Patterning” set. The attenuation of patterning is evident from the global distribution of test statistics between frontal and temporal CTX in ASD and CTL and genes in this set do not show a greater difference in variability in ASD versus controls compared to other genes (Kolmogorov-Smirnov test, two-tailed P = 0.11, Extended Data Fig. 6a).

We complemented this analysis with a machine learning approach using all 123 cortical samples, training a regularized regression model[24] to classify frontal versus temporal CTX with independent gene expression data from BrainSpan[25] (Extended Data Fig. 6b, Supplementary Information). Multiple approaches to training the classifier with BrainSpan can differentiate between frontal and temporal CTX in both CTL and ASD (Extended Data Fig. 6c-e), demonstrating that dissection and sample quality in our samples are of high quality. Loss of classification accuracy in ASD compared to CTL was observed when restricting the model to the genes with the most attenuated patterning in ASD (Extended Data Fig. 6f), demonstrating that attenuation of patterning generalizes across all samples. The Attenuated Cortical Patterning set includes multiple genes known to be involved in cell-cell communication and cortical patterning, such as *PCDH10*, *PCDH17*, *CDH12, MET*, and *PDGFD*, which was recently shown to mediate human specific aspects of cerebral cortical development[26]. GO term enrichment analysis of the Attenuated Cortical Patterning set identified enrichment for G protein coupled signaling, Wnt receptor signaling, and calcium binding, among several developmental processes (Figure 3b), and cell type enrichment analysis did not identify a strong preference for a particular cell type (Extended Data Fig. 6g).

To identify potential drivers of the alteration in cortical patterning, we evaluated transcription factor binding site enrichment upstream of genes in the Attenuated Cortical Patterning set (Supplementary Information), and found an enrichment of *SOX5* binding motifs (upstream of 364/523 genes, Figure 3c). Remarkably, *SOX5* itself belongs to the Attenuated Cortical Patterning set: while *SOX5* is differentially expressed between frontal and temporal CTX in CTL, it is not in ASD (Figure 3d). We thus predicted that if *SOX5* regulates cortically patterned genes, its expression should correlate with target gene transcript levels. Consistent with this prediction, we found that genes in the Attenuated Cortical Patterning set are anti-correlated with *SOX5* in CTL CTX, but not in ASD CTX (Figure 3e, top left; Wilcoxon rank sum test of R values, P = 0.01), suggesting that the normal role of SOX5 as a transcriptional repressor may be disrupted in ASD. We reasoned that a true loss of SOX5-mediated cortical patterning would be specific to the predicted SOX5 targets. Consistent with this, we find a loss of correlations between *SOX5* and predicted targets, but no difference in correlations between *SOX5* and non-targets in the Attenuated Cortical Patterning set (Figure 3e). Taken together, these findings show that a loss of regional patterning downstream of the transcriptional repressor *SOX5*, which plays a crucial role in glutamatergic neuron development[27,28], contributes to the loss of regional identity in ASD.

Gene expression changes in postmortem brain may be a consequence of genetic factors, environmental factors, or both. Brain tissue from individuals with ASD that harbor known, penetrant genetic causes are very rare. However, we were able to identify postmortem brain tissue from 8 subjects with one of the more common recurrent forms of ASD, Duplication 15q Syndrome (dup15q, which is present in about 0.5-1% of ASD cases, see Extended Data Fig. 7a for characterization of duplications). We performed RNA-seq across frontal and temporal cortex and compared DGE changes in dup15q with those observed in individuals with idiopathic ASD to better understand the extent to which the observed molecular pathology overlaps. As expected, most genes in the 15q11.1-13.2 duplicated region have higher expression in dup15q CTX compared to CTL (Figure 4a), although *SNRPN* and *SNURF* were notably downregulated. Conversely, no significant upregulation of genes in this region were identified in idiopathic ASD or controls. Strikingly, when we assessed genome-wide expression changes, we observed a strong signal of DGE in dup15q that widely overlaps with that of idiopathic ASD (fold changes at FDR $\leq 0.05$ in dup15q correlate with idiopathic ASD with $R^2 = 0.79$, Figure 4b). Moreover, the slope of the best-fit line through these changes is 2.0, indicating that on average, the transcriptional changes in dup15q CTX are highly similar, but twice the magnitude of those observed in ASD CTX.

Next, we sought to evaluate AS changes in dup15q. There is only one significant splicing change in the dup15q region (Supplementary Table 3), consistent with the idea that duplication in this region duplicates all isoforms of the genes, resulting in minimal alteration of transcript structure. Similar to DGE, global AS analysis in dup15q CTX vs to CTL CTX revealed a stronger, but highly overlapping signature with idiopathic ASD CTX (fold changes at FDR ≤ 0.2 in dup15q agree correlate with idiopathic ASD with $R^2 = 0.66$) indicating that splicing changes in dup15q syndrome recapitulate those of idiopathic ASD (Figure 4c). The slope of the best-fit line through the PSI for spliced exons in dup15q CTX compared to those in ASD CTX is 2.5 similar to DGE. Notably, both gene expression and AS changes in dup15q implicated similar pathways as those found in idiopathic ASD (Extended Data Fig. 7c-d). Clustering dup15q samples and CTL samples using both the DGE set and the differential AS set showed that all dup15q samples cluster together (Figure 4d), as opposed to the more variable clustering of idiopathic ASD, supporting the hypothesis that this shared genetic abnormality leads to a more homogeneous molecular phenotype.

Next, to test whether this molecular ASD signature may be due to independent of postmortem or reactive effects (Supplementary Information), we compared our data with gene expression profiles from a iPSC-derived neurons (nIPSCs)[29] from dup15q were available, we could use these data to definitively reveal which changes in dup15q CTX are independent of postmortem or reactive effects (Supplementary Information), since such effects are not present *in vitro*. We observe that DGE in the 15q region is concordant with that seen in the nIPSCs (Figure 4e), even though the sample size is small and the analysis is likely underpowered. Upregulated changes in dup15q are also seen in nIPSCs (Figure 4f), consistent with our other statistical analyses showing limited effects of potential confounders. The very immature, fetal state of the nIPSCs[30] likely explains the absence of an enrichment signal for genes downregulated in postnatal ASD brain, which are enriched for genes involved in neurons with more mature synapses.

We next applied gene network analysis to construct an organizing framework to understand shared biological functions across idiopathic ASD and dup15q (combining the ASD Discovery Set, ASD Replication Set, and dup15q set). We utilized Weighted Gene Co-expression Network Analysis (WGCNA), which identifies groups of genes with shared expression patterns across samples (modules) from which shared biological function is inferred. Modules identified via WGCNA can than be related to a range of relevant phenotypes and potential confounders[31,32]. We applied signed co-expression analysis and used bootstrapping to ensure the network was robust, and not dependent on any subset of samples (Supplementary Information), while controlling for technical factors and RNA quality ("Adjusted FPKM" levels, Methods). WGCNA identified 16 co-expression modules (Extended Data Fig. 8a, Supplementary Table 2), which are further characterized by their association to ASD (Extended Data Fig. 8b), enrichment for cell-type specific genes (Extended Data Fig. 8c), and enrichment for GO terms (Extended Data Fig. 9). Of the downregulated modules, three are associated with ASD and dup15q (M1/10/17) and one with dup15q only (M11). Five of the upregulated modules are associated with ASD and dup15q (M4/5/6/9/12) and one is specific to dup15q (M13) (Figure 5a, top). Additionally, we identified a module strongly enriched for genes from the Attenuated Cortical Patterning set and *Wnt* signaling that contains *SOX5* (M12; fold enrichment = 3.0, $P = 3x10^{-8}$), verifying the strong relationship observed between the *Wnt* pathway regulating TF *SOX5* and attenuation of cortical patterning[33].

Notably, the modules identified here significantly overlap with previous patterns identified in ASD (asdM12$_{array}$ and asdM16$_{array}$[7]; Figure 5a, middle). We found that the ASD-associated modules identified by our larger sample size and RNA-seq provide significant refinement of previous observations by identifying more discrete biological processes related to cortical development[34], the post-synaptic density[35], and

lncRNAs (Figure 5a, bottom). For example, M1 overlaps a subset of asdM12$_{array}$ (fold enrichment = 5.7) and developmental modules (devM16 fold enrichment = 3.7), and is enriched for proteins found in the PSD and genes involved in calcium signaling and gated ion channel signaling. Another subset of asdM12$_{array}$, M10 (fold enrichment = 11) overlaps more with a mid-fetal upregulated cortical development module (devM13 fold enrichment = 4.0), and genes involved in secretory pathways and intracellular signaling. A third module, M17 shows the least overlap with asdM12$_{array}$ (fold enrichment = 2.2) and is related to energy metabolism. Notably, these three modules are enriched for neuron-specific genes (Extended Data Fig. 8c), but not all neuronal modules are down regulated in ASD (M3 is not altered in ASD CTX). Taken together, specific neurobiological processes are affected in individuals with ASD related to developmentally regulated neurodevelopmental processes.

The most upregulated modules, M5 and M9, both strongly overlap (fold enrichments > 20) with previously identified upregulated co-expression module asdM16$_{array}$. M5 is enriched for microglial cell markers and immune response pathways, whereas M9 is enriched for astrocyte markers and immune-mediated signaling and immune cell activation (Extended Data Fig. 8c, Extended Data Fig. 9). This analysis clearly separates the contributions of the coordinated biological processes of microglial activation and reactive astrocytosis, which were previously not distinguishable as separate modules[7]. Thus, our analysis pinpoints more specific biological pathways in idiopathic ASD than those previously identified and reveals that similar changes occur downstream of the genetic perturbation in dup15q.

We evaluated the relationship between the five modules most strongly associated with ASD (M1/5/9/10/17, which are supported by module-trait association analysis and gene set enrichment analysis, Supplementary Information), and found that there was a remarkably high anti-correlation between the eigengene of M5 and downregulated modules, particularly M1 ($R^2 = 0.76$) (Figure 5b). M1 (Figure 5c) is downregulated in ASD and enriched for genes at the PSD and genes involved in synaptic transmission, while M5 (Figure 5d) is enriched for microglial genes and cytokine activation. This strong anti-correlation between microglial signaling and synaptic signaling in ASD and dup15q provides evidence in humans for dysregulation of microglia-mediated synaptic pruning, as previously suggested[36].

Next, to determine the role of causal genetic variation, we evaluated enrichment of both rare genetic variants, focusing on genes affected by ASD associated gene disrupting (LGD) *de novo* mutations[37], and common variants[38,39]. Genes within three modules, M1, M3, and M12, show enrichment for common variation signal for ASD (Figure 5e, Methods). Remarkably, M12 (Figure 5f), which is related to cortical patterning and Wnt signaling, also exhibit GWAS signal enrichment, providing the first evidence that risk conferred by common variation in ASD may affect regionalization of the cortex. Interestingly, M3 is significantly enriched for both schizophrenia (SCZ) and ASD common variants, is related to synaptic transmission, nervous system development, and regulation of ion channel activity (Extended Data Fig. 9), consistent with the notion that ASD and SCZ share common and rare genetic risk[1,40-43].

We only identified one module, M2 (Figure 5g), as significantly enriched in protein disrupting (nonsense, splice site, or frameshift) rare *de novo* variants previously associated with SCZ and ASD. M2 overlaps with a cortical developmental module implicated in ASD[34] (devM2 fold enrichment = 5.1). Notably, M2 is not differential between ASD and CTL in our dataset, consistent with the observation that these genes are primarily expressed during early neuronal development in fetal brain[34]. Remarkably, M2 contains an unusually large fraction of lncRNAs (15% of the genes in M2 are classified as lncRNAs, while other modules are 1-5% lncRNA). We hypothesize that, in addition to protein coding genes involved in transcriptional and chromatin regulation, rare *de novo* variants may also affect lncRNAs in ASD, a prediction that will be testable once large sets of whole genome sequences are available.

These combined transcriptomic and genetic analyses reveal that different forms of genetic variation affect biological processes involved in multiple stages of cortical development. Common genetic risk is enriched in M3, M1, and M12, which reflect early glutamatergic neurogenesis, later neuronal function, and cortical patterning, respectively. We also observe that rare *de novo* variation, which is enriched in M2, affects distinct biology related to transcriptional regulation and chromatin modification. These findings are consistent with transcriptomic analyses of early prenatal brain development and ASD risk mutations that implicate chromatin regulation and glutamatergic neuron development[34,44].

We provide the first comprehensive picture of largely unexplored aspects of transcription in ASD, lncRNA and alternative splicing, and identify a strong convergent signal in these, as well as protein coding genes[7]. These results will aid in interpreting genetic variation outside of the known exome, as whole genome sequencing supplants current methods. A role of lncRNAs has been previously explored in ASD[45], but only two individuals were evaluated with targeted microarrays. We evaluate lncRNAs in an unbiased manner across many individuals, notably identifying an enrichment of lncRNAs in M2, most of which are uncharacterized in brain and arose on the primate lineage. The involvement of lncRNAs in this early developmental program that is enriched for *de novo* mutations implicated in ASD suggests their study will be particularly relevant to understanding the emergence of primate higher cognition on the mammalian lineage, and by extension human brain evolution[15,46,47].

We also provide the first confirmation of an attenuation of genes that typically show differential expression between frontal and temporal lobe in ASD CTX and further identified *SOX5*, known to regulate cortical laminar development[50,51], as a putative regulator of this disruption. That M12, which is enriched for genes exhibiting cortical regionalization and is also enriched in ASD GWAS signal, supports the prediction that attenuation of patterning may be mediated by common genetic variation in or near the *SOX5* target genes. Disruption of cortical lamination by direct effects on glutamatergic neurogenesis and function has been predicted by independent data, including network analyses of rare ASD associated variants identified in exome sequencing studies[34,44].

These data, in conjunction with previous studies, reveal a consistent picture of the ASD's emerging postnatal and adult pathology. Specific neuronal signaling and synaptic molecules are downregulated and astrocyte and microglial genes are upregulated in over 2/3 of cases. Microglial infiltration has been observed in ASD cortex with independent methods[52], and normal microglial pruning has been shown to be necessary for brain development[36]. Our findings further suggest that aberrant microglial-neuronal interactions may be pervasive in ASD and related to the gene expression signature seen in a majority of individuals. In our comprehensive AS analysis, we identify three splicing factors upstream of the altered splicing signature observed in ASD CTX. These factors are known to be involved in coordinating sequential processes in neuronal development[17,21] and maintaining neuronal function[48,49]. It may therefore be sufficient to disrupt any one of these factors to induce a similar outcome during brain development, which would be consistent with the shared downstream perturbation observed here.

Finally, evaluation of the transcriptome in dup15q supports the enormous value of the "genotype first" approach of studying syndromic forms of ASD, with known penetrant genetic lesions[53]. It is highly unlikely that the shared transcriptional dysregulation in dup15q is due to a shared environmental insult. Thus, the most parsimonious explanation for the convergent transcriptomic pathology seen in all dup15q and over 2/3 of the cases of idiopathic ASD is that it represents an adaptive or maladaptive response to a primary genetic insult, which in most cases of ASD will be genetic[2,54]. As future investigations pursue the full range of causal genetic variation contributing to ASD risk, these analyses and data will be valuable for interpreting genetic and epigenetic studies of ASD as well as those of other neuropsychiatric disorders.

Figure Legends

Figure 1 | Transcriptome-wide differential gene expression in ASD. a, Average linkage hierarchical clustering of samples in the ASD Discovery Set using the top 100 upregulated and top 100 downregulated protein coding genes. b, Gene Ontology (GO) term enrichment analysis of upregulated and downregulated genes in ASD. *FDR ≤ 0.05 across all GO terms and gene sets. c, 1st principal component of the CTX DGE set (CTX DGE PC1) is able to distinguish ASD and CTL samples, including independent samples from the ASD Replication Set. d, CTX DGE PC1 is primarily associated with diagnosis, and not other factors. e, Average linkage hierarchical clustering of ASD Discovery Set using all lncRNAs in the DGE set. f, UCSC genome browser track displaying reads per million (RPM) in a representative ASD and CTL sample, superimposed over the gene models and sequence conservation for genomic regions including *LINC00693* and *LINC00689*. g, *LINC00693* and *LINC00689* are upregulated across ASD samples and downregulated during frontal cortex development. Abbreviations: FC, frontal cortex; TC, temporal cortex; RIN, RNA integrity number; ADI-R score, Autism Diagnostic Interview Revised score; FPKM, fragments per kilobase million mapped reads.

Figure 2 | Alteration of alternative splicing in ASD. a, Average linkage hierarchical clustering of ASD discovery set using top 100 differentially included and top 100 differentially excluded exons from the differential splicing (DS) set across the ASD Discovery Set. b, Gene Ontology term enrichment analysis of genes with DS in ASD. c, 1st principal component 1 of the CTX differential alternative splicing set (CTX DS PC1) is able to distinguish ASD and CTL samples using independent samples from the ASD Replication Set. d, CTX DS PC1 is primarily associated with diagnosis, and not other factors. e, Correlation between CTX DS PC1 and gene expression of neuronal splicing factors in CTX. f, Correlation between 1st principal component of cerebellum differential splicing (CB DS PC1) and gene expression of neuronal splicing factors in cerebellum. g, Overlap between DS set and splicing events regulated by splicing factors where experimental data was available. h, Scatterplots and correlations between the 1st principal component across the ASD versus CTL DGE sets for different transcriptome subcategories. Abbreviations: FC, frontal cortex; TC, temporal cortex; RIN, RNA integrity number; ADI-R score, Autism Diagnostic Interview Revised score; FPKM, fragments per kilobase million mapped reads.

Figure 3 | Attenuation of cortical patterning in ASD cortex. a, Heatmap of 551 genes exhibiting cortical patterning between frontal cortex (FC) and temporal cortex (TC) in ASD, with samples sorted by diagnostic status and brain region. b, Gene ontology term enrichment analysis of genes exhibiting attenuated cortical patterning (ACP). c, Schematic of transcription factor motif enrichment upstream of genes in the ACP set, with the *SOX5* motif sequence logo. d, The *SOX5* gene exhibits attenuated cortical patterning in ASD CTX compared to CTLs. Lines connect FC-TC pairs that are from the same individual. e, Correlation between *SOX5* gene expression and predicted targets in CTL and ASD, with all ACP genes (top left), SOX5 targets from the ACP set (top right), SOX5 non-targets from the ACP set (bottom left), and all genes not in the ACP set (bottom right). Plots show the difference in correlation between *SOX5* and other genes in ASD and CTL (ΔR).

Figure 4 | Duplication 15q Syndrome recapitulates transcriptomic changes in idiopathic ASD. a, DGE changes across the 15q11-13.2 region for ASD and dup15q compared to CTL, error bars are +/- 95% confidence intervals for the fold changes. b, Comparison of effect sizes in dup15q vs CTL and ASD vs

CTL, with changes in dup15q at FDR ≤ 0.05 highlighted. c, Comparison of differential splicing (DS) changes in dup15q vs CTL and ASD vs CTL, highlighting 402 events at FDR ≤ 0.2 in dup15q. d, Average linkage hierarchical clustering of dup15q samples and controls using the DGE and DS gene sets. e, Plot of fold changes between induced pluripotent stem cells differentiated into neurons (nIPSCs) from dup15q vs CTL and postmortem CTX DGE from dup15q vs CTL in the 15q region. f,  Heatmap overlapping the top 1000 genes up- and down- regulated in the nIPSC comparison to the up- and down- regulated genes in dup15q and idiopathic ASD CTX.

Figure 5 | Co-expression network analysis across all ASD and CTL samples in CTX. a, Gene set enrichment analyses comparing the 16 co-expression modules with multiple gene sets from this RNA-seq study, from postmortem ASD CTX microarray, from human brain development, from the postsynaptic density and set of all brain-expressed lncRNAs. b, Comparison of five ASD-associated modules against each other by correlating module eigengenes. c, Module plot of M1 displaying the top 25 hub genes along with the module's Gene Ontology term enrichment. d, similar to c, but for M5. e, Gene set enrichment analysis with genome-wide whole-exome sequencing data (Rare *de novo* hit genes) and genome-wide association study (GWAS) results in ASD, schizophrenia (SCZ), and intellectual disability (ID). Boxes are filled if the odds ratio is greater than 0, and the enrichment $P < 0.05$. Asterisks* indicate FDR ≤ 0.05 across all comparisons in a and e. f,g, similar to c, but for M12 and M2, respectively. Abbreviations: LGD, likely gene disrupting, genes affected by nonsense, nonsynonymous, or splice-site mutations or frame-shift indels; AGRE, AGP/CHOP, and PGC refer to consortia that collect genetic data (Supplementary Information for details).

## Methods

Sample description: Brain tissue for ASD and control individuals was acquired from the Autism Tissue Program (ATP) brain bank at the Harvard Brain and Tissue Bank and the University of Maryland Brain and Tissue Bank (a Brain and Tissue Repository of the NIH NeuroBioBank). Sample acquisition protocols were followed for each brain bank, and samples were de-identified prior to acquisition. Brain sample and individual level metadata is available in Supplementary Table 1.

RNA-seq methodology: Starting with 1ug of total RNA, samples were rRNA depleted (RiboZero Gold, Illumina) and libraries were prepared using the TruSeq v2 kit (Illumina) to construct unstranded libraries with a mean fragment size of 150bp (range 100-300bp) that underwent 50bp paired end sequencing on an Illumina HiSeq 2000 or 2500 machine. Paired-end reads were mapped to hg19 using Gencode v18 annotations[55] via Tophat2[56]. Gene expression levels were quantified using union exon models with HTSeq[57]. For additional and information on sequencing and read alignment parameters, please see Supplementary Information.

Sample sets for analysis: For differential gene expression and splicing analysis, we defined an age matched set, referred to as the ASD Discovery Set (106 samples in CTX, 51 in cerebellum) of idiopathic ASD and control samples for the discovery set, and held out younger or unmatched samples as the ASD Discovery Set (17 in CTX, 8 in cerebellum). Dup15q individuals were analysed separately, utilizing the full set of controls from the ASD Discovery Set. For co-expression network analysis, we combined the discovery set, replication set, and dup15q individuals for a total of 137 CTX samples and 59 cerebellum samples.

Differential Gene Expression (DGE): DGE analysis was performed with expression levels adjusted for gene length, library size, and G+C content (referred to as "Normalized FPKM") Supplementary Information.

CTX samples (frontal and temporal) were analyzed separately from cerebellum samples. A linear mixed effects model framework was used to assess differential expression in log2(Normalized FPKM) values for each gene for cortical regions (as multiple brain regions were available from the same individuals) and a linear model was used for cerebellum (where one brain region was available in each individual, with a handful of technical replicates removed). Individual brain ID was treated as a random effect, while age, sex, brain region (except in the case of cerebellum, where there is only one region), and diagnoses were treated as fixed effects. We also used technical covariates accounting for RNA quality, library preparation, and batch effects as fixed effects into this model (Supplementary Information).

Reproducibility analyses: We assessed replication between datasets by evaluating the concordance between independent sample sets by comparing the squared correlation ($R^2$) of fold changes of genes in each sample set at a non-stringent P value threshold. This general approach has been shown to be effective for identifying reproducible gene expression patterns[58], and we modify it such that the P value threshold is set in one sample set (the *x* axis in the scatterplots), and the $R^2$ with fold changes in these genes are evaluated in an independent sample set (the *y* axis in the scatterplots).

Differential Splicing Analysis: Alternative splicing was quantified using the percent spliced in (PSI) metric using Multivariate Analysis of Transcript Splicing (MATS, v3.08)[18]. For each event, MATS reports counts supporting the inclusion (I) or exclusion (E) of a splicing event. To reduce spurious events due to low counts, we required at least 80% of samples to have I + S >= 10. For these events, the percent spliced in is calculated as PSI = I / (I + S) (Extended Data Fig. 4a). Statistical analysis for differential splicing was performed utilizing the linear mixed effects model regression framework as described above for DGE. This approach is advantageous over existing methods as it allows modeling of covariates and takes into consideration the variability in PSI across samples when assessing event significance with ASD (Supplementary Information).

Genotyping dup15q: For Dup15q samples, the type of duplication and copy number in the breakpoint 2-3 region were available for these brains[59]. To expand this to the regions between each of the recurrent breakpoint in these samples, 7/8 dup15q brains were genotyped (one was not genotyped due to limitations in tissue availability). The number of copies between each of the breakpoints is reported in Extended Data Fig. 7a.

Co-expression network analysis: The R package weighted gene co-expression network analysis (WGCNA) was used to construct co-expression networks using the technical variation normalized data[31,60] (referred to as “Adjusted FPKM”). We used the biweight midcorrelation to assess correlations between log2(Normalized FPKM) and parameters for network analysis are described in Supplementary Information. Notably, we utilized a modified version of WGCNA that involves bootstrapping the underlying dataset 100 times and constructing 100 networks. The consensus of these networks (50th percentile across all edges) was then used as the final network [32], ensuring that a handful of samples do not determine the network structure. For module-trait analyses, 1st principal component of each module (eigengene) was related to ASD diagnosis, age, sex, and brain region in a linear mixed effects framework as above, only replacing the expression values of each gene with the eigengene.

Enrichment analysis of gene sets and GWAS: Enrichment analyses were performed either with Fisher’s exact test (cell type and splicing factor enrichments) or logistic regression (all enrichment analyses in Figure 5). We used logistic regression in the latter case to control for gene length or other biases that may influence enrichment analysis (Supplementary Information). All GO term enrichment analysis was performed using GO Elite[61] with 10,000 permutations. We focused on molecular function and biological process terms for display purposes.

## Extended Data Figure Legends

Extended Data Figure 1 | Methodology, quality control, and differential expression replication analysis. a, RNA-seq workflow, including RNA extraction, library preparation, sequencing, read alignment, and quality control. b, RNA-seq quality and alignment statistics from this study, including RNA integrity number (RIN), number of aligned reads, proportion of reads mapping to different genomic features (mRNA, intronic, intergenic), and bias in coverage from the 5' to the 3' end of the top 1000 expressed transcripts (statistics compiled using PicardTools). c, Similar statistics as in b for another RNA-seq study that utilized polyA tail selection mRNA-seq to evaluate the transcriptome in ASD cortex[11] (primarily BA19, visual cortex, but also including some BA10/44 samples, frontal cortex). d, RNA-seq read coverage relative to normalized gene length across transcripts from the 5' to the 3' end in this study. e, Dependence between coverage and RIN across gene body (correlation between RIN and coverage in d across samples). f, Correlation of ASD vs CTL fold changes between previously evaluated and new ASD samples in CTX by microarray (left) and RNA-seq (right) using genes that were at $P < 0.05$ the samples from Voineagu et al., 2011. g, Correlation between effect sizes as in f, but for cerebellum (CB) samples. h,i, Correlation between covariates and ASD vs CTL status in CTX (h) and CB (i) in the ASD Discovery Set.

Extended Data Figure 2 | Transcriptome-wide differential gene expression (DGE) analysis in CTX. a, Comparison of P value rankings across different methods for DGE with Spearman's correlation. From left to right: removal of three additional principal components of sequencing statistics (Supplementary Information) related to RNA-sequencing quality, application of a permutation analysis for DGE P value computation, application of variance-weighted linear regression for DGE[62], and using surrogate variable analysis for DGE[63]. b, Average linkage hierarchical clustering heatmap using all genes DGE in the ASD Discovery Set, but including all idiopathic ASD frontal cortex (FC) and temporal cortex (TC) samples across 123 samples, combining the ASD Discovery set and the ASD Replication set. Bolded samples in the dendrogram are used for validation in d. c, Enrichment analysis of cell-type specific gene sets (5-fold enriched in the cell type compared to all other cells) with genes decreased and increased in ASD. d, RT-PCR validation of the two lincRNAs shown in Figure 1f-g, P values are computed with the Wilcoxon rank-sum test.

Extended Data Figure 3 | Gene browser tracks for selected primate-specific lncRNAs. For each lncRNA, expression for representative samples for ASD vs CTL (top) in human, macaque (middle), and mouse (bottom) are shown. The genome location for macaque and mouse displayed is syntenic to the human region, with the expected location of the lncRNA highlighted.

Extended Data Figure 4 | Splicing analyses and validation in ASD. a, Schematic describing how the percent spliced in (PSI) metric is computed. b, Distribution of *P* values for changes in the PSI between ASD and CTL in CTX for all events (left) and event subtypes (SE, spiced exon; A5SS, alternative 5' splice site; A3SS, alternative 3' splice site; MXE, mutually exclusive exons). c, Comparison of the CTX splicing analyses in when using PSI values obtained via read alignment by TopHat2[64] followed by the MATS[18] pipeline (used throughout this study) against read alignment by OLego followed by Quantas[65]. d, Comparison of ΔPSI values in nine splicing events between PCR and RNA-seq. e, PCR validation and sashimi plots for the nine splicing events delineated in d, from the samples highlighted in Extended Data Fig. 5a.

Extended Data Figure 5 | Additional splicing analyses in ASD. a, Average linkage hierarchical clustering heatmap using all differentially spiced (DS) events from the ASD Discovery Set, but including all idiopathic ASD neocortical samples (FC and TC) across 123 samples, combining the ASD Discovery set and the ASD Replication set. Bolded samples in the dendrogram were used for PCR validation in Extended Data Fig. 4. b, Top: difference between ASD and CTL in the DS set based on PC1 of the DS set at the PSI level, and PC1 of the gene expression levels of genes in the DS set. Bottom: Same comparison after differentially expressed genes ($p < 0.05$) are removed. c, Distribution of P values for changes in the PSI between ASD and CTL in cerebellum. d, Cell-type enrichment analysis of splicing events from CTX. e, Average-linkage hierarchical clustering using 1-(Pearson's correlation) to compare the gene expression patterns of the splicing factors investigated in Figure 2.

Extended Data Figure 6 | Attenuation of cortical patterning in ASD. a, Histograms of P values from paired Wilcoxon rank-sum test differential gene expression between 16 frontal cortex (FC) and 16 temporal cortex (TC) in CTL and ASD and a histogram of Bartlett's test P values for differences in gene expression variance between ASD and CTL for all genes (white) and genes in the Attenuated Cortical Patterning (ACP) set (red). c, Approach to training the elastic net model on BrainSpan and application of the model on 123 cortical samples in this study. c-e, Results of learned cortical region classifications with different starting gene sets, with the BrainSpan training set (left), CTL samples (middle), and ASD samples (right) in each panel and the Wilcoxon rank-sum test P value of FC vs TC difference for each comparison. f, Summary of results form c-e. g, Cell type enrichment analysis for genes in the ACP set. Abbreviations: A1C, primary auditory cortex; DFC, dorsolateral prefrontal cortex; MFC, medial prefrontal cortex; STC, superior temporal cortex; FC, frontal cortex; TC, temporal cortex; AUROC, area under the receiver-operator characteristic curve.

Extended Data Figure 7 | Dup15q syndrome analyses. a, Copy number between breakpoints (BP) in the 15q region. Genome-wide CNV analysis allowed evaluation of copy number in additional regions from previous studies[59,66]. b, Differential expression across the 15q region of interest in dup15q vs CTL and ASD vs CTL cerebellum, note only 3 samples were available for dup15q cerebellum so additional analyses were not pursued. c, Gene Ontology term enrichment analysis for the dup15q CTX differential expression set. d, Gene Ontology term enrichment analysis for the dup15q CTX differential splicing (DS) set. e, Hierarchical clustering of iPSC-derived neurons from dup15q, Angelman syndrome, and a control[29].

Extended Data Figure 8 | Co-expression network analysis in ASD CTX. a, Modules identified from a dendrogram constructed from a consensus of 100 bootstrapped datasets using the 137 CTX samples. Correlations for each gene to each measured factor are delineated below the dendrogram (blue = negative, red = positive correlation). b, Module-trait associations as computed by a linear mixed effects model with all factors on the x-axis used as covariates. All P values are displayed where the coefficient passed $p < 0.01$. Note that this alternative approach to module-trait association agrees with the Fisher's exact test used in Figure 5a when the fold enrichment for module overlap with DGE sets is > 2.8, and we use an intersection of both methods for the modules focused on in Figure 5b. c, Module enrichments for cell type specific gene expression patterns.

Extended Data Figure 9 | GO term enrichments for all modules. *FDR  < 0.05 across all GO enrichments across all modules.

## References

1. Geschwind, D. H. Genetics of autism spectrum disorders. *Trends Cogn. Sci. (Regul. Ed.)* 15, 409–416 (2011).
2. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* 46, 881–885 (2014).
3. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* (2015). doi:10.1016/S1474-4422(15)00044-7
4. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* 17, 782–790 (2014).
5. Chen, J. A., Peñagarikano, O., Belgard, T. G., Swarup, V. & Geschwind, D. H. The emerging picture of autism spectrum disorder: genetics and pathology. *Annu Rev Pathol* 10, 111–144 (2015).
6. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 9, 341–355 (2008).
7. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384 (2011).
8. Purcell, A. E., Jeon, O. H., Zimmerman, A. W., Blue, M. E. & Pevsner, J. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* 57, 1618–1628 (2001).
9. Garbett, K. *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiology of Disease* 30, 303–311 (2008).
10. Chow, M. L. *et al.* Age-Dependent Brain Gene Expression and Copy Number Anomalies in Autism Suggest Distinct Pathological Processes at Young Versus Mature Ages. *PLoS Genet.* 8, e1002592 (2012).
11. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Comms* 5, 5748 (2014).
12. Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Publishing Group* 18, 154–161 (2015).
13. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).
14. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* 47, 199–208 (2015).
15. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640 (2014).
16. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Reports* 6, 1139–1152 (2014).
17. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159, 1511–1523 (2014).
18. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* 40, e61–e61 (2012).
19. Lionel, A. C. *et al.* Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. *Human Molecular Genetics* 23, 2752–2768 (2014).
20. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* 20, 1434–1442 (2013).
21. Raj, B. *et al.* A Global Regulatory Mechanism for Activating an Exon Network Required for Neurogenesis. *Molecular Cell* 56, 90–103 (2014).
22. Gueroussov, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* 349, 868–873 (2015).
23. Dittmar, K. A. *et al.* Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Molecular and Cellular Biology* 32, 1468–1482 (2012).

24. Tibshirani, R., Johnstone, I., Hastie, T. & Efron, B. Least angle regression. *The Annals of Statistics* 32, 407–499 (2004).

25. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* 41, D996–D1008 (2013).

26. Lui, J. H. *et al.* Radial glia require PDGFD–PDGFRβ signalling in human but not mouse neocortex. *Nature* 515, 264–268 (2014).

27. Lai, T. *et al.* SOX5 Controls the Sequential Generation of Distinct Corticofugal Neuron Subtypes. *Neuron* 57, 232–247 (2008).

28. Kwan, K. Y. *et al.* SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16021–16026 (2008).

29. Germain, N. D. *et al.* Gene expression analysis of human induced pluripotent stem cell-derived neurons carrying copy number variants of chromosome 15q11-q13.1. *Mol Autism* 5, 44 (2014).

30. Stein, J. L. *et al.* A Quantitative Framework to Evaluate Modeling of Cortical Development by Neural Stem Cells. *Neuron* 83, 69–86 (2014).

31. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* 4, Article17 (2005).

32. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 1, 54 (2007).

33. Morales, P. L. M., Quiroga, A. C., Barbas, J. A. & Morales, A. V. SOX5 controls cell cycle progression in neural progenitors by interfering with the WNT–β-catenin pathway. *EMBO reports* 11, 466–472 (2010).

34. Parikshak, N. N. *et al.* Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* 155, 1008–1021 (2013).

35. Bayés, À. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* 14, 19–21 (2010).

36. Schafer, D. P. *et al.* Microglia Sculpt Postnatal Neural Circuits in an Activity and Complement-Dependent Manner. *Neuron* 74, 691–705 (2012).

37. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221 (2014).

38. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Human Molecular Genetics* 21, 4781–4792 (2012).

39. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. 459, 528–533 (2009).

40. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* 45, 984–994 (2013).

41. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506, 185–190 (2014).

42. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res* 25, 142–154 (2015).

43. Gilman, S. R. *et al.* Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nat. Neurosci.* 15, 1723–1728 (2012).

44. Willsey, A. J. *et al.* Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell* 155, 997–1007 (2013).

45. Ziats, M. N. & Rennert, O. M. Aberrant Expression of Long Noncoding RNAs in Autistic Brain. *J Mol Neurosci* 49, 589–593 (2012).

46. Geschwind, D. H. & Rakic, P. Cortical Evolution: Judge the Brain by Its Cover. *Neuron* 80, 633–647 (2013).

47. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated Recruitment of New Brain Development Genes into the Human Genome. *PLoS Biol* 9, e1001179 (2011).

48. Fogel, B. L. *et al.* RBFOX1 regulates both splicing and transcriptional networks in human neuronal

development. *Human Molecular Genetics* 21, 4171–4186 (2012).

49. Gehman, L. T. *et al.* The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nat Genet* 43, 706–711 (2011).

50. Greig, L. C., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat Rev Neurosci* 14, 755–769 (2013).

51. Srinivasan, K. *et al.* A network of genetic repression and derepression specifies projection fates in the developing neocortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19071–19078 (2012).

52. Morgan, J. T. *et al.* Abnormal microglial–neuronal spatial organization in the dorsolateral prefrontal cortex in autism. *Brain Research* 1456, 72–81 (2012).

53. Stessman, H. A., Bernier, R. & Eichler, E. E. A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell* 156, 872–877 (2014).

54. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet* 15, 133–141 (2014).

55. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7, S4–9 (2006).

56. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31, 46–53 (2012).

57. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).

58. Shi, L. *et al.* The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics* 9, S10 (2008).

59. Scoles, H. A., Urraca, N., Chadwick, S. W., Reiter, L. T. & LaSalle, J. M. Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. *Mol Autism* 2, 19 (2011).

60. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559 (2008).

61. Zambon, A. C. *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* 28, 2209–2210 (2012).

62. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15, R29 (2014).

63. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* 3, e161 (2007).

64. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578 (2012).

65. Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res* 41, 5149–5163 (2013).

66. Wintle, R. F. *et al.* A genotype resource for postmortem brain samples from the Autism Tissue Program. *Autism Res* 4, 89–97 (2011).

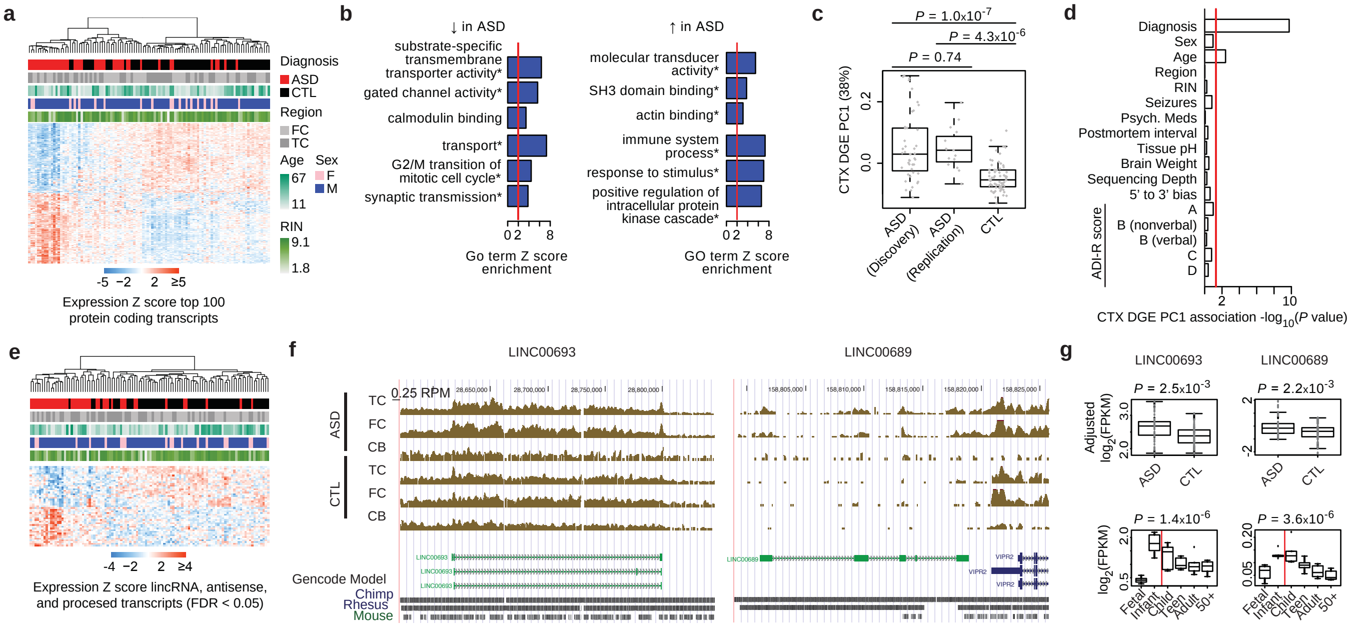
End Notes

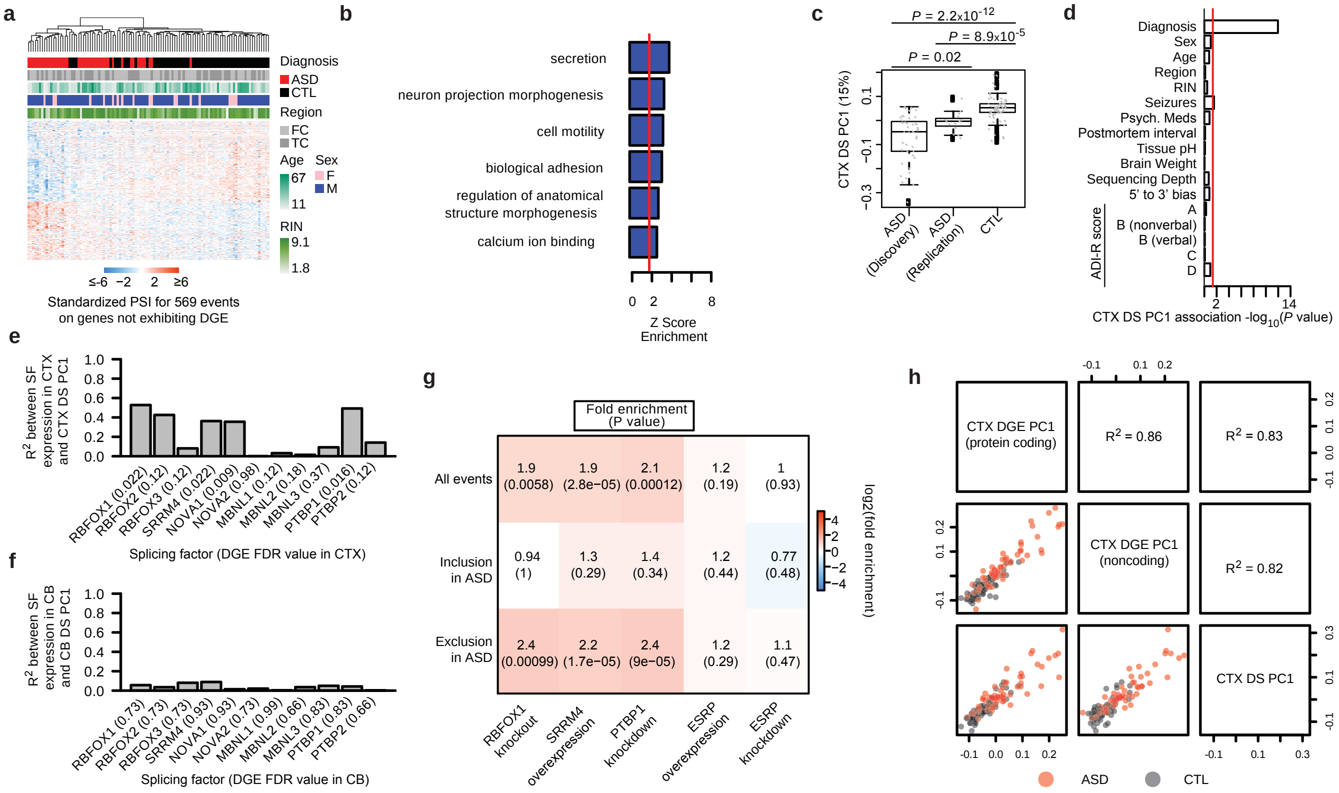
None.

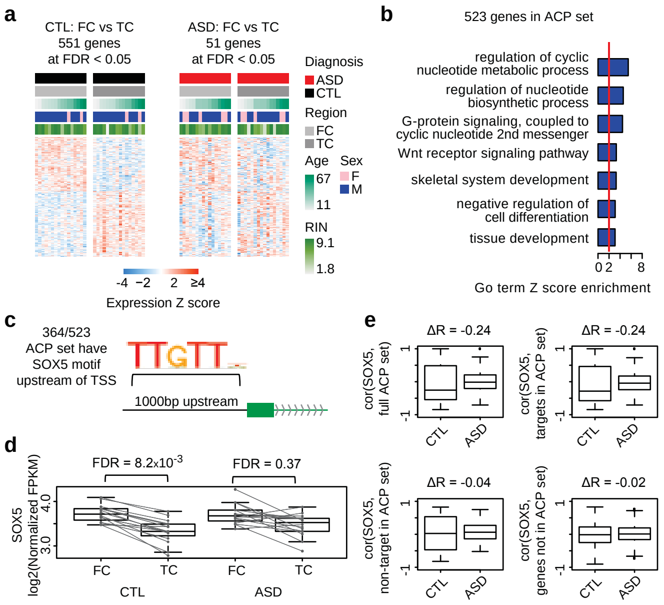
Author Contributions

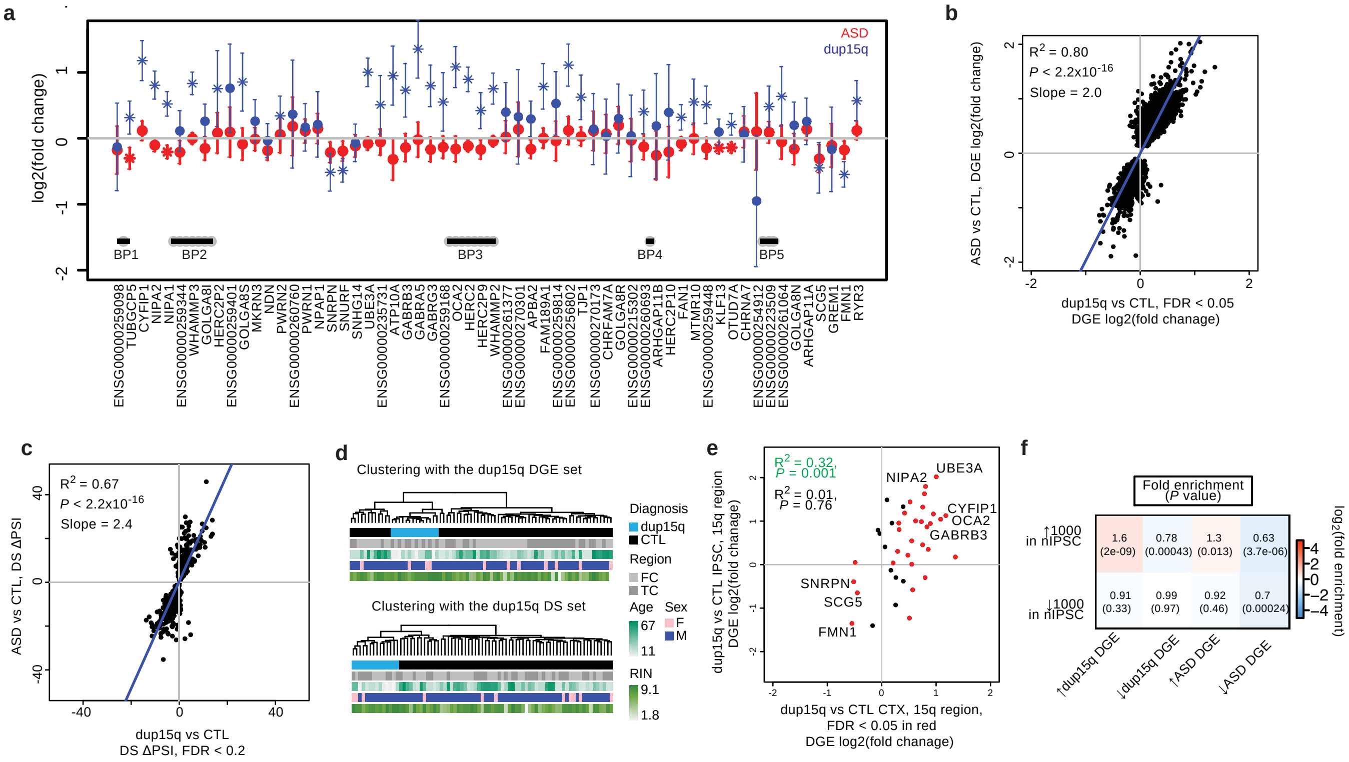
NNP, VS, and TGB performed dissections and RNA-seq analyses and differential gene expression analysis. NNP and VS performed splicing and co-expression network analysis. NNP and TGB performed analyses with Duplication 15q Syndrome individuals. NNP and MG reviewed clinical information and performed meta-analysis of ASD gene expression studies. VL and JKL performed

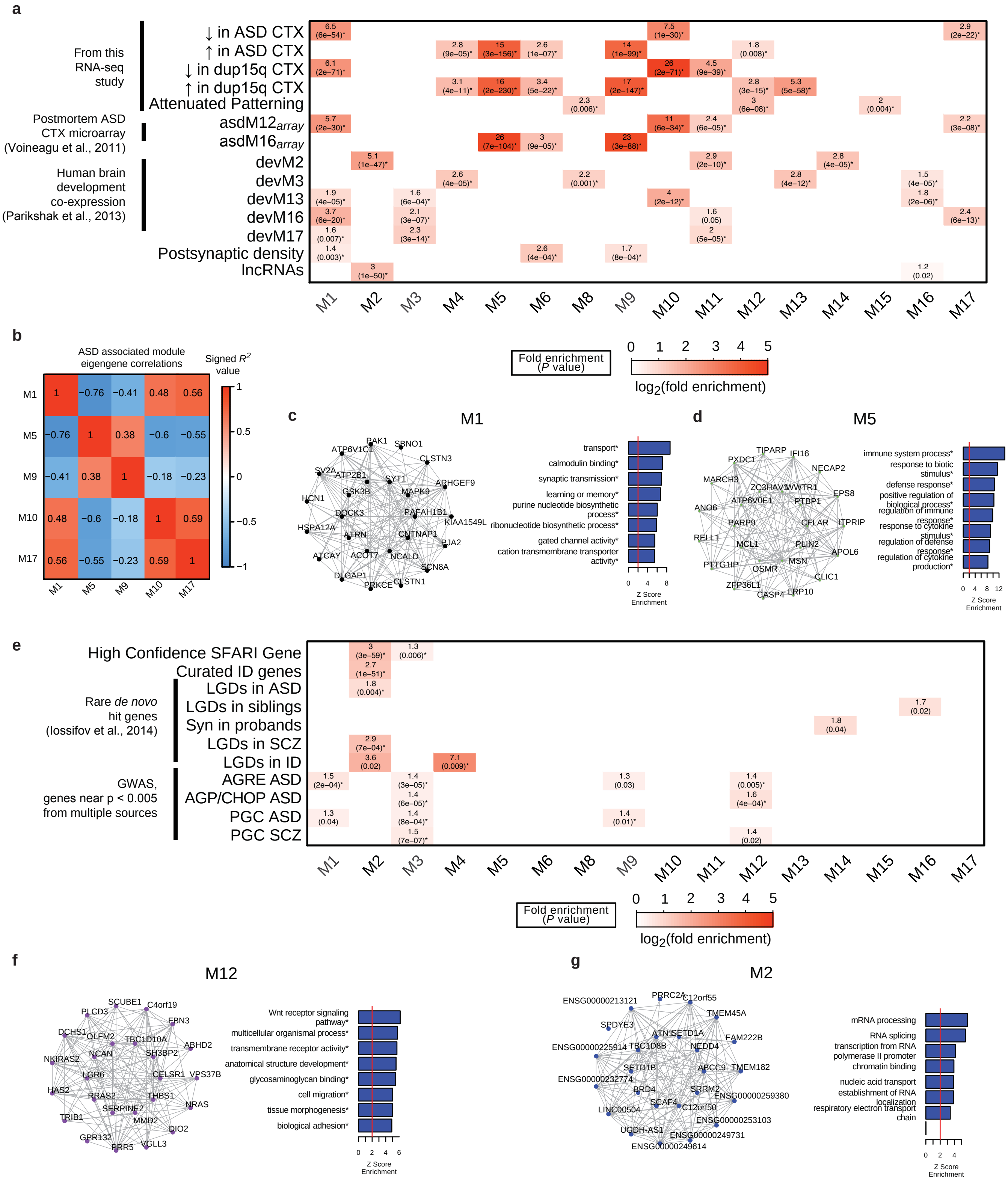
genotyping and CNV analysis on dup15q samples. VS performed validation experiments for gene and splicing level alterations in ASD. MI and BJB assisted with splicing analyses.  RJ performed dissections. SH provided guidance on differential gene expression and co-expression analyses. DHG provided guidance on all experiments and analyses. NNP and DHG wrote the manuscript. All authors contributed to revising and finalizing the manuscript.

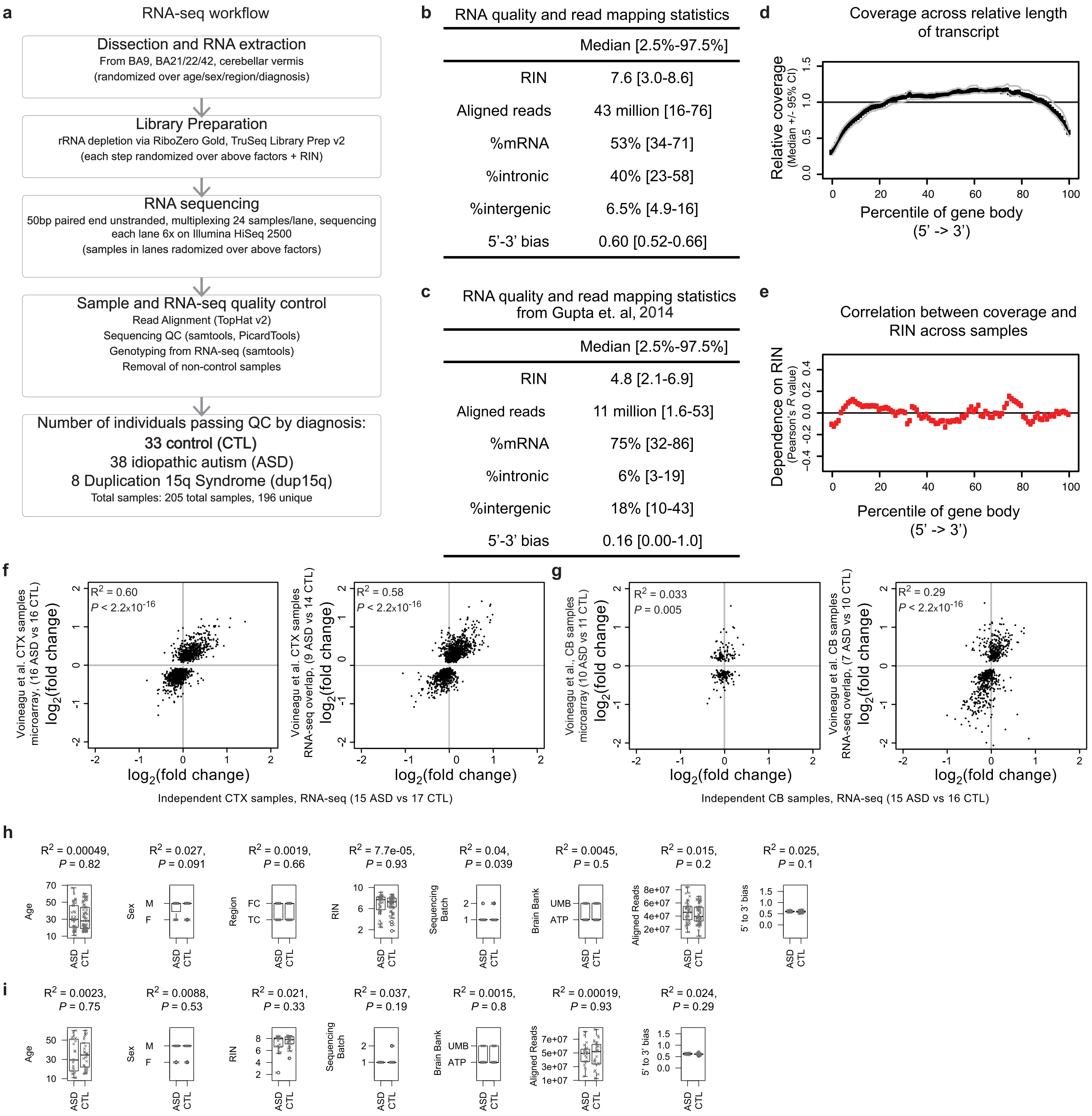

a
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
-5 −2 2 ≥5
Expression Z score top 100 protein coding transcripts
b
↓ in ASD
substrate-specific transmembrane transporter activity*
gated channel activity*
calmodulin binding
transport*
G2/M transition of mitotic cell cycle*
synaptic transmission*
0 2 8
Go term Z score enrichment
↑ in ASD
molecular transducer activity*
SH3 domain binding*
actin binding*
immune system process*
response to stimulus*
positive regulation of intracellular protein kinase cascade*
0 2 8
GO term Z score enrichment
c
P = 1.0x10-7
P = 4.3x10-6
P = 0.74
CTX DGE PC1 (38%)
0.2
0.0
ASD (Discovery)
ASD (Replication)
CTL
d
Diagnosis
Sex
Age
Region
RIN
Seizures
Psych. Meds
Postmortem interval
Tissue pH
Brain Weight
Sequencing Depth
5' to 3' bias
ADI-R score
A
B (nonverbal)
B (verbal)
C
D
2
10
CTX DGE PC1 association -log10(P value)
e
-4 −2 2 ≥4
Expression Z score lincRNA, antisense, and processed transcripts (FDR < 0.05)
f
LINC00693
LINC00689
0.25 RPM
28,650,000
28,700,000
28,750,000
28,800,000
158,805,000
158,810,000
158,815,000
158,820,000
158,825,000
ASD
TC
FC
CB
CTL
TC
FC
CB
Gencode Model
LINC00693
LINC00689
VIPR2
Chimp
Rhesus
Mouse
g
LINC00693
LINC00689
P = 2.5x10-3
P = 2.2x10-3
Adjusted log2(FPKM)
3.0
2.0
2
-2
ASD
CTL
P = 1.4x10-6
P = 3.6x10-6
log2(FPKM)
2.0
0.5
0.20
0.05
Fetal
Infant
Child
Teen
Adult
50+


a
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
≤-6 −2 2 ≥6
Standardized PSI for 569 events
on genes not exhibiting DGE
b
secretion
neuron projection morphogenesis
cell motility
biological adhesion
regulation of anatomical
structure morphogenesis
calcium ion binding
0 2 8
Z Score
Enrichment
c
P = 2.2x10-12
P = 8.9x10-5
P = 0.02
CTX DS PC1 (15%)
0.1
−0.1
−0.3
ASD
(Discovery)
ASD
(Replication)
CTL
d
Diagnosis
Sex
Age
Region
RIN
Seizures
Psych. Meds
Postmortem interval
Tissue pH
Brain Weight
Sequencing Depth
5' to 3' bias
ADI-R score
A
B (nonverbal)
B (verbal)
C
D
2
14
CTX DS PC1 association -log10(P value)
e
R2 between SF
expression in CTX
and CTX DS PC1
1.0
0.8
0.6
0.4
0.2
0.0
RBFOX1 (0.022)
RBFOX2 (0.12)
RBFOX3 (0.12)
SRRM4 (0.022)
NOVA1 (0.009)
NOVA2 (0.98)
MBNL1 (0.12)
MBNL2 (0.18)
MBNL3 (0.37)
PTBP1 (0.016)
PTBP2 (0.12)
Splicing factor (DGE FDR value in CTX)
f
R2 between SF
expression in CB
and CB DS PC1
1.0
0.8
0.6
0.4
0.2
0.0
RBFOX1 (0.73)
RBFOX2 (0.73)
RBFOX3 (0.73)
SRRM4 (0.93)
NOVA1 (0.93)
NOVA2 (0.73)
MBNL1 (0.99)
MBNL2 (0.66)
MBNL3 (0.83)
PTBP1 (0.83)
PTBP2 (0.66)
Splicing factor (DGE FDR value in CB)
g
Fold enrichment
(P value)
All events
Inclusion
in ASD
Exclusion
in ASD
1.9
(0.0058)
1.9
(2.8e−05)
2.1
(0.00012)
1.2
(0.19)
1
(0.93)
0.94
(1)
1.3
(0.29)
1.4
(0.34)
1.2
(0.44)
0.77
(0.48)
2.4
(0.00099)
2.2
(1.7e−05)
2.4
(9e−05)
1.2
(0.29)
1.1
(0.47)
RBFOX1
knockout
SRRM4
overexpression
PTBP1
knockdown
ESRP
overexpression
ESRP
knockdown
4
2
0
−2
−4
log2(fold enrichment)
h
CTX DGE PC1
(protein coding)
R2 = 0.86
R2 = 0.83
CTX DGE PC1
(noncoding)
R2 = 0.82
CTX DS PC1
-0.1
0.1
0.2
0.0
0.3
ASD
CTL


a
CTL: FC vs TC
551 genes
at FDR < 0.05
ASD: FC vs TC
51 genes
at FDR < 0.05
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
-4 −2 2 ≥4
Expression Z score
b
523 genes in ACP set
regulation of cyclic nucleotide metabolic process
regulation of nucleotide biosynthetic process
G-protein signaling, coupled to cyclic nucleotide 2nd messenger
Wnt receptor signaling pathway
skeletal system development
negative regulation of cell differentiation
tissue development
0 2 8
Go term Z score enrichment
c
364/523 ACP set have SOX5 motif upstream of TSS
TTGTT
1000bp upstream
d
FDR = 8.2x10^-3
FDR = 0.37
SOX5 log2(Normalized FPKM)
4.0
3.0
FC
TC
FC
TC
CTL
ASD
e
ΔR = -0.24
cor(SOX5, full ACP set)
ΔR = -0.24
cor(SOX5, targets in ACP set)
ΔR = -0.04
cor(SOX5, non-target in ACP set)
ΔR = -0.02
cor(SOX5, genes not in ACP set)
1
-1
CTL
ASD


a
ASD
dup15q
log2(fold change)
1
0
-1
-2
BP1
BP2
BP3
BP4
BP5
ENSG00000259098
TUBGCP5
CYFIP1
NIPA2
NIPA1
ENSG00000259344
WHAMMP3
GOLGA8I
HERC2P2
ENSG00000259401
GOLGA8S
MKRN3
NDN
PWRN2
ENSG00000260760
PWRN1
NPAP1
SNRPN
SNURF
SNHG14
UBE3A
ENSG00000235731
ATP10A
GABRB3
GABRA5
GABRG3
ENSG00000259168
OCA2
HERC2
HERC2P9
WHAMMP2
ENSG00000261377
ENSG00000270301
APBA2
FAM189A1
ENSG00000259814
ENSG00000256802
TJP1
ENSG00000270173
CHRFAM7A
GOLGA8R
ENSG00000215302
ENSG00000260693
ARHGAP11B
HERC2P10
FAN1
MTMR10
ENSG00000259448
KLF13
OTUD7A
CHRNA7
ENSG00000254912
ENSG00000223509
ENSG00000261064
GOLGA8N
ARHGAP11A
SCG5
GREM1
FMN1
RYR3
b
R² = 0.80
P < 2.2x10⁻¹⁶
Slope = 2.0
ASD vs CTL, DGE log2(fold change)
2
0
-2
-2
0
2
dup15q vs CTL, FDR < 0.05
DGE log2(fold chanage)
c
R² = 0.67
P < 2.2x10⁻¹⁶
Slope = 2.4
ASD vs CTL, DS ΔPSI
40
0
-40
-40
0
40
dup15q vs CTL
DS ΔPSI, FDR < 0.2
d
Clustering with the dup15q DGE set
Clustering with the dup15q DS set
Diagnosis
dup15q
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
e
R² = 0.32,
P = 0.001
R² = 0.01,
P = 0.76
UBE3A
NIPA2
CYFIP1
OCA2
GABRB3
SNRPN
SCG5
FMN1
dup15q vs CTL IPSC, 15q region
DGE log2(fold chanage)
2
1
0
-1
-2
-2
-1
0
1
2
dup15q vs CTL CTX, 15q region,
FDR < 0.05 in red
DGE log2(fold chanage)
f
Fold enrichment
(P value)
↑1000
in nIPSC
↓1000
in nIPSC
1.6
(2e-09)
0.78
(0.00043)
1.3
(0.013)
0.63
(3.7e-06)
0.91
(0.33)
0.99
(0.97)
0.92
(0.46)
0.7
(0.00024)
↑dup15q DGE
↓dup15q DGE
↑ASD DGE
↓ASD DGE
4
2
0
−2
−4
log₂(fold enrichment)


a
From this RNA-seq study
↓ in ASD CTX
↑ in ASD CTX
↓ in dup15q CTX
↑ in dup15q CTX
Attenuated Patterning
Postmortem ASD CTX microarray (Voineagu et al., 2011)
asdM12array
asdM16array
Human brain development co-expression (Parikshak et al., 2013)
devM2
devM3
devM13
devM16
devM17
Postsynaptic density
lncRNAs
↓ in ASD CTX: M1 6.5 (6e−54)*; M10 7.5 (1e−30)*; M17 2.9 (2e−22)*
↑ in ASD CTX: M4 2.8 (9e−05)*; M5 15 (3e−156)*; M6 2.6 (1e−07)*; M9 14 (1e−99)*; M12 1.8 (0.008)*
↓ in dup15q CTX: M1 6.1 (2e−71)*; M10 26 (2e−71)*; M11 4.5 (9e−39)*
↑ in dup15q CTX: M4 3.1 (4e−11)*; M5 16 (2e−230)*; M6 3.4 (5e−22)*; M9 17 (2e−147)*; M12 2.8 (3e−15)*; M13 5.3 (5e−58)*
Attenuated Patterning: M8 2.3 (0.006)*; M12 3 (6e−08)*; M15 2 (0.004)*
asdM12array: M1 5.7 (2e−30)*; M10 11 (6e−34)*; M11 2.4 (6e−05)*; M17 2.2 (3e−08)*
asdM16array: M5 26 (7e−104)*; M6 3 (9e−05)*; M9 23 (3e−88)*
devM2: M2 5.1 (1e−47)*; M11 2.9 (2e−10)*; M14 2.8 (4e−05)*
devM3: M4 2.6 (4e−05)*; M8 2.2 (0.001)*; M13 2.8 (4e−12)*; M16 1.5 (4e−05)*
devM13: M1 1.9 (4e−05)*; M3 1.6 (6e−04)*; M10 4 (2e−12)*; M16 1.8 (2e−06)*
devM16: M1 3.7 (6e−20)*; M3 2.1 (3e−07)*; M11 1.6 (0.05); M17 2.4 (6e−13)*
devM17: M1 1.6 (0.007)*; M3 2.3 (3e−14)*; M11 2 (5e−05)*
Postsynaptic density: M1 1.4 (0.003)*; M6 2.6 (4e−04)*; M9 1.7 (8e−04)*
lncRNAs: M2 3 (1e−50)*; M16 1.2 (0.02)
M1 M2 M3 M4 M5 M6 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17
Fold enrichment (P value)
0 1 2 3 4 5
log2(fold enrichment)
b
ASD associated module eigengene correlations
Signed R2 value
1 0.5 0 −0.5 −1
M1: 1, −0.76, −0.41, 0.48, 0.56
M5: −0.76, 1, 0.38, −0.6, −0.55
M9: −0.41, 0.38, 1, −0.18, −0.23
M10: 0.48, −0.6, −0.18, 1, 0.59
M17: 0.56, −0.55, −0.23, 0.59, 1
M1 M5 M9 M10 M17
c
M1
PAK1 SBNO1 ATP6V1C1 CLSTN3 SV2A ATP2B1 SYT1 ARHGEF9 HCN1 GSK3B MAPK9 DOCK3 PAFAH1B1 KIAA1549L HSPA12A ATRN CNTNAP1 PJA2 ATCAY ACOT7 NCALD SCN8A DLGAP1 PRKCE CLSTN1
transport*
calmodulin binding*
synaptic transmission*
learning or memory*
purine nucleotide biosynthetic process*
ribonucleotide biosynthetic process*
gated channel activity*
cation transmembrane transporter activity*
0 4 8
Z Score Enrichment
d
M5
TIPARP IFI16 PXDC1 NECAP2 MARCH3 ZC3HAV1 WWTR1 EPS8 ATP6V0E1 PTBP1 ANO6 PARP9 CFLAR ITPRIP RELL1 MCL1 PLIN2 APOL6 PTTG1IP OSMR MSN CLIC1 ZFP36L1 CASP4 LRP10
immune system process*
response to biotic stimulus*
defense response*
positive regulation of biological process*
regulation of immune response*
response to cytokine stimulus*
regulation of defense response*
regulation of cytokine production*
0 4 8 12
Z Score Enrichment
e
High Confidence SFARI Gene
Curated ID genes
Rare de novo hit genes (Iossifov et al., 2014)
LGDs in ASD
LGDs in siblings
Syn in probands
LGDs in SCZ
LGDs in ID
GWAS, genes near p < 0.005 from multiple sources
AGRE ASD
AGP/CHOP ASD
PGC ASD
PGC SCZ
High Confidence SFARI Gene: M2 3 (3e−59)*; M3 1.3 (0.006)*
Curated ID genes: M2 2.7 (1e−51)*
LGDs in ASD: M2 1.8 (0.004)*
LGDs in siblings: M16 1.7 (0.02)
Syn in probands: M14 1.8 (0.04)
LGDs in SCZ: M2 2.9 (7e−04)*
LGDs in ID: M2 3.6 (0.02); M4 7.1 (0.009)*
AGRE ASD: M1 1.5 (2e−04)*; M3 1.4 (3e−05)*; M9 1.3 (0.03); M12 1.4 (0.005)*
AGP/CHOP ASD: M3 1.4 (6e−05)*; M12 1.6 (4e−04)*
PGC ASD: M1 1.3 (0.04); M3 1.4 (8e−04)*; M9 1.4 (0.01)*
PGC SCZ: M3 1.5 (7e−07)*; M12 1.4 (0.02)
M1 M2 M3 M4 M5 M6 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17
Fold enrichment (P value)
0 1 2 3 4 5
log2(fold enrichment)
f
M12
SCUBE1 C4orf19 PLCD3 FBN3 DCHS1 OLFM2 TBC1D10A ABHD2 SH3BP2 NKIRAS2 NCAN LGR6 CELSR1 VPS37B HAS2 RRAS2 THBS1 SERPINE2 NRAS TRIB1 MMD2 DIO2 GPR132 PRR5 VGLL3
Wnt receptor signaling pathway*
multicellular organismal process*
transmembrane receptor activity*
anatomical structure development*
glycosaminoglycan binding*
cell migration*
tissue morphogenesis*
biological adhesion*
0 2 4 6
Z Score Enrichment
g
M2
PRRC2A C12orf55 ENSG00000213121 TMEM45A SPDYE3 ATN1 SETD1A FAM222B ENSG00000225914 TBC1D8B NEDD4 SETD1B ABCC9 TMEM182 ENSG00000232774 BRD4 SRRM2 ENSG00000259380 SCAF4 C12orf50 LINC00504 ENSG00000253103 UGDH-AS1 ENSG00000249731 ENSG00000249614
mRNA processing
RNA splicing
transcription from RNA polymerase II promoter
chromatin binding
nucleic acid transport
establishment of RNA localization
respiratory electron transport chain
0 2 4
Z Score Enrichment



**a** RNA-seq workflow

Dissection and RNA extraction
From BA9, BA21/22/42, cerebellar vermis
(randomized over age/sex/region/diagnosis)

Library Preparation
rRNA depletion via RiboZero Gold, TruSeq Library Prep v2
(each step randomized over above factors + RIN)

RNA sequencing
50bp paired end unstranded, multiplexing 24 samples/lane, sequencing each lane 6x on Illumina HiSeq 2500
(samples in lanes radomized over above factors)

Sample and RNA-seq quality control
Read Alignment (TopHat v2)
Sequencing QC (samtools, PicardTools)
Genotyping from RNA-seq (samtools)
Removal of non-control samples

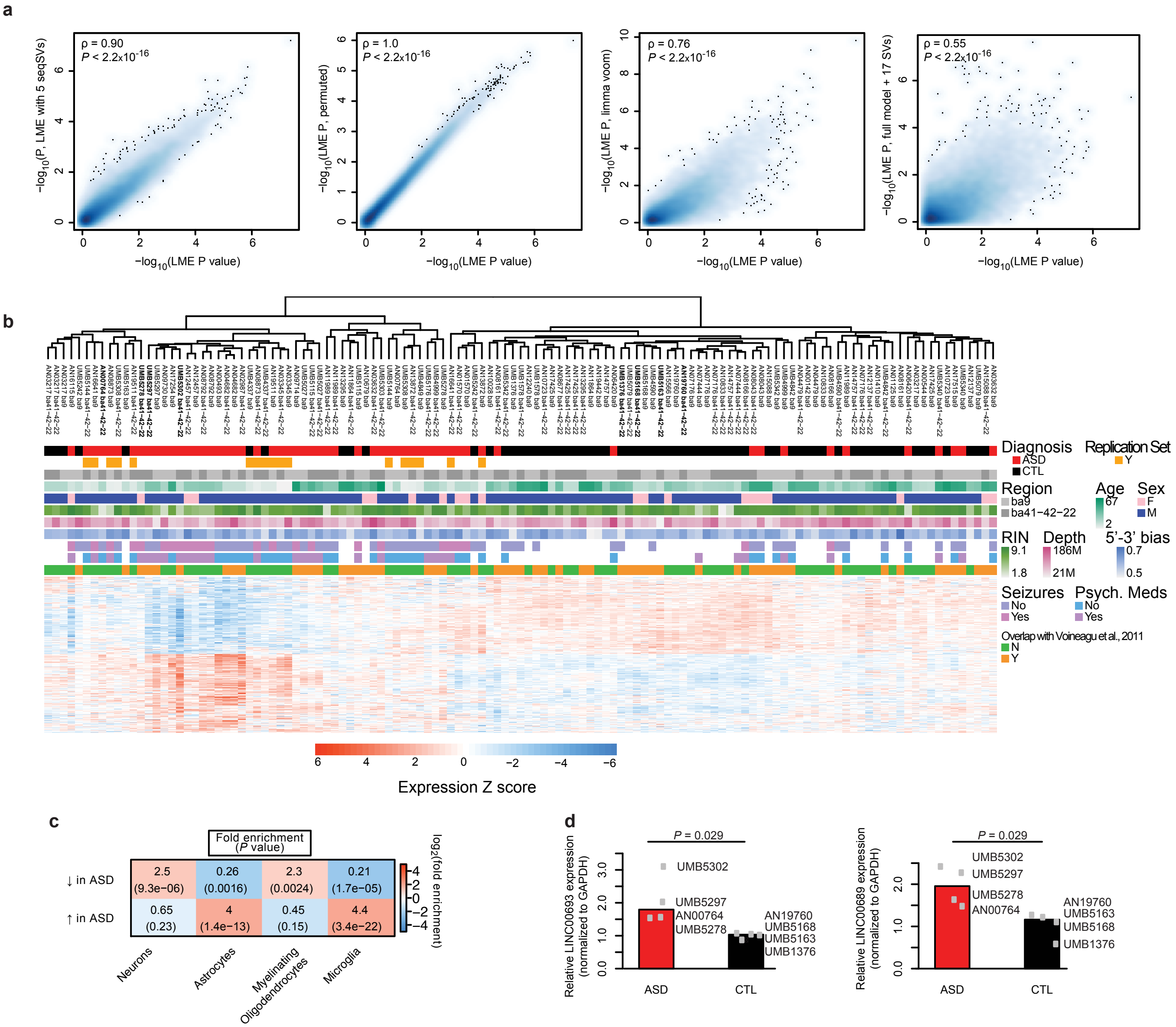
Number of individuals passing QC by diagnosis:
33 control (CTL)
38 idiopathic autism (ASD)
8 Duplication 15q Syndrome (dup15q)
Total samples: 205 total samples, 196 unique

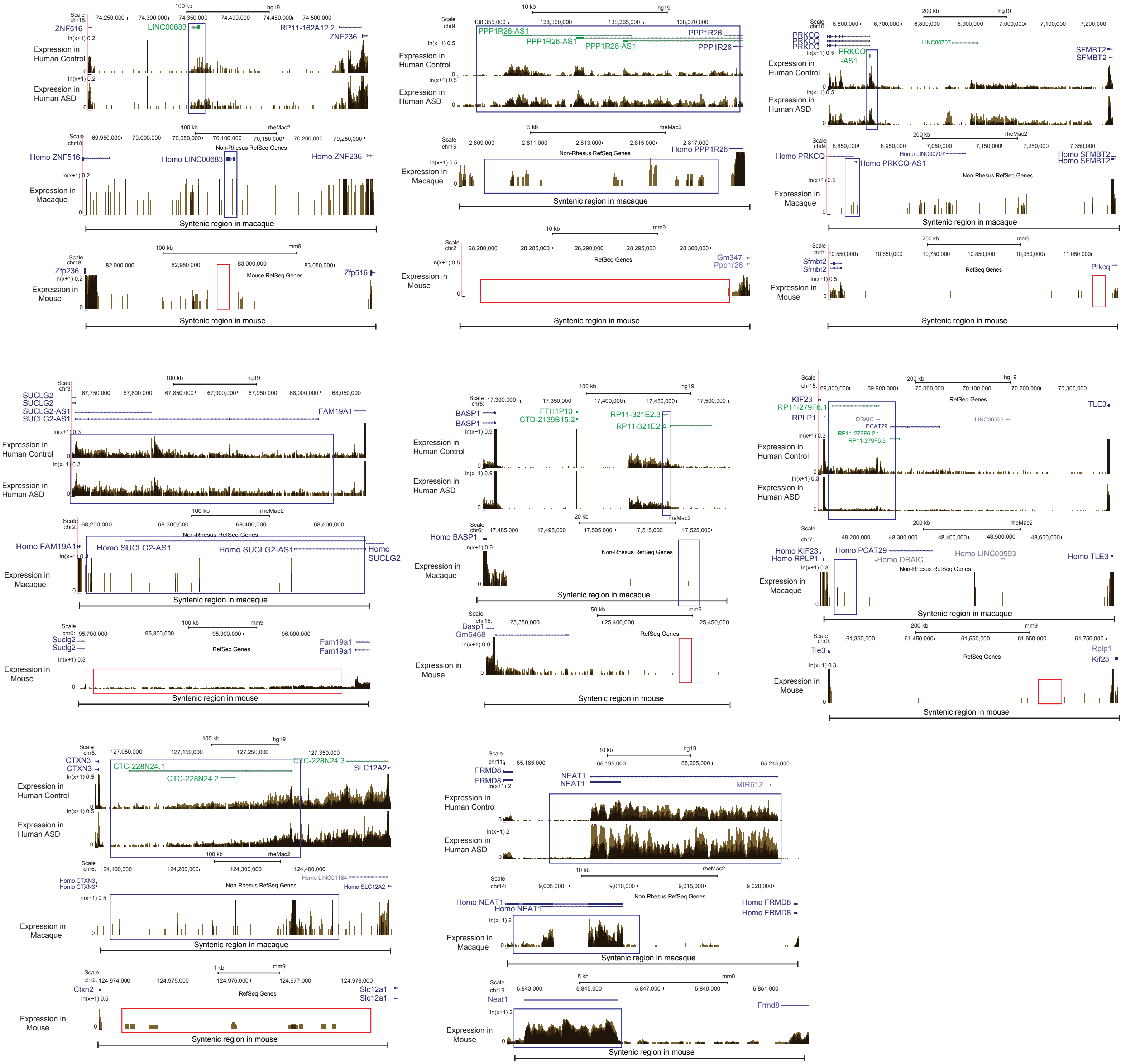
**b** RNA quality and read mapping statistics

| | Median [2.5%-97.5%] |
|---|---|
| RIN | 7.6 [3.0-8.6] |
| Aligned reads | 43 million [16-76] |
| %mRNA | 53% [34-71] |
| %intronic | 40% [23-58] |
| %intergenic | 6.5% [4.9-16] |
| 5'-3' bias | 0.60 [0.52-0.66] |

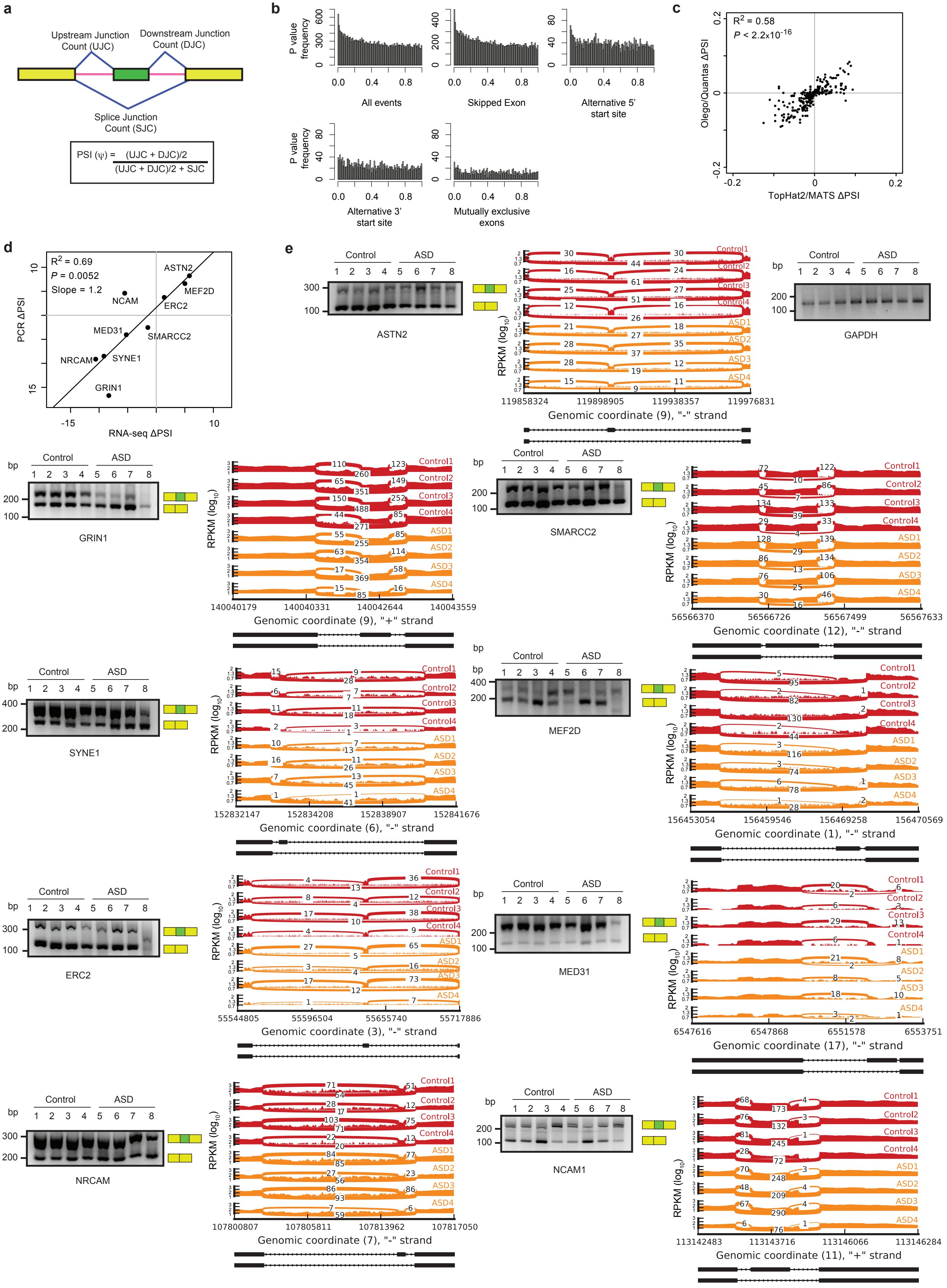
**c** RNA quality and read mapping statistics from Gupta et. al, 2014

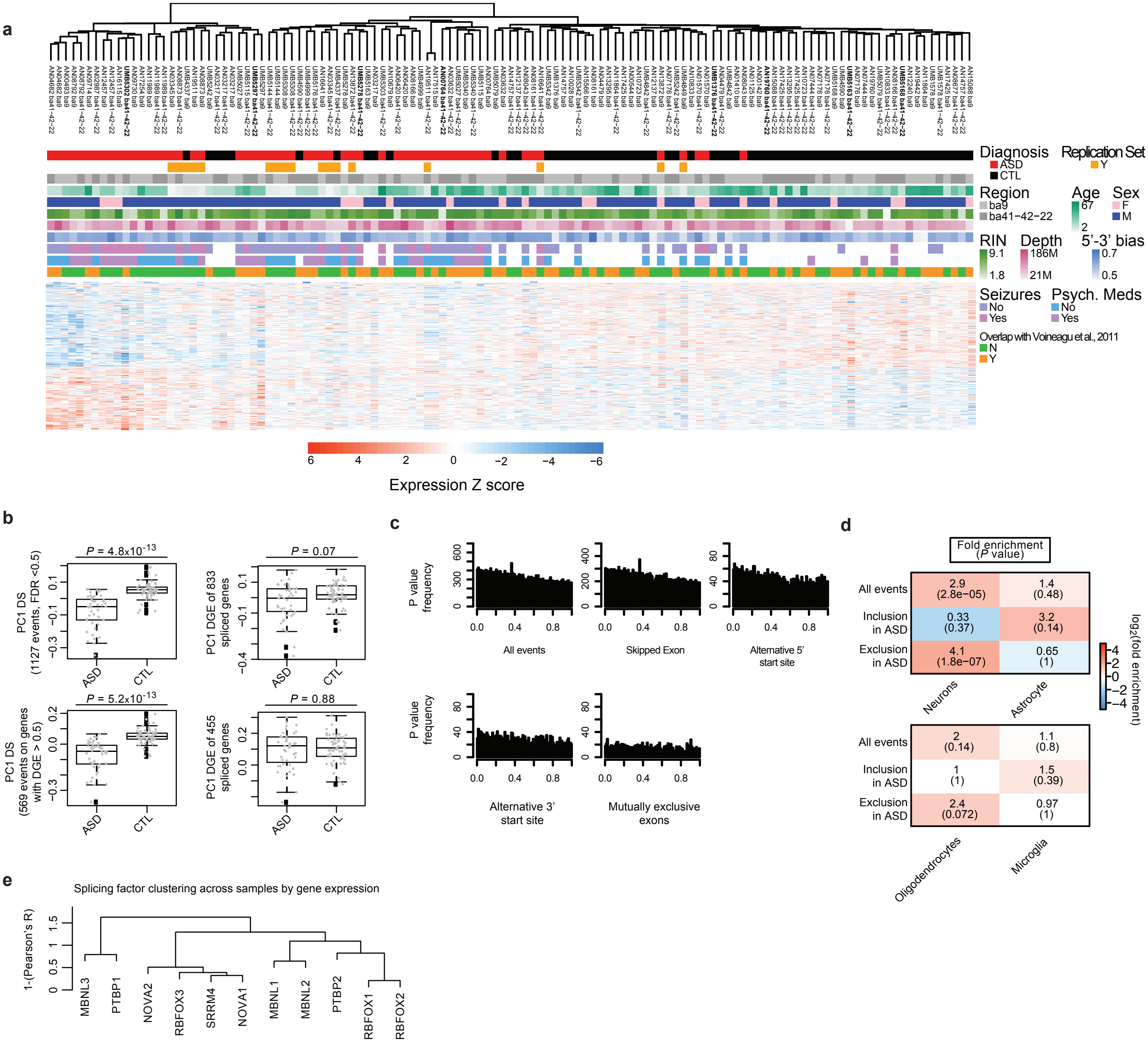
| | Median [2.5%-97.5%] |
|---|---|
| RIN | 4.8 [2.1-6.9] |
| Aligned reads | 11 million [1.6-53] |
| %mRNA | 75% [32-86] |
| %intronic | 6% [3-19] |
| %intergenic | 18% [10-43] |
| 5'-3' bias | 0.16 [0.00-1.0] |

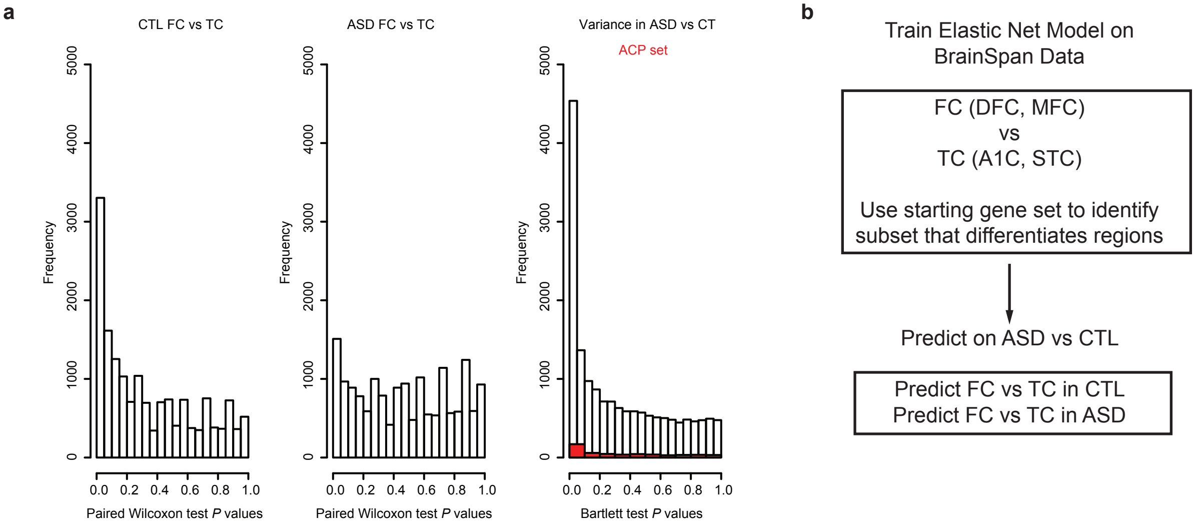

a
ρ = 0.90
P < 2.2x10-16
−log10(P, LME with 5 seqSVs)
−log10(LME P value)
ρ = 1.0
P < 2.2x10-16
−log10(LME P, permuted)
−log10(LME P value)
ρ = 0.76
P < 2.2x10-16
−log10(LME P, limma voom)
−log10(LME P value)
ρ = 0.55
P < 2.2x10-16
−log10(LME P, full model + 17 SVs)
−log10(LME P value)
b
Diagnosis
ASD
CTL
Replication Set
Y
Region
ba9
ba41−42−22
Age
67
2
Sex
F
M
RIN
9.1
1.8
Depth
186M
21M
5'-3' bias
0.7
0.5
Seizures
No
Yes
Psych. Meds
No
Yes
Overlap with Voineagu et al., 2011
N
Y
6
4
2
0
−2
−4
−6
Expression Z score
c
Fold enrichment
(P value)
↓ in ASD
↑ in ASD
2.5 (9.3e−06)
0.26 (0.0016)
2.3 (0.0024)
0.21 (1.7e−05)
0.65 (0.23)
4 (1.4e−13)
0.45 (0.15)
4.4 (3.4e−22)
Neurons
Astrocytes
Myelinating Oligodendrocytes
Microglia
log2(fold enrichment)
4
2
0
−2
−4
d
P = 0.029
Relative LINC00693 expression (normalized to GAPDH)
UMB5302
UMB5297
AN00764
UMB5278
AN19760
UMB5168
UMB5163
UMB1376
ASD
CTL
P = 0.029
Relative LINC00689 expression (normalized to GAPDH)
UMB5302
UMB5297
UMB5278
AN00764
AN19760
UMB5163
UMB5168
UMB1376
ASD
CTL

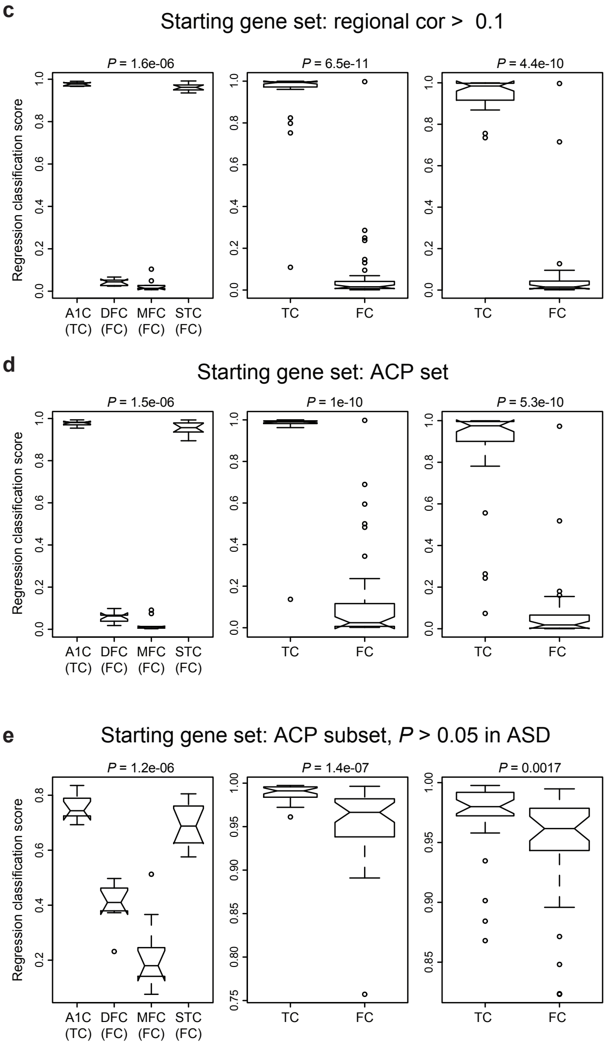


ZNF516
LINC00683
RP11-162A12.2
ZNF236
Expression in Human Control
Expression in Human ASD
Homo ZNF516
Homo LINC00683
Homo ZNF236
Non-Rhesus RefSeq Genes
Expression in Macaque
Syntenic region in macaque
Zfp236
Zfp516
Mouse RefSeq Genes
Expression in Mouse
Syntenic region in mouse
PPP1R26-AS1
PPP1R26
Homo PPP1R26
Gm347
Ppp1r26
RefSeq Genes
PRKCQ
PRKCQ-AS1
LINC00707
SFMBT2
Homo PRKCQ
Homo PRKCQ-AS1
Homo LINC00707
Homo SFMBT2
Sfmbt2
Prkcq
SUCLG2
SUCLG2-AS1
FAM19A1
Homo FAM19A1
Homo SUCLG2-AS1
Homo SUCLG2
Suclg2
Fam19a1
BASP1
FTH1P10
CTD-2139B15.2
RP11-321E2.3
RP11-321E2.4
Homo BASP1
Basp1
Gm5468
KIF23
RP11-279F6.1
RPLP1
DRAIC
PCAT29
RP11-279F6.2
RP11-279F6.3
LINC00593
TLE3
Homo KIF23
Homo RPLP1
Homo PCAT29
Homo DRAIC
Homo LINC00593
Homo TLE3
Tle3
Rplp1
Kif23
CTXN3
CTC-228N24.1
CTC-228N24.2
CTC-228N24.3
SLC12A2
Homo CTXN3
Homo LINC01184
Homo SLC12A2
Ctxn2
Slc12a1
FRMD8
NEAT1
MIR612
Homo NEAT1
Homo FRMD8
Neat1
Frmd8
hg19
rheMac2
mm9


a
Upstream Junction Count (UJC)
Downstream Junction Count (DJC)
Splice Junction Count (SJC)
PSI (ψ) = (UJC + DJC)/2 / ((UJC + DJC)/2 + SJC)
b
P value frequency
All events
Skipped Exon
Alternative 5' start site
Alternative 3' start site
Mutually exclusive exons
c
R² = 0.58
P < 2.2x10⁻¹⁶
Olego/Quantas ΔPSI
TopHat2/MATS ΔPSI
d
R² = 0.69
P = 0.0052
Slope = 1.2
PCR ΔPSI
RNA-seq ΔPSI
ASTN2
MEF2D
NCAM
ERC2
MED31
SMARCC2
NRCAM
SYNE1
GRIN1
e
Control
ASD
bp
1 2 3 4 5 6 7 8
ASTN2
GAPDH
GRIN1
SMARCC2
SYNE1
MEF2D
ERC2
MED31
NRCAM
NCAM1
RPKM (log10)
Control1
Control2
Control3
Control4
ASD1
ASD2
ASD3
ASD4
Genomic coordinate (9), "-" strand
Genomic coordinate (9), "+" strand
Genomic coordinate (12), "-" strand
Genomic coordinate (6), "-" strand
Genomic coordinate (1), "-" strand
Genomic coordinate (3), "-" strand
Genomic coordinate (17), "-" strand
Genomic coordinate (7), "-" strand
Genomic coordinate (11), "+" strand

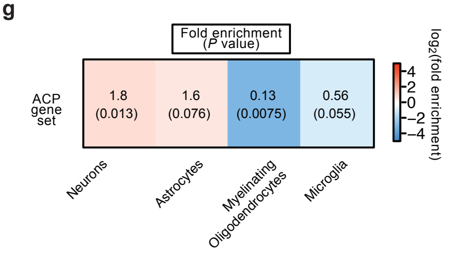

a
Diagnosis
ASD
CTL
Replication Set
Y
Region
ba9
ba41-42-22
Age
67
2
Sex
F
M
RIN
9.1
1.8
Depth
186M
21M
5'-3' bias
0.7
0.5
Seizures
No
Yes
Psych. Meds
No
Yes
Overlap with Voineagu et al., 2011
N
Y
6
4
2
0
−2
−4
−6
Expression Z score
b
P = 4.8x10^-13
PC1 DS (1127 events, FDR <0.5)
P = 0.07
PC1 DGE of 833 spliced genes
P = 5.2x10^-13
PC1 DS (569 events on genes with DGE > 0.5)
P = 0.88
PC1 DGE of 455 spliced genes
ASD
CTL
c
P value frequency
All events
Skipped Exon
Alternative 5' start site
Alternative 3' start site
Mutually exclusive exons
d
Fold enrichment (P value)
All events
Inclusion in ASD
Exclusion in ASD
2.9 (2.8e−05)
1.4 (0.48)
0.33 (0.37)
3.2 (0.14)
4.1 (1.8e−07)
0.65 (1)
Neurons
Astrocyte
2 (0.14)
1.1 (0.8)
1 (1)
1.5 (0.39)
2.4 (0.072)
0.97 (1)
Oligodendrocytes
Microglia
log2(fold enrichment)
4
2
0
−2
−4
e
Splicing factor clustering across samples by gene expression
1-(Pearson's R)
MBNL3
PTBP1
NOVA2
RBFOX3
SRRM4
NOVA1
MBNL1
MBNL2
PTBP2
RBFOX1
RBFOX2





f

| Starting gene set | #Genes kept | AUROC BrainSpan | AUROC CTL | AUROC ASD |
|---|---|---|---|---|
| Regional cor > 0.1 | 71 | 1 | 0.97 | 0.98 |
| ACP set | 46 | 1 | 0.97 | 0.98 |
| ACP subset, *P* > 0.05 in ASD | 48 | 1 | 0.88 | 0.74 |



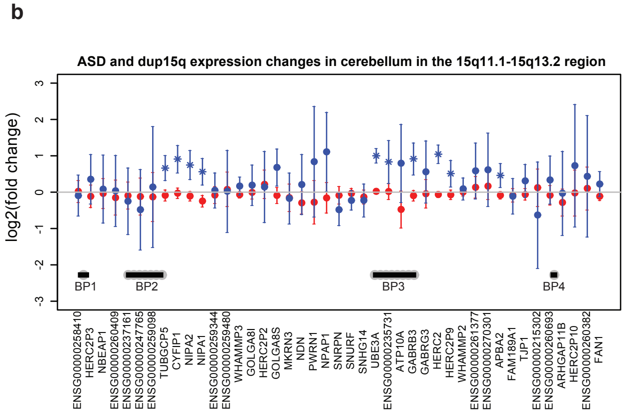
**a**

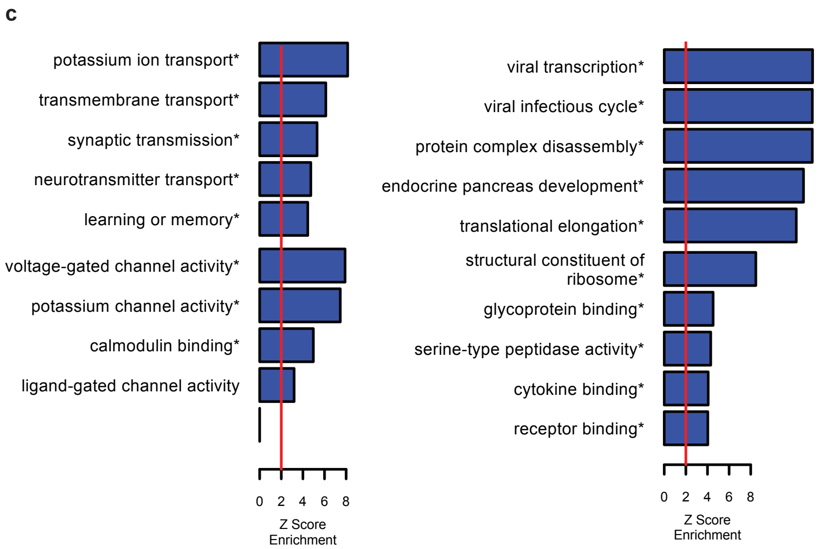
Duplication 15q breakpoints across individuals

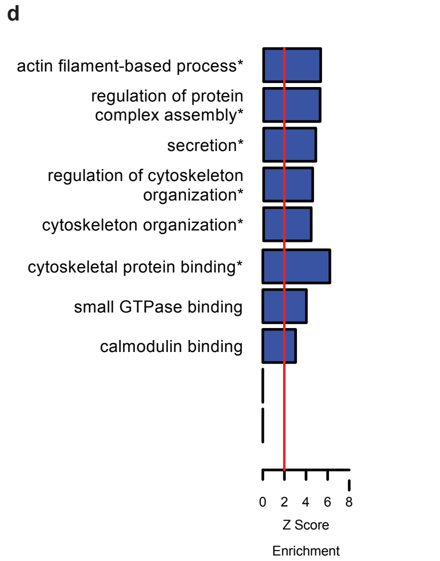
| Sample | BP1-2 | BP2-3 | BP3-4 | BP4-5 |
|---|---|---|---|---|
| AN09402 | 4 | 4,b | 2 | 2 |
| AN14829 | 4 | 4 | 4 | 3 |
| AN17138 | 4 | 4 | 2 | 2 |
| AN03935 | 4 | 4 | 4 | 3 |
| AN05983 | 4 | 4 | 4 | 3 |
| AN06365 | 4 | 4 | 4 | 3 |
| AN11931 | 4 | 4 | 4 | 3 |
| AN14762 | - | 4,a | - | - |

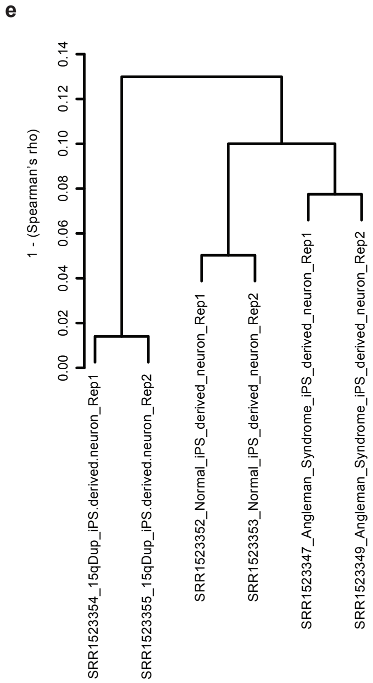
a,Obtained from Scoles et al., 2011 who evaluated duplication in this region by RT-PCR of SNRPN/GABRB3/UBE3A vs B2M

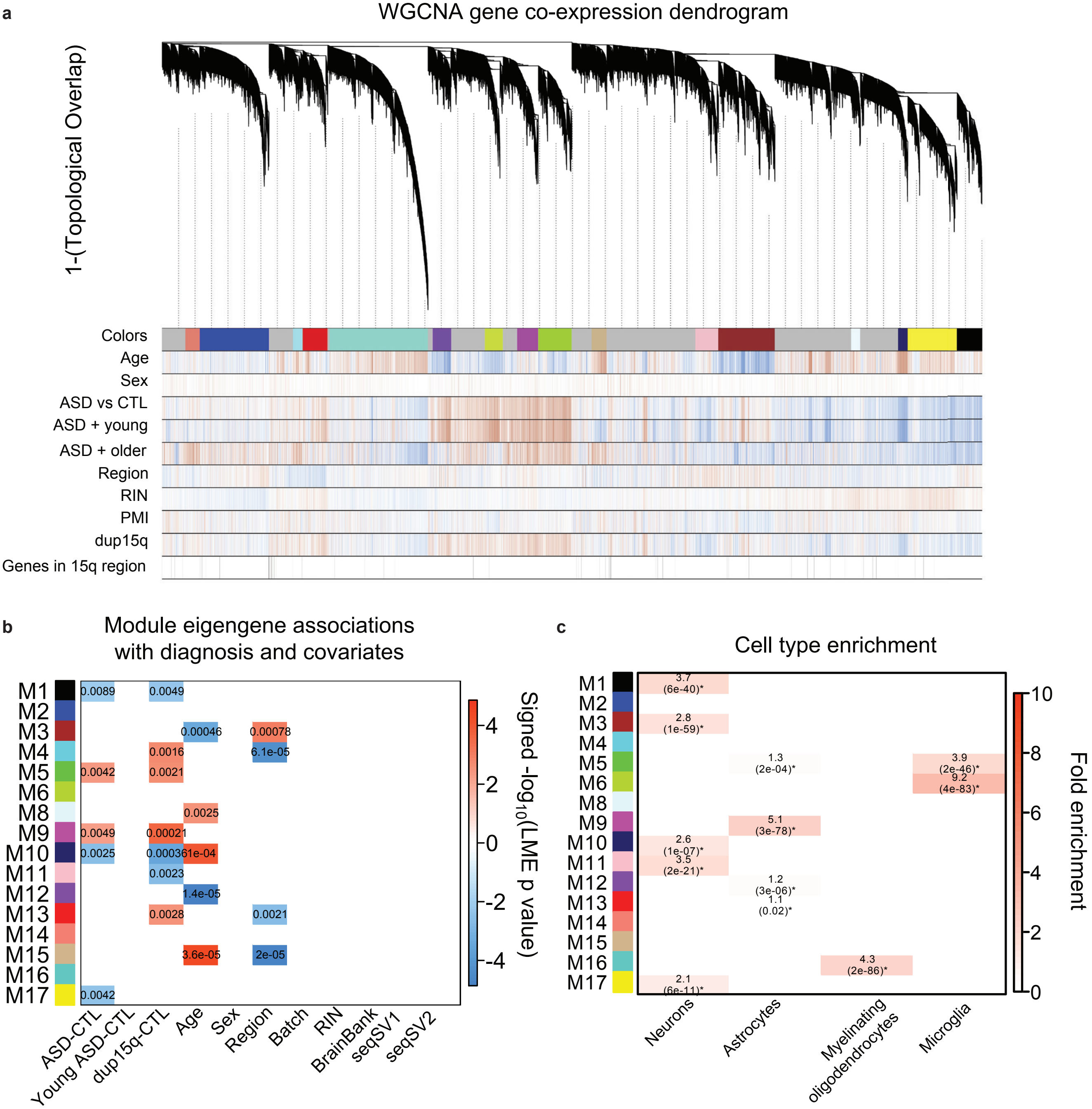
b,Discrepancy with Scoles et al., who report 5 here

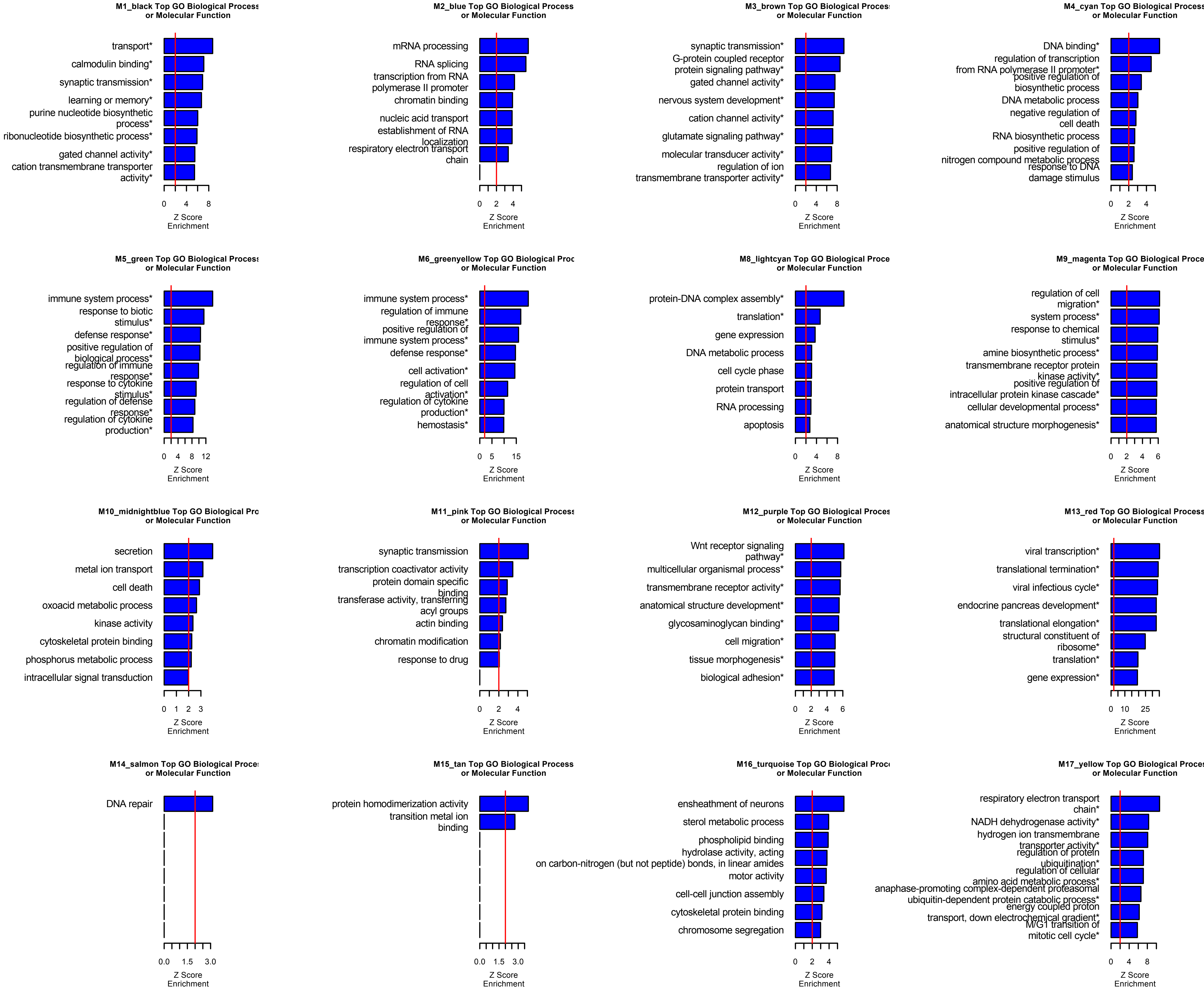









a
WGCNA gene co-expression dendrogram
1-(Topological Overlap)
Colors
Age
Sex
ASD vs CTL
ASD + young
ASD + older
Region
RIN
PMI
dup15q
Genes in 15q region
b
Module eigengene associations
with diagnosis and covariates
M1
M2
M3
M4
M5
M6
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
0.0089
0.0049
0.00046
0.00078
0.0016
6.1e-05
0.0042
0.0021
0.0025
0.0049
0.00021
0.0025
0.00036
1e-04
0.0023
1.4e-05
0.0028
0.0021
3.6e-05
2e-05
0.0042
ASD-CTL
Young ASD-CTL
dup15q-CTL
Age
Sex
Region
Batch
RIN
BrainBank
seqSV1
seqSV2
4
2
0
-2
-4
Signed -log10(LME p value)
c
Cell type enrichment
M1
M2
M3
M4
M5
M6
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
3.7
(6e-40)*
2.8
(1e-59)*
1.3
(2e-04)*
3.9
(2e-46)*
9.2
(4e-83)*
5.1
(3e-78)*
2.6
(1e-07)*
3.5
(2e-21)*
1.2
(3e-06)*
1.1
(0.02)*
4.3
(2e-86)*
2.1
(6e-11)*
Neurons
Astrocytes
Myelinating
oligodendrocytes
Microglia
10
8
6
4
2
0
Fold enrichment


M1_black Top GO Biological Process
or Molecular Function
transport*
calmodulin binding*
synaptic transmission*
learning or memory*
purine nucleotide biosynthetic process*
ribonucleotide biosynthetic process*
gated channel activity*
cation transmembrane transporter activity*
0 4 8
Z Score Enrichment
M2_blue Top GO Biological Process
or Molecular Function
mRNA processing
RNA splicing
transcription from RNA polymerase II promoter
chromatin binding
nucleic acid transport
establishment of RNA localization
respiratory electron transport chain
0 2 4
Z Score Enrichment
M3_brown Top GO Biological Proces
or Molecular Function
synaptic transmission*
G-protein coupled receptor protein signaling pathway*
gated channel activity*
nervous system development*
cation channel activity*
glutamate signaling pathway*
molecular transducer activity*
regulation of ion transmembrane transporter activity*
0 4 8
Z Score Enrichment
M4_cyan Top GO Biological Process
or Molecular Function
DNA binding*
regulation of transcription from RNA polymerase II promoter*
positive regulation of biosynthetic process
DNA metabolic process
negative regulation of cell death
RNA biosynthetic process
positive regulation of nitrogen compound metabolic process
response to DNA damage stimulus
0 2 4
Z Score Enrichment
M5_green Top GO Biological Process
or Molecular Function
immune system process*
response to biotic stimulus*
defense response*
positive regulation of biological process*
regulation of immune response*
response to cytokine stimulus*
regulation of defense response*
regulation of cytokine production*
0 4 8 12
Z Score Enrichment
M6_greenyellow Top GO Biological Proc
or Molecular Function
immune system process*
regulation of immune response*
positive regulation of immune system process*
defense response*
cell activation*
regulation of cell activation*
regulation of cytokine production*
hemostasis*
0 5 15
Z Score Enrichment
M8_lightcyan Top GO Biological Proce
or Molecular Function
protein-DNA complex assembly*
translation*
gene expression
DNA metabolic process
cell cycle phase
protein transport
RNA processing
apoptosis
0 4 8
Z Score Enrichment
M9_magenta Top GO Biological Proce
or Molecular Function
regulation of cell migration*
system process*
response to chemical stimulus*
amine biosynthetic process*
transmembrane receptor protein kinase activity*
positive regulation of intracellular protein kinase cascade*
cellular developmental process*
anatomical structure morphogenesis*
0 2 4 6
Z Score Enrichment
M10_midnightblue Top GO Biological Pro
or Molecular Function
secretion
metal ion transport
cell death
oxoacid metabolic process
kinase activity
cytoskeletal protein binding
phosphorus metabolic process
intracellular signal transduction
0 1 2 3
Z Score Enrichment
M11_pink Top GO Biological Process
or Molecular Function
synaptic transmission
transcription coactivator activity
protein domain specific binding
transferase activity, transferring acyl groups
actin binding
chromatin modification
response to drug
0 2 4
Z Score Enrichment
M12_purple Top GO Biological Proces
or Molecular Function
Wnt receptor signaling pathway*
multicellular organismal process*
transmembrane receptor activity*
anatomical structure development*
glycosaminoglycan binding*
cell migration*
tissue morphogenesis*
biological adhesion*
0 2 4 6
Z Score Enrichment
M13_red Top GO Biological Process
or Molecular Function
viral transcription*
translational termination*
viral infectious cycle*
endocrine pancreas development*
translational elongation*
structural constituent of ribosome*
translation*
gene expression*
0 10 25
Z Score Enrichment
M14_salmon Top GO Biological Proce
or Molecular Function
DNA repair
0.0 1.5 3.0
Z Score Enrichment
M15_tan Top GO Biological Process
or Molecular Function
protein homodimerization activity
transition metal ion binding
0.0 1.5 3.0
Z Score Enrichment
M16_turquoise Top GO Biological Proc
or Molecular Function
ensheathment of neurons
sterol metabolic process
phospholipid binding
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides
motor activity
cell-cell junction assembly
cytoskeletal protein binding
chromosome segregation
0 2 4
Z Score Enrichment
M17_yellow Top GO Biological Proces
or Molecular Function
respiratory electron transport chain*
NADH dehydrogenase activity*
hydrogen ion transmembrane transporter activity*
regulation of protein ubiquitination*
regulation of cellular amino acid metabolic process*
anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process*
energy coupled proton transport, down electrochemical gradient*
M/G1 transition of mitotic cell cycle*
0 4 8
Z Score Enrichment

## Title

Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development

## Authors and affiliations

Hyejung Won[1], Luis de la Torre-Ubieta[1], Jason L. Stein[1], Neelroop N. Parikshak[1], Farhad Hormozdiari[3], Changhoon Lee[1], Eleazar Eskin[3,4], Jason Ernst[2,4], Daniel H. Geschwind[1,4*]

[1] Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles

[2] Department of Biological Chemistry, David Geffen School of Medicine, University of California Los Angeles

[3] Department of Computer Science, University of California Los Angeles

[4] Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles

[5]Department of Molecular, Cell and Developmental Biology, University of California Los Angeles, Los Angeles

[*] Correspondence: dhg@mednet.ucla.edu

## Introduction

The demonstration that chromatin exhibits a complex 3 dimensional organization, whereby short and long distance physical interactions correspond to complex gene regulatory processes has opened a new window on understanding the functional organization of the human genome[1-4]. Recently, chromatin remodeling has also been causally implicated in several neurodevelopmental disorders, including autism and schizophrenia[5-7]. However, it remains unclear whether knowledge of chromosome organization in a tissue specific manner might inform our understanding of gene regulation in brain development or disease. Here we determined the genome-wide landscape of chromosome conformation during early human cortical development by performing Hi-C analysis in the mitotically active and post mitotic laminae of human fetal brain. We integrate Hi-C data with transcriptomic and epigenomic data and utilize chromosome contact information to delineate physical gene-gene regulatory interactions for non-coding regulatory elements. We show how these data permit large-scale functional annotation of non-coding variants identified in schizophrenia GWAS and of human specific enhancers[8,9]. These data provide a rubric that illustrates the power of tissue-specific annotation of non-coding regulatory elements, as well as novel insights into the pathogenic mechanisms of neurodevelopmental disorders and the evolution of higher cognition.

Recent advances in high-throughput sequencing have unveiled the epigenomic landscape of multiple human cell types, as well as 3 dimensional folding principles of chromatin[10,11]. In particular, chromosome conformation capture experiments demonstrate that chromatin is organized into hierarchical structures, which include compartments (a few megabase (Mb))[1], topological associating domains (TADs, sub-Mb)[12], and loops (ranging from few kilobase (kb) to few hundred kb)[2,4]. These structures are thought to play a role in gene regulation and biological function by defining functional genomic units and mediating the effects of *cis*-regulatory elements via both short- and long-range physical interactions (e.g. promotor-enhancer interactions), relationships that cannot simply be predicted by linear adjacency in chromosomes. Coupled with epigenomic data, such higher order chromatin interactions should facilitate systemic annotation of *cis*-regulatory elements, as well as intergenic and intronic variants, which will further expand our understanding of tissue specific developmental programs, as well as disease pathogenesis.

We constructed multiple Hi-C libraries in mid gestation fetal cerebral cortex from three individuals during the peak of neurogenesis and migration (gestation week, GW17-18). We reasoned that it would be useful to analyze mitotically active neuronal precursors involved in neurogenesis separately from post-mitotic migrating and maturing neurons, so we dissected the cortical anlage into two major structures: the cortical and subcortical plate (CP), consisting primarily of post mitotic neurons and the germinal zone (GZ), containing primarily mitotically active neural progenitors (representative heatmap in **Fig. 1a**, **Extended Data Fig. 1a-b**). For comparison with non-neuronal cell types, we also used publicly available Hi-C data on human embryonic stem (ES) cells and IMR90 cells[11,12]. To provide grounding for our data and compare global chromosome architecture between different cell types, we performed principal component analysis (PCA)[13] on the genome-wide inter-chromosomal contact matrices of CP, GZ, ES, and IMR90. As previously demonstrated, global chromosome architecture does not change dramatically between different cell types[13]. However, the first principal components (PC1s) from neuronal tissues (CP and GZ) have significantly higher correlation than the PC1s between different cell types (**Fig. 1b**), consistent with the higher similarity between tissues from brain, versus the two other cell lines.

**3D chromatin structure reflects gene regulation during neural differentiation.**

Previous studies have shown that genome-wide chromosome conformation captures multiple levels of genomic features related to biological function, ranging from GC content and gene number to marks of open chromatin, such as DNase I hypersensitivity sites (DHS)[13]. Most human-relevant Hi-C has been conducted in cell lines[1,2,4,11,12,14] and not in complex tissue, such as developing brain. As an initial first step to insure the quality and validity of our data, we analyzed the relationships between the major component of the inter-chromosomal interaction matrix with these major genomic features, finding high correlation with GC content, gene number, DHS[10], and to a lesser extent, gene expression[15] (**Fig. 1c**, **Extended Data Fig. 2a**), as has been previously observed in non-neural cell lines[13].

To further explore the biological significance of chromosome contact changes during neural differentiation, we explored whether the genes in regions of dynamic chromatin structure were related to neural differentiation by comparing the inter-chromosome contact matrices (binned to 100kb) in different cell types and selecting bins with the highest chromatin contact count changes between two cell types (**Methods**). Genes located in the regions of highest inter-chromosomal interaction changes between CP and GZ were enriched for neuronal genes, represented by the gene ontology (GO) categories of neuron recognition, axon guidance, central nervous system (CNS)

development, and synapse (**Fig. 1d**, **Extended Data Fig. 2b**; **Methods**). Genes located in regions with highest inter-chromosomal interaction changes between CP and ES cells were enriched for developmental genes involved in forebrain development and chromatin organization (**Fig. 1d**, **Extended Data Fig. 2b**), indicating that these interactions reflect tissue relevant developmental gene regulation.

To further explore how these physical chromatin interactions relate to biological function, we hypothesized that highly interacting chromatin regions would be more likely to be co-regulated. To test this, we compared the distribution of correlation patterns for genes locating in (1) the regions of highest interaction values in both CP and GZ, (2) the lowest interacting regions in both CP and GZ, and (3) the regions of differential interaction values (the regions of highest interaction values in CP and lowest interaction values in GZ and vice versa). Highly interacting regions tend to be biased toward positive correlations, while there was no bias in correlation for low and differential interacting regions (**Fig. 1e**). Interestingly, the positive correlation for high interacting regions becomes even higher when more stringent cutoffs are used, supportive of the quantitative nature of interaction-driven co-expression, whereby the relationship between physical 3D chromatin interactions and expression is mostly driven by the top percentiles of interacting regions (**Extended Data Fig. 2c**). To further elucidate the epigenetic regulatory mechanisms behind the apparent interaction-mediated co-expression, we marked bins in which epigenetic marks from two loci appear together. By comparing the epigenetic mark combination matrix with the Hi-C contact matrix, we observed that interacting regions exhibit shared epigenetic patterns at the level of both inter- and intra-chromosomal interactions (**Fig. 1f**, **Extended Data Fig. 3**; **Methods**). In particular, regions associated with positive transcriptional regulation and enhancers are more likely to physically interact with each other, consistent with their co-regulation.

One of the core functional units of general genome organization recently uncovered by chromatin capture methods across a wide variety of cell types is the compartment, a relatively large, dynamic domain[1], which is comprised of smaller, sub-Mb regions of topologically associating domains (TADs)[12]. Compartments are divided into two types, type A compartments that consist primarily of euchromatin and actively transcribed genes and type B compartments, which are heterochromatic and repressed. TADs have been previously shown to be relatively stable, whereas compartments have been shown to change during lineage specification in stem cells[11]. Consistent with this, we observed dynamic compartment switching between CP and GZ, enriched for GO categories related to neuronal genes and phosphatase activity (**Fig. 2c**), as well as compartment switching between CP and ES (**Fig. 2a,d**). Genes that change compartments from ES to CP are decreased for A to B transitions across differentiation and increased for changes from the B to A compartments (**Fig. 2b**), as expected. Compartment changes are also accompanied by epigenetic changes, so that the B to A compartment shift is associated with increased DHS and active epigenetic marks indicative of open chromatin, whereas the A to B shift is associated with decreased DHS and increased repressive marks (**Fig. 2b,e**). The same pattern was observed for GZ vs. ES and CP vs. GZ (**Fig. 2b,e**, **Extended Data Fig. 2d**), demonstrating that gene expression changes across development are tightly linked to epigenetic changes coupled with compartment switching.

TADs are thought to mediate co-transcriptional regulation primarily within their boundaries (100kb-1Mb) through physical "looping" interactions of promotors and enhancers in co-regulated genes[4,16]. Since TAD boundaries are conserved across different cell types[12], we hypothesized that changes in epigenetic marks in TADs, rather than the boundaries of TADs, would be most associated with gene expression changes

across development. To test this, we divided genes based on their fold change in expression between ES and differentiated neurons[17] (both increased and decreased), and assessed changes in epigenetic marks within the TADs where these genes reside (**Extended Data Fig. 1c-e**, **Methods**). Notably, active marks including enhancers and elements related to transcribed regions are increased in TADs that contain upregulated genes, whereas repressive marks are increased in TADs that contain downregulated genes (**Fig. 2f**). Collectively, these results indicate that our Hi-C data reflects the major elements of global chromosome architecture in fetal brains, providing a framework for exploring gene regulatory mechanism related to human neural development and function.

Next, to demonstrate how knowledge of intra-chromosomal contacts could significantly advance understanding of important gene regulatory relationships in the nervous system, we performed two integrative experiments. In the first, we used these chromatin contact data to functionally annotate specific non-coding regulatory elements in the developing brain. We leveraged recent efforts that have identified >2000 developmental enhancers gained specifically in the human cerebral cortex, providing a remarkable resource for understanding the evolution of human cognition[8]. Usually, in the absence of such tissue specific data, regulatory elements are assigned to the closest gene[18,19], a convention that we compared with our Hi-C derived interactions. We reasoned that our Hi-C data from fetal brain could be used to identify the target genes for many of these enhancers, which based on previously chromatin looping analyses in cell lines are often not the closest gene[4,16,18,19].

We derived an interaction map of human-gained enhancers, defined as significant interacting regions (at a 1% false discovery rate, FDR) compared to the null distribution generated by fitting the contact frequencies of all fetal brain enhancers identified in the same study[8] (**Extended Data Fig. 4a**, **Methods**). We defined the search space as including the 1Mb flanking regions, since most enhancer-promoter interactions are within this range[4]. Although statistically significant interactions are increased upon proximity to the enhancer, the majority of interactions are at relatively long-ranges (>100kb, **Extended Data Fig. 4b**) and are not restricted to the adjacent genes. Indeed, ~65% of the closest genes to human-gained enhancers are not identified through fetal brain Hi-C interactions, revealing that the majority of enhancers are not interacting with the most adjacent gene (**Fig. 3c**). Compared to the original study[8], which used human-gained enhancer hotspot TADs in ES cells and IMR90 cells due to the lack of Hi-C data from relevant tissue, our approach provides genes of action with higher resolution in the matching tissue (fetal cortices) from which evolutionary enhancers were identified. Human-gained enhancer-interacting regions were enriched with enhancers, promoters, and transcription start sites (TSSs) (**Fig. 3a**, **Extended Data Fig. 4c**), consistent with the previous findings that enhancers interact with promoters, as well as other enhancers[16]. The majority of interactions (>75%) were in the same TADs (**Fig. 3b**), also consistent with observations in cell lines that most enhancer-promoter interactions are in the same TAD[16,19]. Human-gained enhancer interacting genes (Hi-$C_{evol}$ genes) are involved in GTPase regulation as well as G-protein coupled receptor (GPCR) and CREB signaling, and are enriched with GO terms representing synaptic and axon guidance genes (**Fig. 3e**, representative interactions in **Fig. 3d**). One striking example is a human-gained enhancer that interacts with *ARHGAP11B*, a human-specific gene implicated in the expansion of human neocortex[20] (**Fig. 3d**).

Given the high conservation of protein-coding genes across the vertebrate lineage, comparative genomics have suggested that human-specific traits most likely result from changes in regulatory elements[8,21]. Indeed, protein-coding Hi-$C_{evol}$ genes have a lower

non-synonymous substitution (dN)/synonymous substitution (dS) ratio compared to Hi-C non-interacting protein-coding genes in multiple lineages (**Extended Data Fig. 5**). These results indicate that human-gained enhancers are interacting with protein-coding genes that undergo purifying selection, further supporting the hypothesis that non-coding elements undergo evolutionary selection to induce species-specific changes in gene expression[8,21]. We also investigated whether human-gained enhancers are interacting with lineage-specific long non-coding RNAs (lncRNAs)[22]. We observed that lineage-specific interactions with human-gained enhancers were enriched for primate-specific lncRNAs, as well as evolutionary conserved lncRNAs (**Fig. 3f**, **Extended Data Fig. 5**). Thus, while human-gained enhancers interact and possibly regulate evolutionary conserved protein-coding genes, they are more likely to interact with primate-specific lncRNAs.

Since the development of human higher cognition is dependent on the development of the human cerebral cortex via elaboration of novel gene regulatory relationships[8,23], we reasoned, as have others[8] that the genes regulated by these human specific enhancers would be associated with intellectual functioning in humans. Remarkably, we found that the Hi-$C_{evol}$ genes in fetal brain, but not the genes defined by proximity to the enhancers are significantly enriched with intellectual disability (ID) risk genes[6] (**Fig. 3g**). This result provides experimental support for the contention that human-gained enhancers are associated with the evolution of human cognitive function[8]. This enrichment was tissue-specific, as Hi-$C_{evol}$ genes defined by Hi-C interactions in ES cells did not show enrichment for ID risk genes (**Fig. 3g**). Indeed, ~56% of the Hi-$C_{evol}$ genes in neuronal tissue were not identified through chromatin contacts in ES cells, emphasizing the importance of defining tissue-relevant chromatin contacts, as well as importance of using the relevant tissue for Hi-C analysis (**Fig. 4c**).

Since most disease related common genetic variation is located in non-protein coding regions, we next assessed the ability of Hi-C data for functional annotation of common single nucleotide polymorphisms (SNPs). As a first line verification that Hi-C data could identify known functional relationships between SNPs and gene expression we used *cis*-expression quantitative trait loci (eQTL) data from adult frontal cortex[24], since such data is not yet available from fetal brain. For each significant eQTL locus, we obtained a set of significant eQTL SNPs with >95% likelihood of containing the causal SNP from association statistics and linkage disequilibrium (LD; 1000 Genomes) structure using CAVIAR[25]. We then identified genes interacting to likely causal eQTL SNPs via the chromatin contact matrix (Hi-$C_{eQTL}$ genes, **Methods**), and compared Hi-$C_{eQTL}$ genes with the known associated gene from the eQTL study, finding that Hi-$C_{eQTL}$ genes significantly overlapped with eQTL transcripts (**Extended Data Fig. 6a**). There were many Hi-$C_{eQTL}$ genes that were not identified as eQTL transcripts, likely due to a combination of factors, including low power of the eQTL sample, limited resolution of Hi-C (SNP-transcript interactions within 20kb cannot be detected), and the difference in age of tissues used for each analysis. Indeed, eQTL SNPs identified by CAVIAR were highly enriched with adult frontal cortex, but not fetal brain, enhancers (**Extended Data Fig. 6b-d**). Despite this, eQTL SNP-transcript pairs exhibit higher chromatin contact frequency than expected by chance across all distance ranges (**Extended Data Fig. 6e**), further supporting the utility of Hi-C to infer the biological function of regulatory variation.

Next, we applied a similar logic to advance our understanding of 108 genome-wide significant schizophrenia-associated loci, most of which are in relatively uncharacterized non-coding regions of the genome[9]. We obtained credible SNPs using CAVIAR, and split SNPs into those without known function and likely functional SNPs (SNPs that cause missense, frameshift, and splice variants and SNPs that fall onto gene promoters;

Methods). Credible SNPs were enriched with enhancers in fetal brain and adult frontal cortex, confirming the likely regulatory role of these SNPs in the brain (**Extended Data Fig. 7**). SNPs defined as likely functional SNPs and promoter SNPs were directly assigned to their target genes. For the remaining intergenic and intronic SNPs that were un-annotated, and therefore without clear function, we used the chromatin contact matrix to find genes with which the regions where the SNPs are located are physically interacting (diagram in **Extended Data Fig. 7**).

Combining genes annotated as functional SNPs, promoter SNPs, and by Hi-C interactions, we obtained a total of ~900 genes (Hi-$C_{SCZ}$ genes) associated with schizophrenia risk variants. Hi-C contacts identified numerous genes that were neither adjacent to index SNPs nor in LD with them (**Fig. 4a-c**, **Extended Data Fig. 9**). While almost 70-80% of the LD genes and closest genes were identified as Hi-$C_{SCZ}$ genes, only half of them were identified by chromatin contacts, indicating that many of them were identified by functional SNPs residing in the genes. Moreover, 70-90% of the Hi-$C_{SCZ}$ genes were not identified by using LD genes or the closest genes to the association signal, consistent with observations that the linear organization of genes and regulatory elements on the chromosome does not reflect regulatory interactions[4,18,19].

Hi-C analysis showed that schizophrenia-associated common variants converge into specific molecular pathways related to neuronal function, including the postsynaptic density, acetylcholine receptors, cell cycle, and chromatin remodelers (**Fig. 4d-e**, **Extended Data Fig. 7-8**). To insure that this was not an artifact of the method used for credible SNP selection, we used a different method to define the set of credible SNPs[9] (**Extended Data Fig. 9**) and found the same enrichments, demonstrating the robustness of the genes identified through the Hi-C analysis. One notable example is illustrated by credible SNPs (rs4245150, rs17602038, rs4938021, rs4936275, rs4936276) that reside upstream of the *Dopamine D2 Receptor* (*DRD2*), the target of antipsychotic drugs. Although these SNPs are close to the *DRD2* TSS, they are not within the gene, which complicates interpretation of their biological function. Hi-C analysis demonstrates for the first time that indeed these SNPs are interacting with the TSS of *DRD2* (**Fig 4e**), providing biological insights into the function of these SNPs.

Another relevant example is an index SNP (rs79212538) interacting with *GRIA1*, an ionotropic glutamate receptor subunit, although *GRIA1* is neither the closest gene nor in LD with the index SNP (**Extended Data Fig. 8**). Additionally, Hi-C shows that schizophrenia associated non-coding SNPs interact with multiple genes involved in excitatory synaptic transmission, including *CACNA1C*, *GRIN2A*, and *NLGN4X*, further supporting glutamatergic transmission defects in schizophrenia pathophysiology (**Extended Data Fig. 8**). Interestingly, Hi-$C_{SCZ}$ genes significantly overlap with ASD *de novo* likely gene-disrupting (LGD) targets (CP: OR=2.4, P=$1.6 \times 10^{-5}$, GZ: OR=1.8, P=0.006), consistent with a shared genetic etiology between ASD and schizophrenia[26]. The fact that genes with LGD mutations in ASD are associated with regulatory variants in schizophrenia suggests that complete abrogation of these genes may cause developmental defects as in ASD, while regulatory changes in these genes may cause later-onset of neuropsychiatric symptoms as in schizophrenia. Collectively, genes annotated by chromatin contact information provide novel insights into schizophrenia pathogenesis.

In conclusion, we demonstrate how a comprehensive analysis of genome-wide chromatin configuration during human neural development informs our view of gene regulation. This chromatin contact landscape provides important biological insights on gene regulatory mechanisms, such that co-expressed genes share epigenetic co-regulation of interacting regions, and that changes in functional epigenetic marks are tightly linked to TADs and compartment switching to induce changes in gene expression.

We also annotated non-coding regulatory elements in the genome based on long-range chromatin contacts to identify enhancer-promoter interactions during human brain development, as well as genes of actions for eQTL. In turn, we show how these interactions can be used to inform our biological interpretation of risk variants for schizophrenia, which serves as a template for understanding the role of non-coding variation more broadly in neuropsychiatric disorders.

## Methods

### Fetal brain layer dissection

Human fetal cortical tissues from three individuals were collected from frontoparietal cortex at gestation week (GW) 17-18 (one sample from GW17 and two samples from GW18). In cold DMEM/F-12 (ThermoFisher, 11320-033), frontoparietal cortex was first dissected to thin (~1mm) slices to visualize layers. Under the light field microscope, cortical slice was dissected to germinal zone (GZ) and cortical plates (CP). GZ contains ventricular zone and subventricular zone, and hence comprised of proliferating neurons. CP refers to intermediate zone, cortical plate, and marginal zone, which are mainly composed of differentiated and migrating neurons. By dissecting layers from same fetal cortices, we can compare progenitors to differentiated neurons with same genotype and minimize intersample heterogeneity.

### Hi-C

Collected tissue was dissociated with trypsin and cell number was counted. Ten million cells were fixed in 1% formaldehyde for 10 min. Cross-linked DNA was digested by restriction enzyme HindIII (NEB, R0104). Digested chromatin ends were filled and marked with biotin-14-dCTP (ThermoFisher, 19518-018). Resulting blunt-end fragments were ligated under dilute concentration to minimize random intermolecular ligations. DNA purified after crosslinking was reversed by proteinase K (NEB, P8107) treatment. Biotins from unligated ends were removed by exonuclease activity of T4 DNA polymerase (ThermoFisher, 18005). DNA was sheared by sonication (Covaris, M220) and 300-600bp fragments were selected. Biotin-tagged DNA, which is intermolecular ligation products, was pulled down with streptavidin beads (Invitrogen, 65001), and ligated with Illumina paired end adapters. Resulting Hi-C library was amplified by PCR (KAPA Biosystems HiFi HotStart PCR kit, KK2502) with the minimum number of cycle (typically 12-13 cycles), and sequenced by Illumina 50bp paired-end sequencing.

### Hi-C reads mapping and pre-processing

Note that mapping and filtering of the reads, as well as normalization of experimental and intrinsic biases of Hi-C contact matrices were conducted with the following method regardless of cell types to minimize potential variance in the data obtained from different platforms. We implemented *hiclib* (https://bitbucket.org/mirnylab/hiclib) to perform initial analysis on Hi-C data from mapping to filtering and bias correction. Briefly, quality analysis was performed using a phred score, and sequenced reads were mapped to hg19 human genome by *Bowtie2* (with increased stringency, *--score-min -L 0.6,0.2--very-sensitive*) through iterative mapping. Read pairs were then allocated to HindIII restriction enzyme fragments. Self-ligated and unligated fragments, fragments from repeated regions of the genome, PCR artifacts, and genome assembly errors were removed. Filtered reads were binned at 10kb, 40kb, and 100kb resolution to build a genome-wide contact matrix at a given bin size. This contact map depicts contact frequency between any two genomic loci. Biases can be introduced to contact matrices by experimental procedures and intrinsic properties of the genome. To decompose biases from the contact matrix and yield a true contact probability map, filtered bins were subjected to iterative correction[13], the basic assumption of which is that each locus has uniform coverage. Bias correction and normalization results in a corrected heatmap of bin-level resolution. 100kb resolution bins were assessed for inter-chromosomal interactions, 40kb for TAD analysis, and 10kb for gene loop detection.

### Inter-chromosomal principal component analysis

Principal component analysis (PCA) was conducted in a genome-wide inter-chromosome contact map (100kb binned) as described previously[13]. Since intra-

chromosome conformation may drive the PCA results, *cis* contacts were iteratively replaced to random *trans* counts. After removing diagonal and poorly covered regions, we performed PCA using *hiclib* command *doEig*.

Pearson's correlations between the first principal components (PC1) from different cell types (CP, GZ, ES, and IMR90[12]) were calculated to compare similarities in inter-chromosomal interactions between different cell types.

Spearman's correlations between PC1/PC2 and biological traits (GC content, gene density, DNase I hypersensitivity (DHS), gene expression) were calculated. GC content (%) for each 100kb bin was calculated by *gcContentCalc* command from R package *Repitools*. Gene density (number of genes in 100kb bin) was obtained based on longest isoforms from GENCODE19. DHS of fetal brains from Epigenomic roadmap[10] and gene expression level of prenatal cortical layers from Miller et al.[15] were used and average values per 100kb bin were calculated.

### Gene enrichment analysis

Gene ontology (GO) enrichment was performed by GO-Elite Pathway Analysis (http://www.genmapp.org/go_elite/). All genes in the genome except the ones located in the chromosome Y and mitochondrial DNA were used as a background gene list. Because Hi-C interaction is measured in bins, sometimes we cannot dissect the individual genes when they are clustered in the genome (i.e. PCDH locus). To prevent several gene clusters overriding entire GO terms, we removed GO mainly defined by gene clusters (for 100kb or 40kb binned data) or we randomly included one gene per cluster (e.g. PCDHA1 for PCDHA1-13 cluster) prior to GO analysis (for 10kb binned data).

Gene enrichment for the curated gene lists was performed using binomial generalized linear model to regress out exome length. Autism spectrum disorder (ASD) *de novo* gene list and intellectual disability (ID) curated gene list from Iossifov et al.[27] and Pariskshak et al.[6] were used for the enrichment test, respectively. Protein-coding genes based on biomaRt were used as a background gene list.

### Identification of the regions with largest inter-chromosomal conformation changes

Chromosome contact matrix was normalized with the total interaction counts between two cell types for comparison. Intra-chromosomal interactions were masked from the genome-wide contact matrix, and top 1000 bins with the largest interaction changes between different cell types (GZ vs. CP or ES vs. CP) were selected. As one bin is comprised of two loci that are interacting with each other, this would give ~2000 sites in the genome. Genes located in those ~2000 sites were combined to perform GO analysis.

### Co-expression of inter-chromosomal interacting regions

Using transcriptome from fetal cortical layers[28], average expression values per 100kb bin were calculated. Pearson correlation matrix was calculated from 100kb binned expression data from all layers to generate gene co-expression matrix. At this step, gene co-expression matrix has the same dimension as inter-chromosomal contact matrix.

We hypothesized that genes would be co-expressed across the layers when they are interacting in all stages (both in CP and GZ), so we selected top 2% highest interacting regions of fetal brains both at GZ and CP (high interacting regions). We also selected (1) low interacting regions: top lowest interacting regions (0 interaction from normalized Hi-C contact matrix) of fetal brains both at GZ and CP, as well as (2) variant interacting regions: top 2% highest interacting regions from one stage (e.g. GZ) that are top 2%

lowest interacting regions from the other stage (e.g. CP) for comparison. Expression correlation values of the same regions were selected from the gene co-expression matrix, and expression correlations between different states (high interacting regions vs. low interacting regions and high interacting regions vs. variant interacting regions) were compared by two-sample Kolmogorov-Smirnov test.

### Epigenetic state enrichment for inter-chromosomal interacting regions

The fetal brain epigenetic 25 state model from Epigenomic roadmap[10] was used to generate the epigenetic state combination matrix, which was generated by marking loci where two interacting chromosomal bins (defined as bins with (1) interaction counts > 75% quantile interaction count for inter-chromosome and (2) interaction counts > 0 for intra-chromosome) share epigenetic signature. For example, the epigenetic combination matrix between the active transcription start site (TssA) and active enhancers (EnhA1) was generated by marking where interacting loci have TssA on one locus and EnhA1 on the other locus. Intra- and inter-chromosomal contact frequency maps were then compared to epigenetic state matrix by Fisher's exact test to calculate enrichment of shared epigenetic combinations in interacting regions.

### Compartment analysis

Expected interaction frequency was calculated from the normalized intra-chromosomal 40kb binned contact matrix based on the distance between two bins. We summed series of submatrices of 400kb window size with 40kb step size from the normalized Hi-C maps to generate observed and expected matrices. The Pearson's correlation matrix was computed from the observed/expected matrix, and PCA was conducted on correlation matrix. PC1 from each chromosome was used to identify compartments. Eigenvalues positively correlated with the gene density were set as compartment A, while those that are negatively correlated were set as compartment B.

### Gene expression and epigenetic state change across different compartments

Genomic regions were classified into three categories according to compartments: compartment A in cell type1 that changes to compartment B in cell type2 (A to B), compartment B in cell type1 that changes to compartment B in cell type2 (B to A), regions that do not change compartment between two cell types (stable).

Genes residing in each compartment category were selected and GO enrichment was performed. Gene expression fold-change (FC) between different cell types was calculated from Miller et al.[15] (comparison for CP vs. GZ) and CORTECON[17] (comparison for ES vs. CP and ES vs. GZ). Distribution of gene expression FC for genes in different compartment categories was compared by one-way ANOVA and Tukey's posthoc test.

15 state epigenetic marks from Epigenomic Roadmap[10] in genomic regions classified based on compartments were averaged across 40kb bins. The DHS FC[10] between different cell types (ES vs. CP and ES vs. GZ) was calculated and statistically evaluated as in the gene expression comparison. Each epigenetic state counts[10] for one compartment category was normalized by total epigenetic mark number of that compartment category and compared between ES and fetal brains.

### TAD analysis

We conducted TAD-level analysis as described previously[12]. Shortly, we quantified the directionality index by calculating the degree of upstream or downstream (2Mb) interaction bias of a given bin, which was processed by a hidden Markov model (HMM) to remove hidden directionality bias.

Regions in between TADs are titled as TAD boundaries when the regions are smaller than 400kb and unorganized chromatin when the regions are larger than 400kb.

### TAD-based epigenetic changes upon differentially expressed genes

Genes were subdivided into 20 groups based on expression FC between ES and most differentiated neuronal states in CORTECON[17]: genes that are upregulated and downregulated upon differentiation were grouped into 10 quantiles, respectively, based on the FC. TADs into which genes from one subdivision reside were selected, and epigenetic state changes (from Epigenomic roadmap's 15 state epigenetic marks in ES and fetal brains[10]) in those TADs were normalized with TAD length and compared between ES and fetal brains. As different types of epigenetic marks have different absolute numbers (e.g. there are more quiescent states than enhancer states in the genome), each epigenetic state change was scaled across different quantiles to allow comparison between different states.

### Identification of Hi-C interacting regions

We identified Hi-C interacting regions and target genes for (1) human-gained enhancers[8], (2) expression quantitative trait loci (eQTL) SNPs[24], and (3) schizophrenia SNPs[9]. As the highest resolution available for the current Hi-C data was 10kb, we assigned these enhancers/SNPs to 10kb bins, obtained Hi-C interaction profile for 1Mb flanking region (1Mb upstream to 1Mb downstream) of each bin. We also made a background Hi-C interaction profile by pooling (1) 255,698 H3K27ac sites from frontal and occipital cortex at 12 PCW for human-gained enhancers[8] and (2) 9,444,230 imputed SNPs for eQTL and schizophrenia SNPs[9]. To avoid significant Hi-C interactions affecting the distribution fitting as well as parameter estimation, we used the lowest 95 percentiles of Hi-C contacts and removed zero contact values. Using these background Hi-C interaction profiles, we fit the distribution of Hi-C contacts at each distance for each chromosome using *fitdistrplus* package (**Extended Data Fig. 4a**). Significance for a given Hi-C contact was calculated as the probability of observing a stronger contact under the fitted Weibull distribution matched by chromosome and distance. P-values were adjusted by computing FDR, and Hi-C contacts with FDR<0.01 were selected as significant interactions. Significant Hi-C interacting regions were overlapped with GENCODE19 gene coordinates (including 2kb upstream to transcription start sites (TSS) to allow detection of enhancer-promoter interactions) to identify interacting genes. Same analysis was performed on Hi-C contact maps from CP, GZ, and ES[11]. To address the functional significance of target genes, GO enrichment was performed for the interacting genes.

### Protein-coding genes interacting with human-specific evolutionary enhancers

Protein-coding genes based on biomaRt (GENCODE19) were selected and non-synonymous substitution (dN)/synonymous substitution (dS) ratio was calculated for homologs in mouse, rhesus macaque, and chimpanzee for representation of mammals, primates, and great apes, respectively. Log2(dN/dS) distributions for protein-coding genes interacting vs. non-interacting to human-specific evolutionary enhancers in each lineage were then compared by two-sample Kolmogorov-Smirnov test.

### LncRNAs interacting with human-specific evolutionary enhancers

Long non-coding RNAs (lncRNAs) classified according to evolutionary lineages[22] were used to assess whether lineage-specific lncRNAs are interacting to human-specific evolutionary enhancers. We randomly selected the same number of enhancers (2,104) to the human-specific ones from the total enhancer pool (255,698), identified interacting regions based on the null distribution generated from a background enhancer interaction profile. Significant interacting regions (FDR<0.01) identified by Hi-C were intersected

with lncRNA coordinates[22] and interacting lncRNAs for each lineage were counted. This step was repeated for 3,000 times to obtain the lncRNA lineage distribution. LncRNAs interacting with human-specific evolutionary enhancers were also identified and enrichment was tested by calculating P-values as the probability of observing more interacting lncRNAs for a given lineage under the null lncRNA lineage distribution.

## Epigenetic state enrichment for Hi-C interacting regions

The functional framework for (1) eQTL SNPs, (2) schizophrenia SNPs, and (3) human-gained enhancers-interacting regions was assessed for epigenetic state enrichment. We implemented the same approach as in GREAT[29] to analyze the epigenetic state enrichment for *cis*-regulatory regions. For example, to evaluate whether schizophrenia SNPs are enriched with DHS, fraction of genome annotated with DHS (p), the number of schizophrenia SNPs (n), and number of schizophrenia SNPs overlapping with DHS (s) were calculated. Significance of the overlaps was tested by binomial probability of $P = Pr_{binom}\ (k \geq s \mid n = n,\ p = p)$[29]. Histone marks and 15-chromatin states from fetal brains, adult frontal cortex, and IMR90[10] were used for epigenetic state enrichment.

## eQTL analysis

To address whether co-localization mediates gene regulation, we compared the association between chromosome conformation with eQTL. Although fetal brain eQTL data would be optimal, since this data is currently not available, we analyzed adult frontal cortex *cis*-acting eQTL data[24]. We selected SNPs associated with gene expression (FDR<0.01) and clustered them with association $P<1\times10^{-5}$ and $r^2>0.6$ to obtain index SNPs. Using summary association statistics and linkage disequilibrium (LD) structure for each index SNP, we applied *CAVIAR*[25] to quantify the probability of each variant to be causal. Among 121,273,364 SNP-transcript pairs from frontal cortex eQTL data, this process resulted in 42,190 SNP-transcript pairs (267 transcripts and 14,882 SNPs) that are potentially credible. We refer to 14,882 credible SNPs as credible SNPs. Credible SNP interacting genes were identified as described in “identification of Hi-C interacting regions” section.

Fisher’s exact test was performed to evaluate the significance of the overlap between Hi-C interacting genes and eQTL transcripts. The background gene list for Fisher’s exact test includes genes located in 1Mb flanking regions to credible SNPs that are also tested in eQTL analysis.

For 42,190 SNP-transcript pairs, we assigned credible SNPs and genes into 10kb bins, and obtained Hi-C contacts between credible SNPs and genes from the 10kb binned Hi-C contact maps. As a gene can span across multiple 10kb bins, the highest interaction in the gene to a credible SNP was selected as Hi-C contacts as previously defined[30]. We also calculated expected interaction frequency from the normalized 10kb binned contact matrix based on the distance between two bins. Opposite interaction frequency was calculated by obtaining Hi-C contacts for the opposite site to the credible SNP with the same distance. Because interaction counts differ in different chromosomes as well as in different cell types, we normalized interaction by chromosomes and cell types. We performed one-way ANOVA and Tukey’s posthoc test for the comparison between different interaction paradigms.

## Identification of credible SNPs for schizophrenia GWAS loci

128 LD-independent SNPs with genome-wide significance ($P<5\times10^{-8}$)[9] were used as index SNPs to obtain schizophrenia credible SNPs. All SNPs that are associated with $P<1\times10^{-5}$ and in LD ($r^2>0.6$) with an index SNP were selected, and correlations among this set of SNPs (LD structure) were calculated. CAVIAR was applied to summary association statistics and LD structure for each index SNP, and potentially causal SNPs

for each index SNP were identified. Among 55,000 SNPs that are in LD with 128 index SNPs, 7,613 SNPs were selected as causal by CAVIAR. Here we refer to these CAVIAR-identified SNPs as credible SNPs. Genes interacting to credible SNPs were identified as described in "identification of Hi-C interacting regions" section for CP, GZ, and ES. A separate set of credible SNPs initially reported from the original study was also processed with the same method[9].

## Identification of schizophrenia GWAS SNP-associated genes

We classified credible SNPs based on potential functionality (flow chart in **Extended Data Fig. 7**). For credible SNPs classified as functional (stop gained variant, frameshift variant, splice donor variant, NMD transcript variant, and missense variant) from biomaRt, we selected genes in which those SNPs locate. For those that are not directly affecting the gene function, we selected SNPs that fall onto the promoter and TSS of genes (2kb upstream-1kb downstream to TSS). Remaining SNPs were tested for Hi-C interaction so that Hi-C interacting genes were identified. This pipeline gives total ~900 genes potentially associated with GWAS SNPs.

## Identification of closest genes and LD genes

Closest genes to human-gained enhancers and schizophrenia index SNPs were obtained by *closestBed* command from *bedtools*. Gene coordinates from GENCODE19 including 2kb upstream to TSS were used to identify the closest genes.

LD genes refer to all genes in the LD. Here, LD is defined as physically distinct schizophrenia-associated 108 genome-wide significant regions[9]. We overlapped gene coordinates from GENCODE19 with LD regions to find genes that reside in LD.

Closest genes and LD genes were compared with Hi-C interacting genes. Venn diagrams were generated by *Vennerable* package in R. Only protein-coding genes were included in plotting Venn diagrams.

## Calculation of distance between SNPs and genes

For LD genes and closest genes, the shortest distance between an index SNP and a target gene was selected. For credible SNPs, (1) the distance between functional credible SNPs and target genes was set as 0, because functional SNPs reside in the gene, (2) the distance between promoter credible SNPs and target genes was calculated as the distance between SNPs and TSS of a gene, (3) the distance between credible SNPs and Hi-C interacting genes was calculated based on the distance between SNPs and Hi-C interacting bins (note that this distance has a unit of 10kb). We then combined the distance distributions from the 3 categories.

## Figure Legends

**Figure 1. Chromosome conformation in fetal brains reflects genomic features. a.** Representative heatmap of the chromosome contact matrix of CP. Normalized contact frequency (contact enrichment) is color-coded according to the legend on the right. **b.** Pearson correlation of the leading principle component (PC1) of inter-chromosomal contacts at 100kb resolution between *in vivo* cortical layers and non-neuronal cell types (ES and IMR90). **c.** Spearman correlation of PC1 of chromatin interaction profile of fetal brain (GZ) with GC content (GC), gene number, DNase I hypersensitivity (DHS) of fetal brain, and gene expression level in fetal laminae. **d.** GO enrichment of genes located in the top 1000 highly interacting inter-chromosomal regions specific to CP vs. GZ (left), and CP vs. ES (right), indicating that genes located on dynamic chromosomal regions are enriched for neuronal development. **e.** The top 2% highest interacting regions of fetal brains both at GZ and CP (High) show positive correlation in gene expression, while the top 2% lowest interacting regions (Low) and top 2% highly variant regions (Variant) have no skew in distribution. P-values from Kolmogorov–Smirnov test. **f.** The epigenetic state combination in inter-chromosomal interacting regions in GZ. Inter-chromosomal contact frequency map is compared to epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1), transcriptional regulators (TxReg), and transcribed regions (Tx) interact highly to each other as marked in red. Colored bars on the left represent epigenetic marks associated with promoters and transcribed regions (orange), enhancers (red), and repressive marks (blue). Chr, chromosome. Annotation for epigenetic marks described in

http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Figure 2. Compartment and TADs provide insights into gene regulatory mechanism. a.** Leading principal component (PC1) of the intra-chromosomal contact matrix in CP, GZ, and ES, with the DNase I hypersensitivity (DHS) fold change (FC) between ES and fetal brain (FB). PC1 values indicate compartment status of a given region, where positive PC1 represents compartment A (red), and negative PC1 represents compartment B (green). **b.** Distribution of gene expression FC (left) and DHS FC (right) for genes/regions that change compartment status ("A to B" or "B to A") or that remain the same ("stable") in different cell types. P-values from one-way ANOVA. **c.** GO enrichment of genes that change compartment status from A to B (top) and B to A (bottom) in CP to GZ. **d.** Heatmap of PC1 values of the genome that change compartment status in different cell types. **e.** Percentage of epigenetic marks for genomic regions that change compartment status between ES and CP. Note that B to A shift in ES to CP is associated with increased proportion of active transcribed regions (TssA and Tx) and enhancers (Enh, top), while A to B shift in ES to CP is associated with increased proportions of repressive marks (Het and ReprPCWk, bottom). P-values from Fisher's exact test. **f.** Epigenetic changes in topological associating domains (TADs) mediate gene expression changes during neuronal differentiation. Genes were divided by expression FC between ES and differentiated neurons, and epigenetic marks in the TADs containing genes in each group were counted and compared between ES and CP. Upregulated genes in neurons locate in TADs with more active epigenetic marks in CP than in ES, while downregulated genes in neurons locate in TADs with more repressive marks in CP than in ES. Epigenetic states associated with activation and transcription of the genes were marked as a red bar, while those associated with repression were marked as blue bars on the right. Annotation for epigenetic marks

described in http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Figure 3. Genetic architecture of human-gained enhancers. a.** Fraction of epigenetic states for regions interacting to human-gained enhancers in CP and GZ. **b.** Proportions of whether human-gained enhancers and interacting regions are within the same topological associating domain (TAD) vs. outside of the TAD. **c.** Overlap between human-gained enhancer interacting genes (Hi-$C_{evol}$ genes) in CP and GZ with closest genes to human-gained enhancers (left) and Hi-$C_{evol}$ genes in ES (right). **d.** Representative interaction map of a 10kb bin, in which human-gained enhancers reside, with the corresponding 1Mb flanking regions. This interactome map provides genes of action that interact with human-gained enhancers. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; Evol, genomic coordinate for a 10kb bin in which human-gained enhancers reside; -log10(P-value), P-value for the significance of the interaction between human-gained enhancers and each 10kb bin, grey dotted line for FDR=0.01; TAD, TAD borders in CP, GZ, and ES. **e.** GO enrichment for Hi-$C_{evol}$ genes in CP (left) and GZ (right). **f.** Number of primate-specific long non-coding RNAs (lncRNAs) interacting with human-gained enhancers in CP (red vertical lines in the graph) against a background control generated from 3,000 permutations, where the number of lncRNAs interacting with the same number of enhancers pooled from all fetal brain enhancers was counted. **g.** Overrepresentation of Hi-$C_{evol}$ genes in different tissues and closest genes with a curated set of intellectual disability (ID) risk genes. *P<0.05, **P<0.01, *** P<0.001. TSS, transcription start site; OR, odds ratio; GPCR, G-protein coupled receptor; Hi-C genes: GZ, CP, ES, Hi-$C_{evol}$ genes in each tissue; Hi-C genes: FB, union of Hi-$C_{evol}$ genes in GZ and CP; Hi-C genes: ES-specific, Hi-$C_{evol}$ genes in ES but not in fetal brain (FB); Hi-C genes: FB-specific, Hi-$C_{evol}$ genes in FB (union) but not in ES; Closest genes, closest genes to human-gained enhancers.

**Figure 4. Annotation of significant chromatin interactions for schizophrenia-associated loci. a.** Overlap between closest genes to index SNPs (Closest), genes locating in linkage disequilibrium (LD), and genes identified through SNP categorization and chromatin contacts in CP and GZ (Hi-$C_{SCZ}$ genes, diagram in **Extended Data Fig. 7**). **b.** Number of closest genes and LD genes that interact to credible SNPs (Hi-C supported) and those that do not interact to credible SNPs (Hi-C non-supported, top). Number of genes that interact to credible SNPs that are closest to or in LD with index SNPs (Hi-C genes), and not closest to or in LD with index SNPs (Hi-C genes not, bottom). Note that Hi-C genes here contain physically interacting genes, but not genes identified by functional SNPs or promoter SNPs. **c.** Distance between CAVIAR/index SNPs and their target genes for closest genes to index SNPs (Closest), genes locating in linkage disequilibrium (LD), and Hi-$C_{SCZ}$ genes in CP (CP) and GZ (GZ) **d.** GO enrichment for Hi-$C_{SCZ}$ genes in CP (left) and GZ (right). **e.** Representative interaction map of a 10kb bin, in which credible SNPs reside, to the corresponding 1Mb flanking regions. This interactome provides target genes interacting to credible SNPs-containing region. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; SNP, genomic coordinate for a 10kb bin in which credible SNPs locate; -log10(P-value), P-value for the significance of the interaction between credible SNPs and each 10kb bin, grey dotted line for FDR=0.01; GWAS loci, LD region for the index SNP; TAD, topological associating domain borders in CP, GZ, and ES.

## Acknowledgements

This work was supported by NIH grants to D.H.G. (5R01MH060233:05; 5R01MH100027) and J.L.S (K99MH102357), Glenn/AFAR Postdoctoral Fellowship Program for Translational Research on Aging (20145357) and Basic Science Research Program through the National Research Foundation of Korea (NRF) (2013024227) to H.W., the California Institute for Regenerative Medicine (CIRM)-BSCRC Training Grant to L.T.U. (TG2-01169). The Hi-C library was sequenced by Broad Stem Cell Research Center (BSCRC), and fetal tissue was collected from Center for Aids Research (CFAR, 5P30 AI028697). eQTL data was kindly provided by Mina Ryten and Adaiklavan Ramasamy. We thank Suhua Feng, Michael Gandal, Yuan Tian, Vivek Swarup, Grant Belgard for helpful discussions and critical reading of the manuscript.

## Author Contributions

H.W. designed and performed experiments, interpreted results, and co-wrote the manuscript. L.T.U. performed sample collection and experiments. J.L.S., N.N.P., and F.H. analyzed data. C.L. helped establishing Hi-C protocol. J.E. and E.E. participated in the discussion of the results. D.H.G. supervised the experimental design and analysis, interpreted results, provided funding, and co-wrote the manuscript.

## Author Information

*Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles*

Hyejung Won, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Changhoon Lee, Daniel H. Geschwind

*Department of Biological Chemistry, University of California California Los Angeles*

Jason Ernst

*Department of Computer Science, University of California Los Angeles*

Farhad Hormozdiari, Eleazar Eskin

*Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles*

Daniel H. Geschwind, Eleazar Eskin, Jason Ernst

## References

1 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

2 Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

3 Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).

4 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294, doi:10.1038/nature12644 (2013).

5 Network & Pathway Analysis Subgroup of Psychiatric Genomics, C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209, doi:10.1038/nn.3922 (2015).

6 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021, doi:10.1016/j.cell.2013.10.031 (2013).

7 Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007, doi:10.1016/j.cell.2013.10.020 (2013).

8 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159, doi:10.1126/science.1260943 (2015).

9 Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

10 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

11 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

12 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

13 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).

14 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64, doi:10.1038/nature12593 (2013).

15 Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206, doi:10.1038/nature13185 (2014).

16 Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-1065, doi:10.1016/j.cell.2015.07.048 (2015).

17 van de Leemput, J. *et al.* CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. *Neuron* **83**, 51-68, doi:10.1016/j.neuron.2014.05.013 (2014).

18 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

19 Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).

20 Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465-1470, doi:10.1126/science.aaa1975 (2015).

21 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).

22 Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635-640, doi:10.1038/nature12943 (2014).

23 Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647, doi:10.1016/j.neuron.2013.10.045 (2013).

24 Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428, doi:10.1038/nn.3801 (2014).

25 Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508, doi:10.1534/genetics.114.167908 (2014).

26 McCarthy, S. E. *et al.* De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658, doi:10.1038/mp.2014.29 (2014).

27 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221, doi:10.1038/nature13908 (2014).

28 Miller, J. A., Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12698-12703, doi:10.1073/pnas.0914257107 (2010).

29 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

30 Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic acids research* **42**, 87-96, doi:10.1093/nar/gkt857 (2014).

**Extended Data Figure 1. Basic characterization of Hi-C libary. a.** Hi-C library sequencing information. Percentage for double-stranded (DS) reads indicates percentage of DS reads to all reads, and percentage for valid pairs and filtered reads indicates percentage of valid pairs and filtered reads to DS reads. **b.** Frequency distribution of Hi-C contacts in GZ (left) and CP (right) **c.** Size distribution of topological associating domains (TADs) in GZ (left) and CP (right). **d.** Size distribution of genomic regions in between TADs that are less than 400kb (TAD boundaries) in GZ (left) and CP (right). **e.** Size distribution of genomic regions in between TADs that are bigger than 400kb (unorganized chromosome) in GZ (left) and CP (right). Cis ratio, ratio of cis (intra-chromosomal) reads to the total number of reads; chr, chromosome.

**Extended Data Figure 2. Chromosome conformation is associated with various genomic features.** a. Spearman correlation of principal components (PCs) of chromatin interaction profile of CP with GC content (GC), gene number, DNase I hypersensitivity (DHS), and gene expression level of fetal brains. **b.** GO enrichment of genes located in the top 1000 regions that gain inter-chromosomal interactions in CP compared to ES (upper left), ES compared to CP (upper right), CP compared to GZ (lower left), and GZ compared to CP (lower right). **c.** Top 5% (left) and 10% (middle) highest interacting regions both in GZ and CP (High) show positive correlation in gene expression, while low interacting regions (Low) and variant interacting regions (Variant) have no skew in distribution. (Right) Mean (top) and median (bottom) values for gene expression correlation for high, low, and variant interacting regions with different cutoffs, indicating that higher the interaction, higher the correlation of gene expression. **d.** Percentage of epigenetic marks for genomic regions that change compartment status between ES and GZ. Note that B to A shift in ES to GZ is associated with increased proportion of active transcribed regions (TssA and Tx) and enhancers (Enh, top), while A to B shift in ES to GZ is associated with increased proportions of repressive marks (Het and ReprPCWk, bottom). P-values from Fisher's exact test. Annotation for epigenetic marks described in a core 15-state model from
http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 3. Interacting regions share epigenetic states. a.** Epigenetic state combination in inter-chromosomal interacting regions in CP. Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1), transcriptional regulatory regions (TxReg), and transcribed regions (Tx) interact highly to each other as marked in red. **b-c.** Epigenetic state combination in intra-chromosomal interacting regions in GZ (**b**) and CP (**c**). Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1) and transcriptional regulatory regions (TxReg) interact highly to promoters (PromD1, PromD2) and transcribed regions (Tx5', Tx) as marked in red. Inter- and intra-chromosomal contact frequency map is compared to epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Colored bars on the left represent epigenetic marks associated with promoters and transcribed regions (orange), enhancers (red), and repressive marks (blue). Annotation for epigenetic marks described in a 25-state model from
http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 4. Characterization of chromatin interactome of human-gained enhancers. a.** Distribution fitting of normalized chromatin interaction frequency between human-gained enhancers with 1Mb upstream (top) and 100kb upstream (bottom) regions. Weibull distribution (red line) fits Hi-C interaction frequency the best for every distance range. **b.** Distribution of the number of significant interacting loci to human-gained enhancers in GZ (top), CP (middle), and ES (bottom). **c.** Fraction of histone states (left) and epigenetic mark enrichment (right) for regions interacting with

human-gained enhancers in GZ and CP. CDF, cumulative distribution function; Annotation for epigenetic marks described in
http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 5. Human-gained enhancers interact to evolutionary lineage-specific long non-coding RNAs (lncRNAs). a.** Protein-coding genes interacting with human-gained enhancers in CP (CP) and GZ (GZ) have lower non-synonymous substitutions (dN)/synonymous substitutions (dS) ratio compared to protein-coding genes non-interacting to human-gained enhancers (All) in mammals (mouse), primates (rhesus macaque), and great apes (chimpanzee), indicative of purifying selection. **b.** Number of lineage-specific lncRNAs interacting to human-gained enhancers (red vertical lines in the graph) in GZ (top) and CP (bottom). Null distribution generated from 3,000 permutations, where the number of lncRNAs interacting to the same number of enhancers pooled from all fetal brain enhancers was counted.

**Extended Data Figure 6. Association between eQTL and Hi-C interaction. a.** Overlap between eQTL transcripts and genes physically interacting to eQTL SNPs in CP and GZ. Significance of the overlap between eQTL transcripts and Hi-C interacting genes described in the upper right (Fisher's exact test). Background gene list for Fisher's exact test is all transcripts assessed in eQTL study within 1Mb from eQTL SNPs. **b-d.** Histone state enrichment for eQTL SNPs in adult frontal cortex (FCTX, **b**), fetal brain (FB, **c**), and IMR90 (**d**). **e.** Hi-C interaction frequency between eQTL SNPs and transcripts is greater than expected by chance in the relevant cell type. Lowess smooth curve plotted with actual data points. CP, chromatin contact frequency in CP; GZ, chromatin contact frequency in GZ; ES, chromatin contact frequency in ES; Exp, expected interaction frequency given the distance between two regions; Opp, opposite interaction frequency: interaction frequency of SNPs and transcripts when the position of genes was mirrored relative to the eQTL SNP. ***P<0.001, P-values from repeated measure of ANOVA.

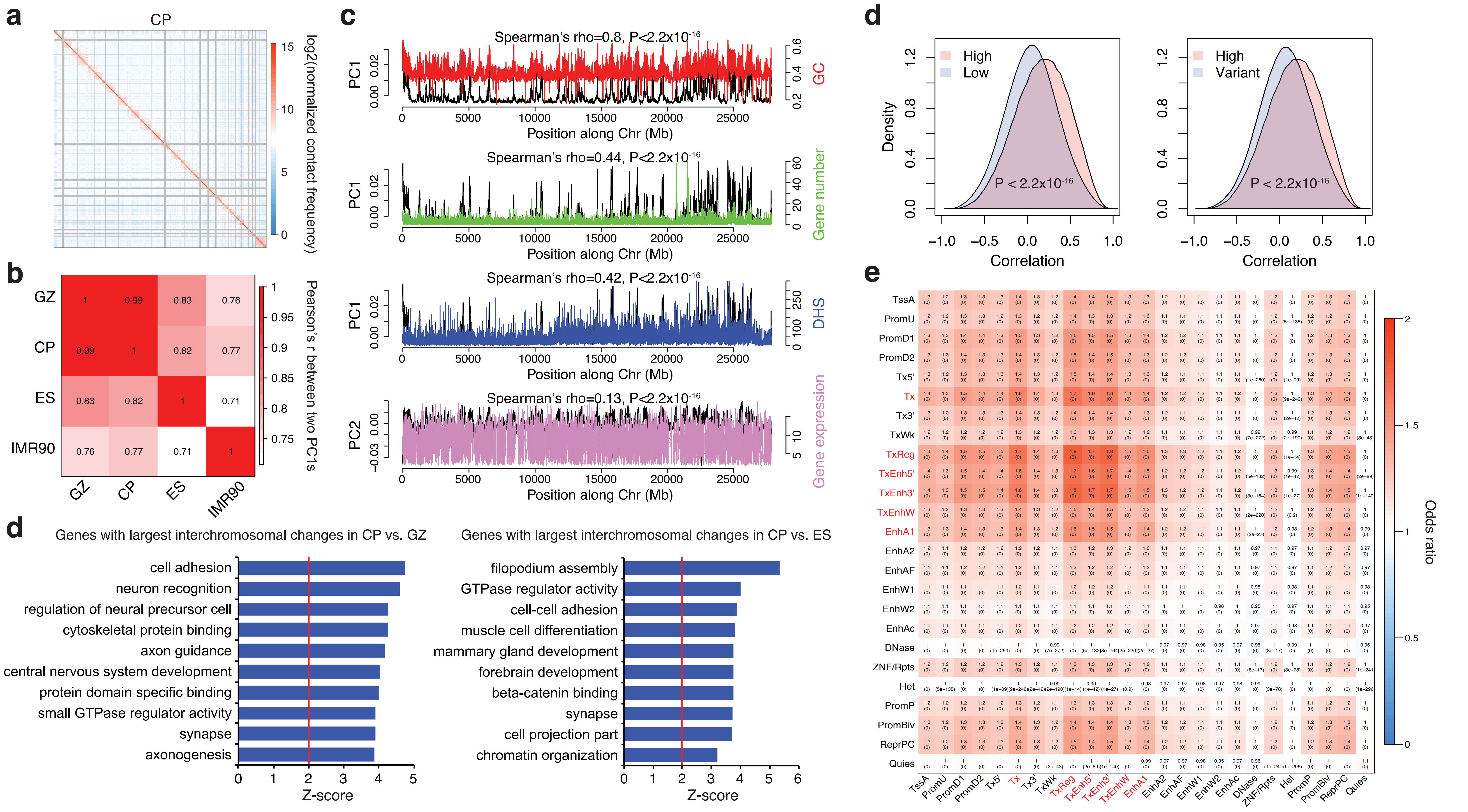
**Extended Data Figure 7. Defining schizophrenia risk genes based on functional annotation of credible SNPs.** Credible SNPs were selected using CAVIAR and categorized into functional SNPs, SNPs that fall onto gene promoters, and un-annotated SNPs. Histone state enrichment of credible SNPs was assessed in fetal brain (FB) and adult frontal cortex (FCTX). Functional SNPs and promoter SNPs were directly assigned to the target genes, while un-annotated SNPs were assigned to the target genes via Hi-C interactions in CP and GZ. GO enrichment for genes identified by each category is shown in the bottom. NMD, nonsense-mediated decay; TSS, transcription start site.

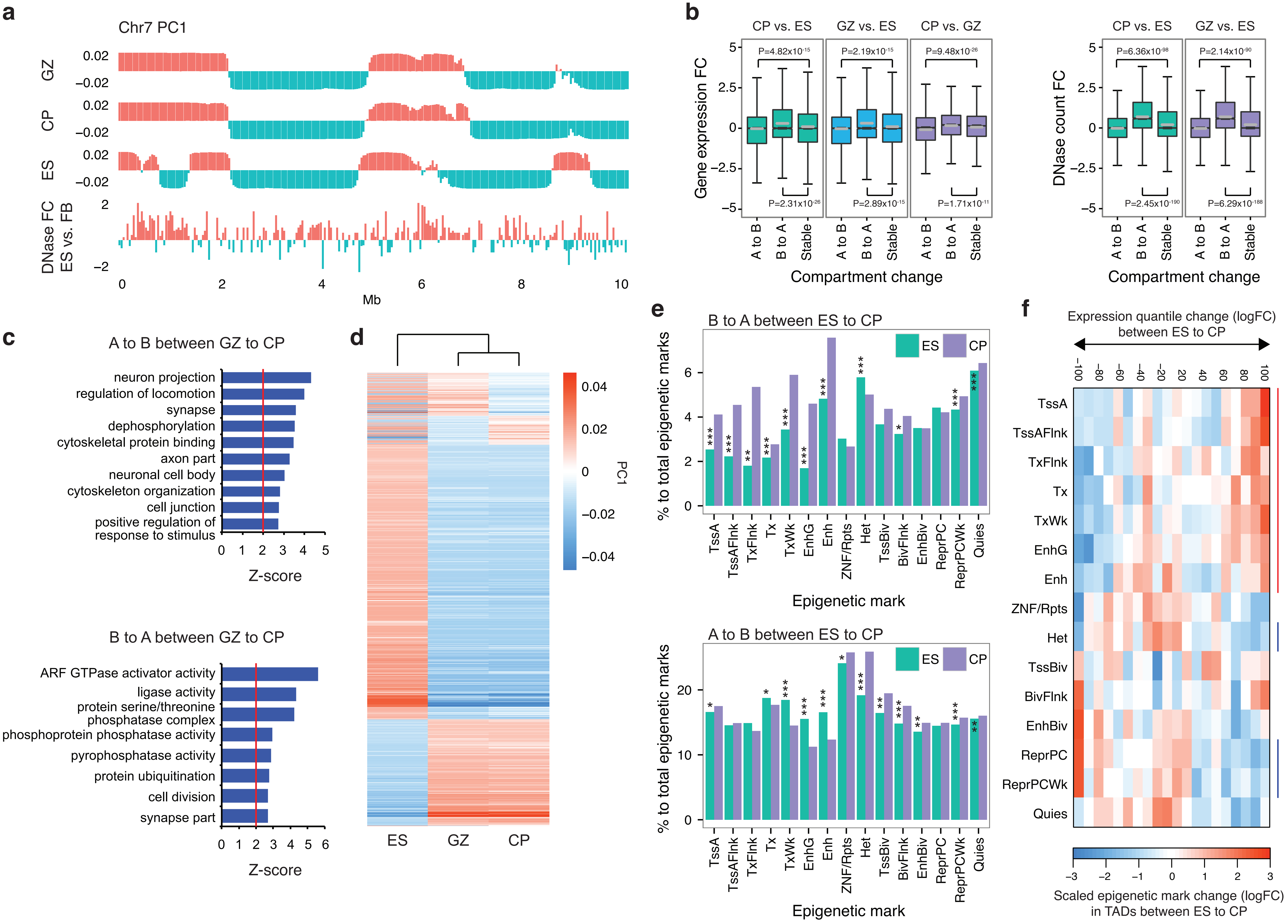
**Extended Data Figure 8. Representative interaction maps for credible SNPs to 1Mb flanking regions.** Interaction maps provide gene of actions for credible SNPs based on physical interaction. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; SNP, genomic coordinate for a 10kb bin in which credible SNPs locate; -log10(P-value), P-value for the significance of the interaction between credible SNPs and each 10kb bin, grey dashed line for FDR=0.01; GWAS loci, linkage disequilibrium (LD) region with the index SNP; TAD, TAD borders in CP, GZ, and ES.

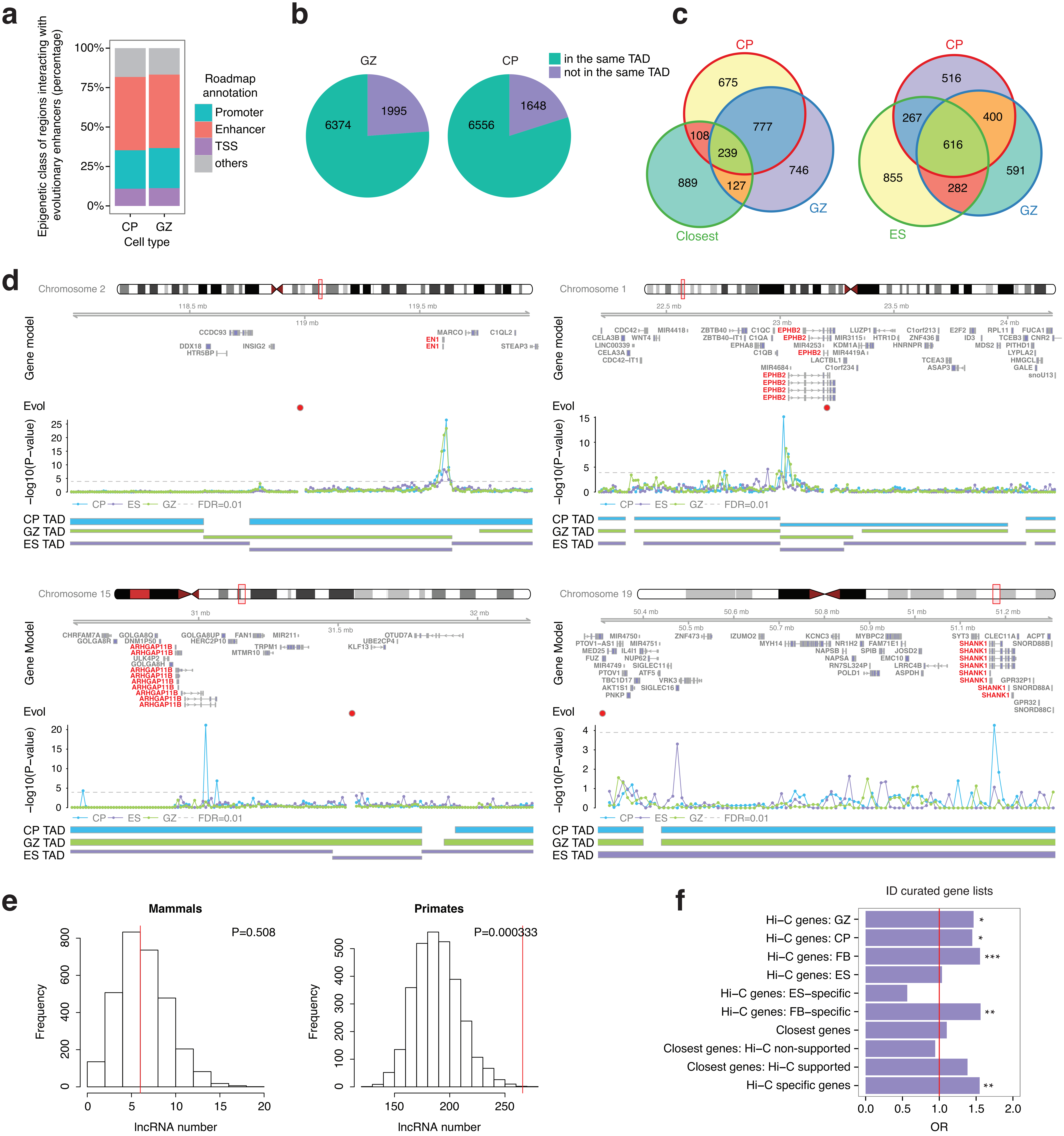
**Extended Data Figure 9. GO enrichment for schizophrenia risk genes curated by various methods. a-b.** GO enrichment for the closest genes to index SNPs (**a**) and genes in linkage disequilibrium (LD) with index SNPs (**b**) that are identified by a schizophrenia risk gene assessment pipeline in **Extended Data Fig. 7** (right) vs. not (left). **c.** GO enrichment for schizophrenia risk genes identified by a pipeline in **Extended Data Fig. 7** that are neither the closest genes nor in LD to index SNPs. Intersect and

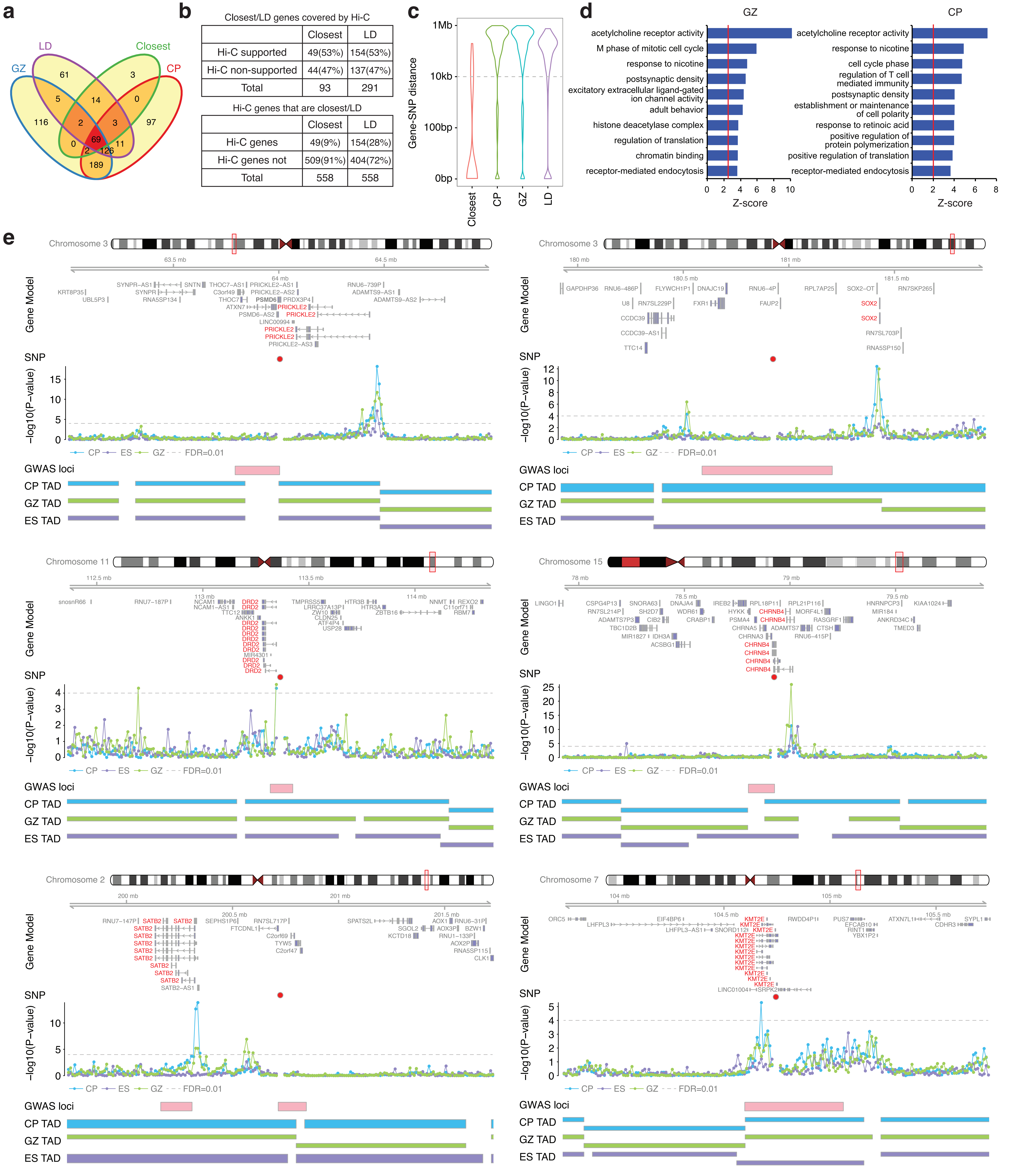
union between CP and GZ in left and right, respectively. Venn diagrams are marked in orange to depict the gene list assessed for GO enrichment.

**Extended Data Figure 10. Defining schizophrenia risk genes based on functional annotation of another set of credible SNPs.** Credible SNPs defined in the original study were categorized into functional SNPs, SNPs that fall onto gene promoters, and un-annotated SNPs. Overlap between credible SNPs identified by CAVIAR and credible SNPs originally identified indicates that two credible SNP lists overlap with each other. Histone state enrichment of credible SNPs in fetal brain (FB) and adult frontal cortex (FCTX). Functional SNPs and promoter SNPs were directly assigned to the target genes, while un-annotated SNPs were assigned to the target genes via Hi-C interactions in CP and GZ. GO enrichment for genes identified by each category and combined gene list is shown in the bottom. NMD, nonsense-mediated decay; TSS, transcription start site.


a
CP
log2(normalized contact frequency)
15
10
5
0
b
GZ
CP
ES
IMR90
1
0.99
0.83
0.76
0.99
1
0.82
0.77
0.83
0.82
1
0.71
0.76
0.77
0.71
1
Pearson's r between two PC1s
1
0.95
0.9
0.85
0.8
0.75
c
Spearman's rho=0.8, P<2.2x10^-16
PC1
GC
Position along Chr (Mb)
Spearman's rho=0.44, P<2.2x10^-16
Gene number
Spearman's rho=0.42, P<2.2x10^-16
DHS
Spearman's rho=0.13, P<2.2x10^-16
PC2
Gene expression
0
5000
10000
15000
20000
25000
d
High
Low
Variant
Density
Correlation
P < 2.2x10^-16
−1.0
−0.5
0.0
0.5
1.0
0.0
0.4
0.8
1.2
e
TssA
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
Odds ratio
2
1.5
1
0.5
0
d
Genes with largest interchromosomal changes in CP vs. GZ
cell adhesion
neuron recognition
regulation of neural precursor cell
cytoskeletal protein binding
axon guidance
central nervous system development
protein domain specific binding
small GTPase regulator activity
synapse
axonogenesis
Z-score
0
1
2
3
4
5
Genes with largest interchromosomal changes in CP vs. ES
filopodium assembly
GTPase regulator activity
cell-cell adhesion
muscle cell differentiation
mammary gland development
forebrain development
beta-catenin binding
synapse
cell projection part
chromatin organization
Z-score
6


a
Chr7 PC1
GZ
CP
ES
0.02
−0.02
DNase FC
ES vs. FB
2
−2
0
2
4
6
8
10
Mb
b
CP vs. ES
GZ vs. ES
CP vs. GZ
P=4.82x10-15
P=2.19x10-15
P=9.48x10-26
P=2.31x10-26
P=2.89x10-15
P=1.71x10-11
Gene expression FC
5
2.5
0
−2.5
−5
A to B
B to A
Stable
Compartment change
P=6.36x10-98
P=2.14x10-90
P=2.45x10-190
P=6.29x10-188
DNase count FC
c
A to B between GZ to CP
neuron projection
regulation of locomotion
synapse
dephosphorylation
cytoskeletal protein binding
axon part
neuronal cell body
cytoskeleton organization
cell junction
positive regulation of response to stimulus
Z-score
B to A between GZ to CP
ARF GTPase activator activity
ligase activity
protein serine/threonine phosphatase complex
phosphoprotein phosphatase activity
pyrophosphatase activity
protein ubiquitination
cell division
synapse part
Z-score
d
ES
GZ
CP
0.04
0.02
0
−0.02
−0.04
PC1
e
B to A between ES to CP
ES
CP
% to total epigenetic marks
Epigenetic mark
A to B between ES to CP
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
*
**
***
f
Expression quantile change (logFC) between ES to CP
−100
−80
−60
−40
−20
20
40
60
80
100
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
−3
−2
−1
0
1
2
3
Scaled epigenetic mark change (logFC) in TADs between ES to CP


a
Epigenetic class of regions interacting with evolutionary enhancers (percentage)
100%
75%
50%
25%
0%
CP
GZ
Cell type
Roadmap annotation
Promoter
Enhancer
TSS
others
b
GZ
CP
6374
1995
6556
1648
in the same TAD
not in the same TAD
c
CP
675
777
108
239
889
127
746
GZ
Closest
CP
516
267
400
616
855
282
591
GZ
ES
d
Chromosome 2
118.5 mb
119 mb
119.5 mb
Gene model
CCDC93
DDX18
HTR5BP
INSIG2
MARCO
EN1
EN1
C1QL2
STEAP3
Evol
−log10(P−value)
CP
ES
GZ
FDR=0.01
CP TAD
GZ TAD
ES TAD
Chromosome 1
22.5 mb
23 mb
23.5 mb
24 mb
EPHB2
Chromosome 15
31 mb
31.5 mb
32 mb
Gene Model
ARHGAP11B
Chromosome 19
50.4 mb
50.5 mb
50.6 mb
50.7 mb
50.8 mb
50.9 mb
51 mb
51.1 mb
51.2 mb
SHANK1
e
Mammals
P=0.508
Frequency
lncRNA number
Primates
P=0.000333
f
ID curated gene lists
Hi−C genes: GZ *
Hi−C genes: CP *
Hi−C genes: FB ***
Hi−C genes: ES
Hi−C genes: ES−specific
Hi−C genes: FB−specific **
Closest genes
Closest genes: Hi−C non-supported
Closest genes: Hi−C supported
Hi−C specific genes **
OR



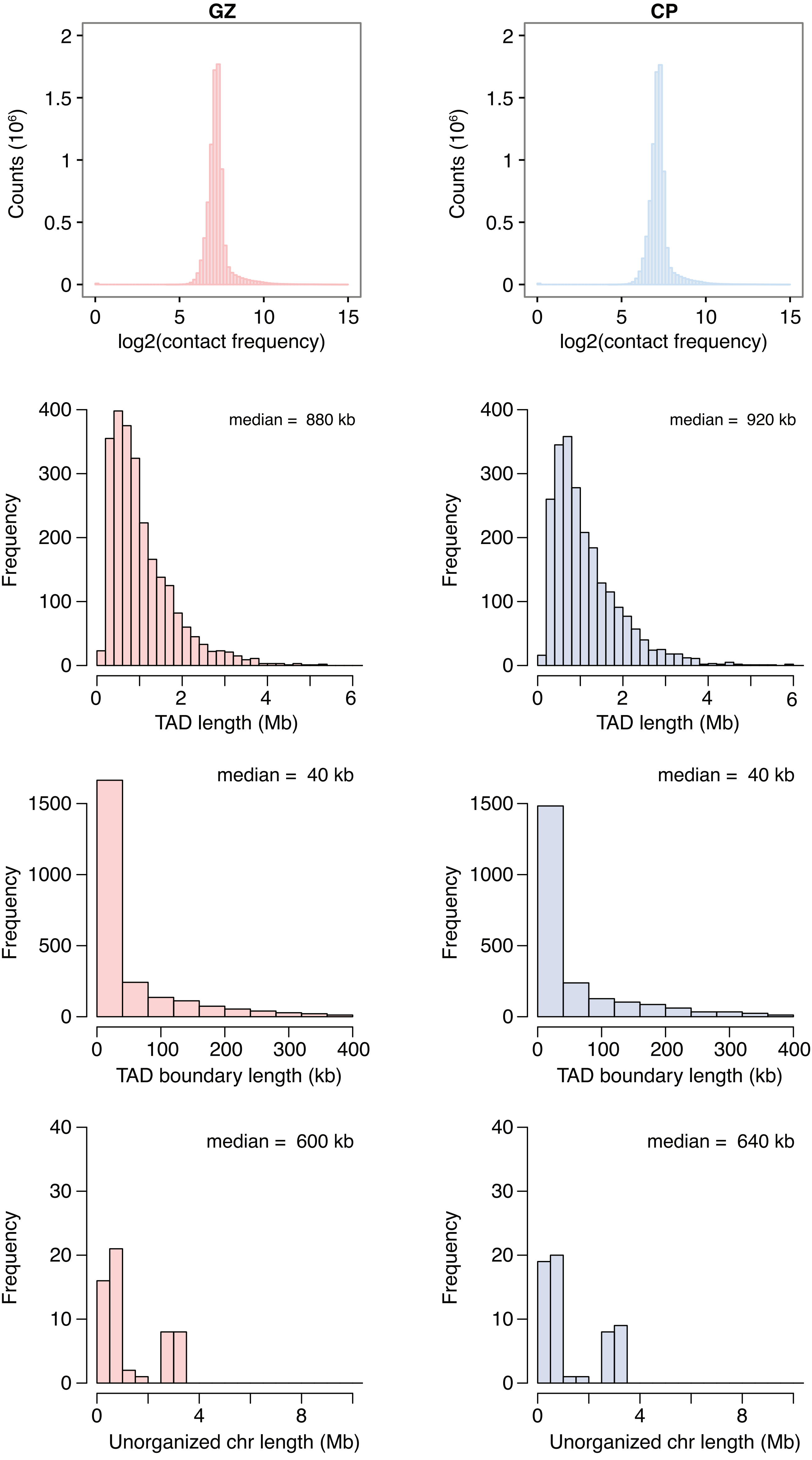
Closest/LD genes covered by Hi-C

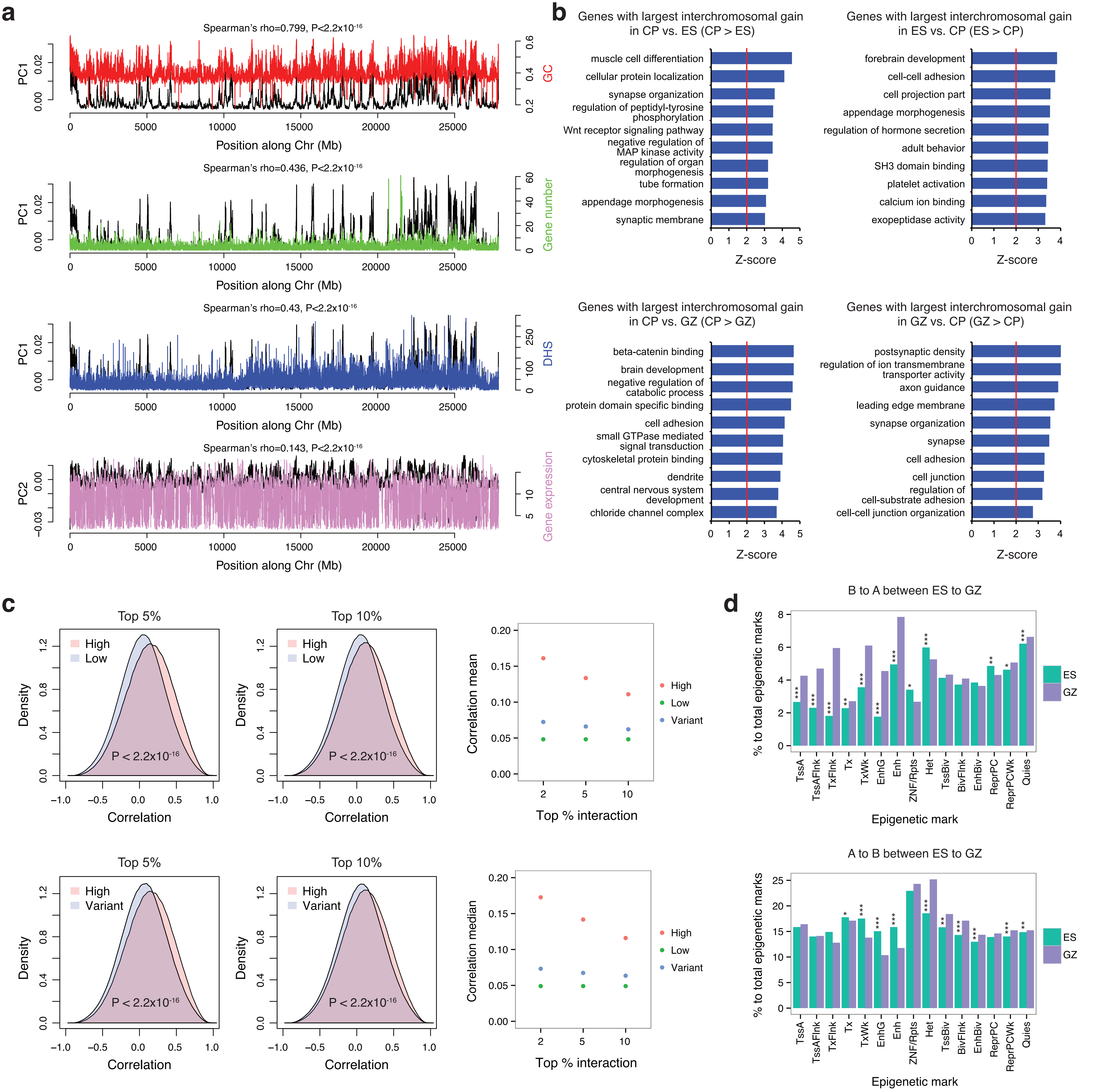
| | Closest | LD |
|---|---|---|
| Hi-C supported | 49(53%) | 154(53%) |
| Hi-C non-supported | 44(47%) | 137(47%) |
| Total | 93 | 291 |

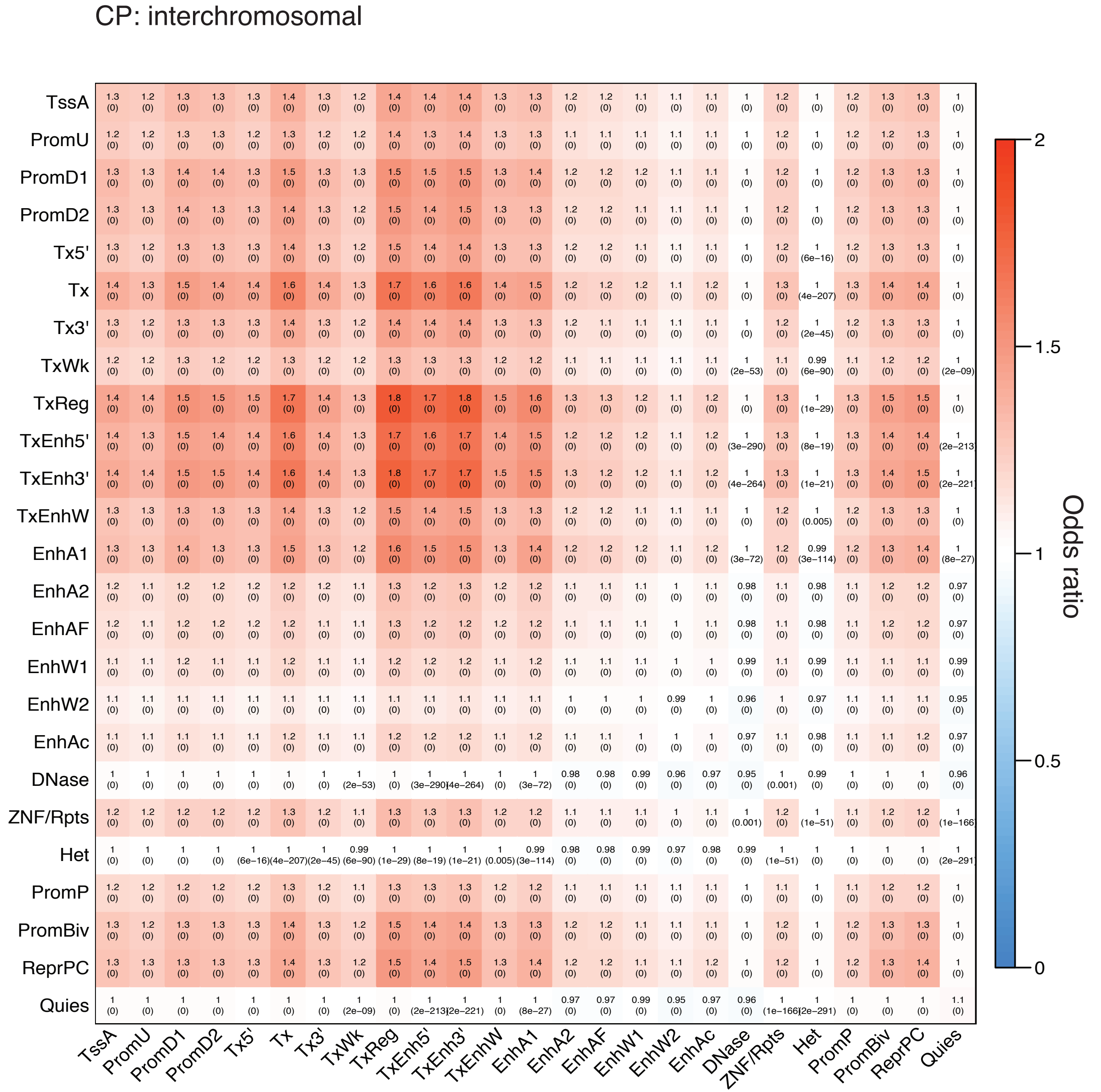
Hi-C genes that are closest/LD

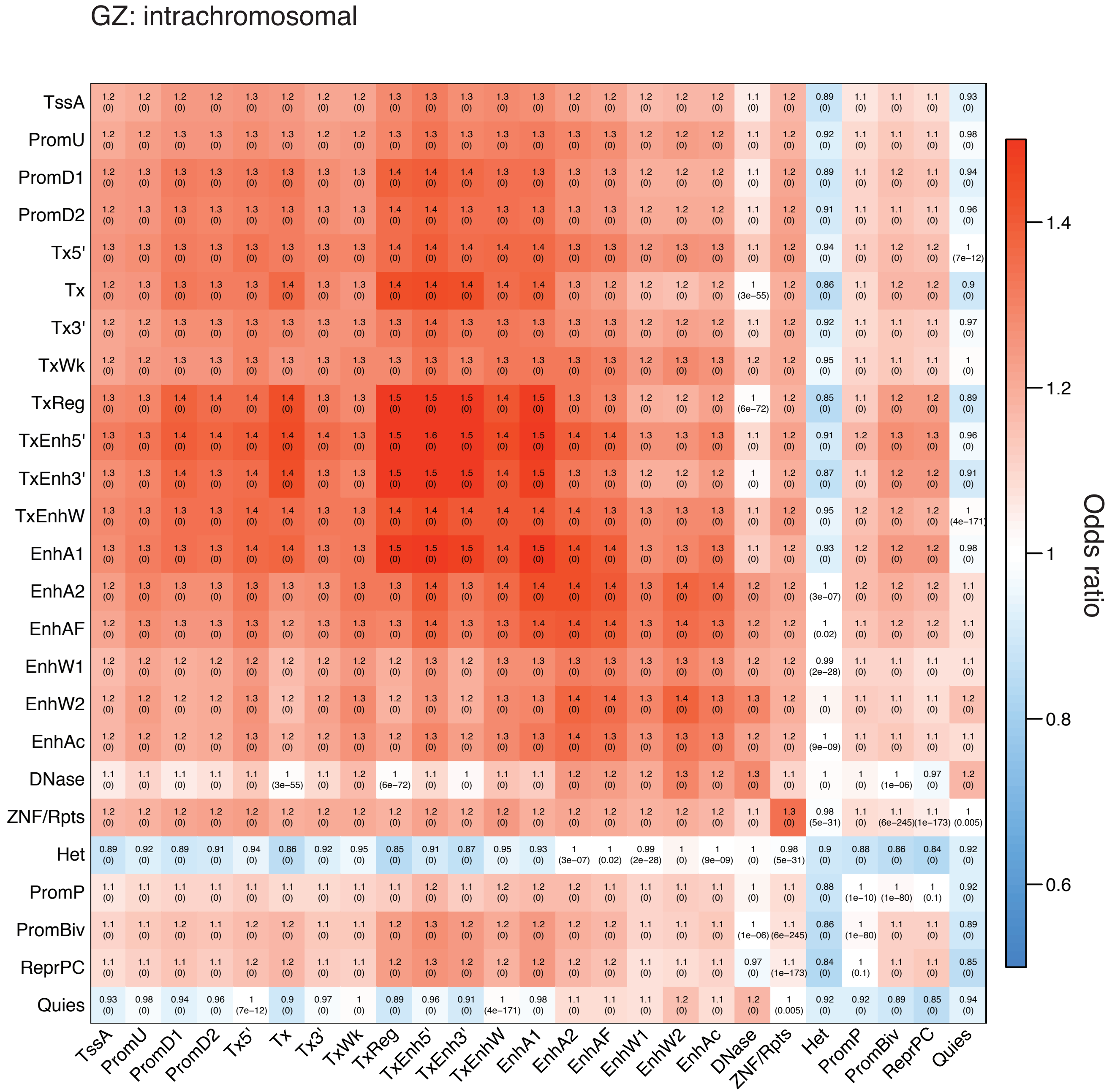
| | Closest | LD |
|---|---|---|
| Hi-C genes | 49(9%) | 154(28%) |
| Hi-C genes not | 509(91%) | 404(72%) |
| Total | 558 | 558 |

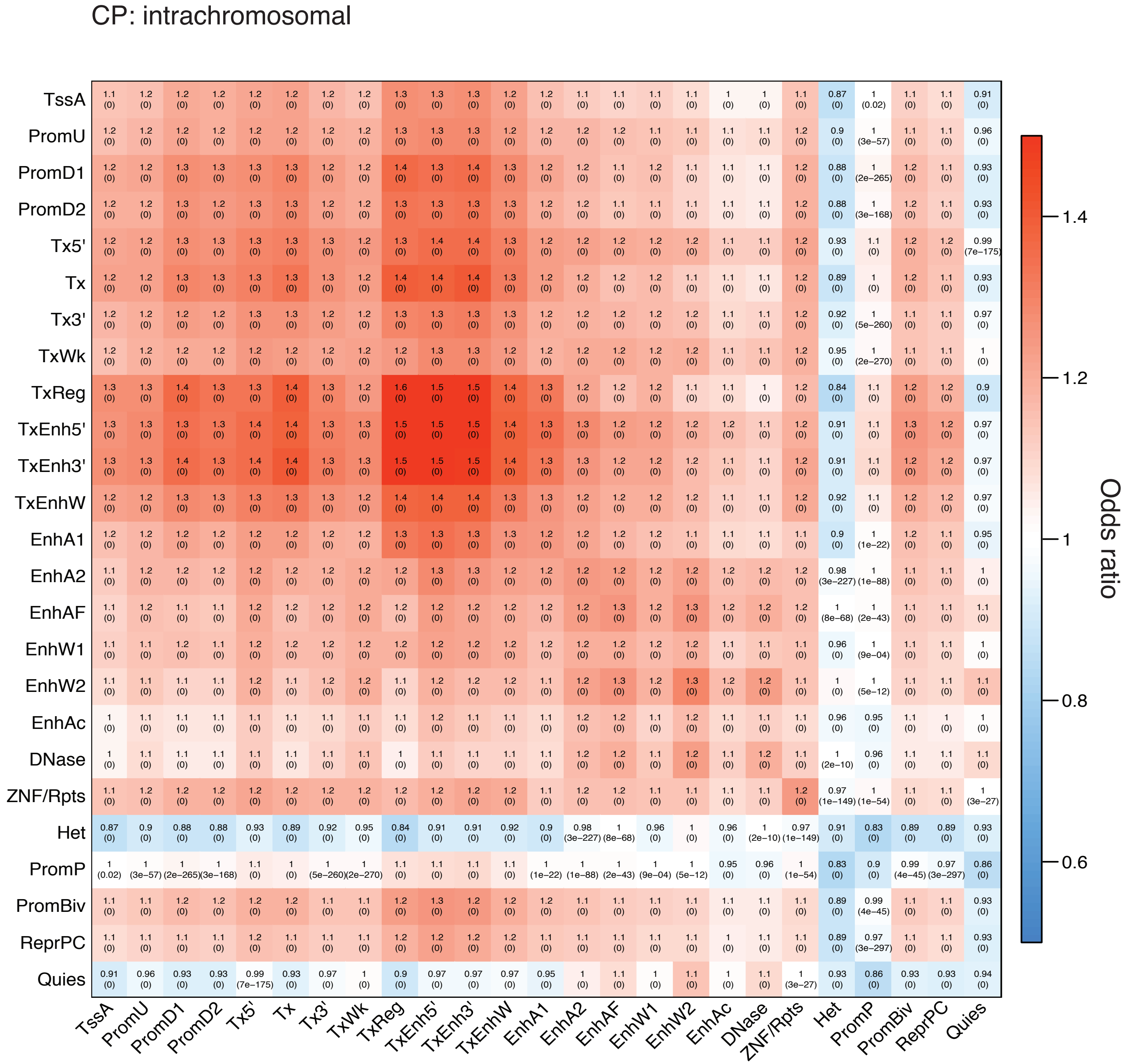
| Cell type | Cis ratio | All reads | DS mapped reads | Valid pairs | Filtered reads |
|---|---|---|---|---|---|
| GZ | 47.45% | 1,991,686,360 | 1,407,918,128<br>(70.69%) | 1,243,116,106<br>(88.29%) | 1,048,911,579<br>(74.50%) |
| CP | 46.40% | 1,958,637,304 | 1,352,951,087<br>(69.08%) | 1,225,315,488<br>(90.57%) | 1,022,593,960<br>(75.58%) |

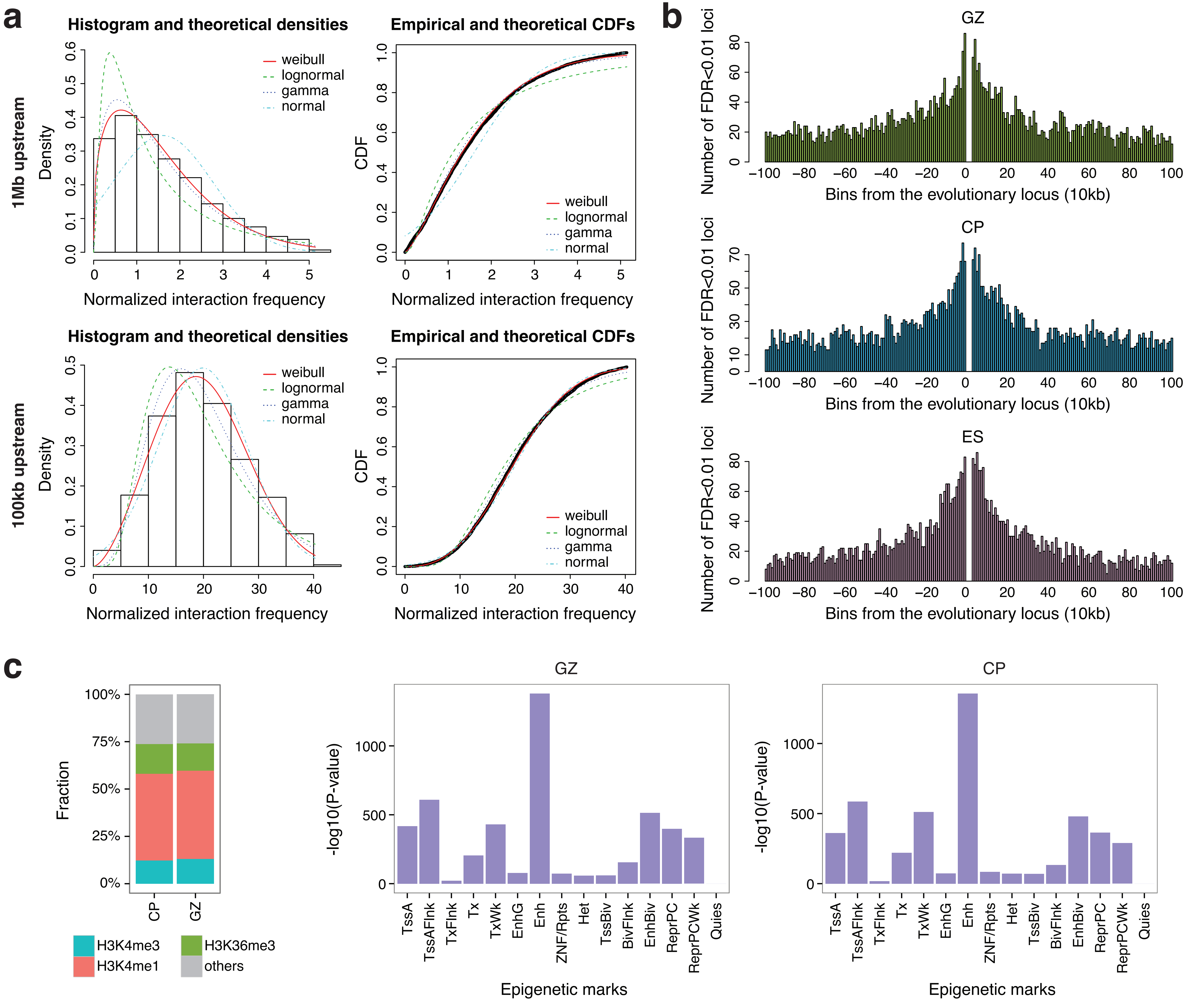


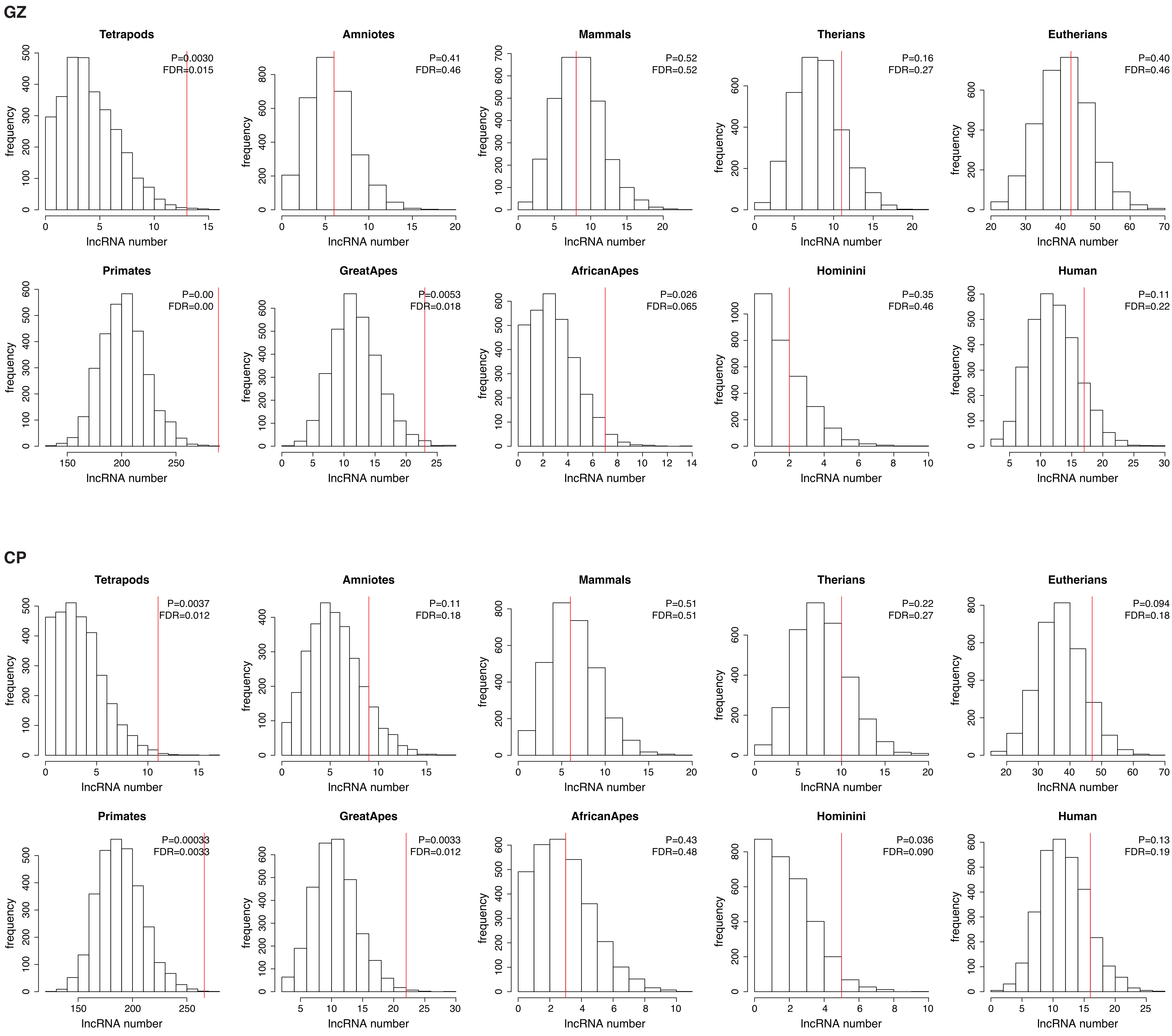

a
Spearman's rho=0.799, P<2.2x10-16
PC1
GC
Position along Chr (Mb)
Spearman's rho=0.436, P<2.2x10-16
Gene number
Spearman's rho=0.43, P<2.2x10-16
DHS
Spearman's rho=0.143, P<2.2x10-16
PC2
Gene expression
0
5000
10000
15000
20000
25000
b
Genes with largest interchromosomal gain in CP vs. ES (CP > ES)
muscle cell differentiation
cellular protein localization
synapse organization
regulation of peptidyl-tyrosine phosphorylation
Wnt receptor signaling pathway
negative regulation of MAP kinase activity
regulation of organ morphogenesis
tube formation
appendage morphogenesis
synaptic membrane
Z-score
Genes with largest interchromosomal gain in ES vs. CP (ES > CP)
forebrain development
cell-cell adhesion
cell projection part
appendage morphogenesis
regulation of hormone secretion
adult behavior
SH3 domain binding
platelet activation
calcium ion binding
exopeptidase activity
Genes with largest interchromosomal gain in CP vs. GZ (CP > GZ)
beta-catenin binding
brain development
negative regulation of catabolic process
protein domain specific binding
cell adhesion
small GTPase mediated signal transduction
cytoskeletal protein binding
dendrite
central nervous system development
chloride channel complex
Genes with largest interchromosomal gain in GZ vs. CP (GZ > CP)
postsynaptic density
regulation of ion transmembrane transporter activity
axon guidance
leading edge membrane
synapse organization
synapse
cell adhesion
cell junction
regulation of cell-substrate adhesion
cell-cell junction organization
c
Top 5%
Top 10%
High
Low
Variant
Density
Correlation
P < 2.2x10-16
Correlation mean
Correlation median
Top % interaction
d
B to A between ES to GZ
A to B between ES to GZ
% to total epigenetic marks
Epigenetic mark
ES
GZ
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies

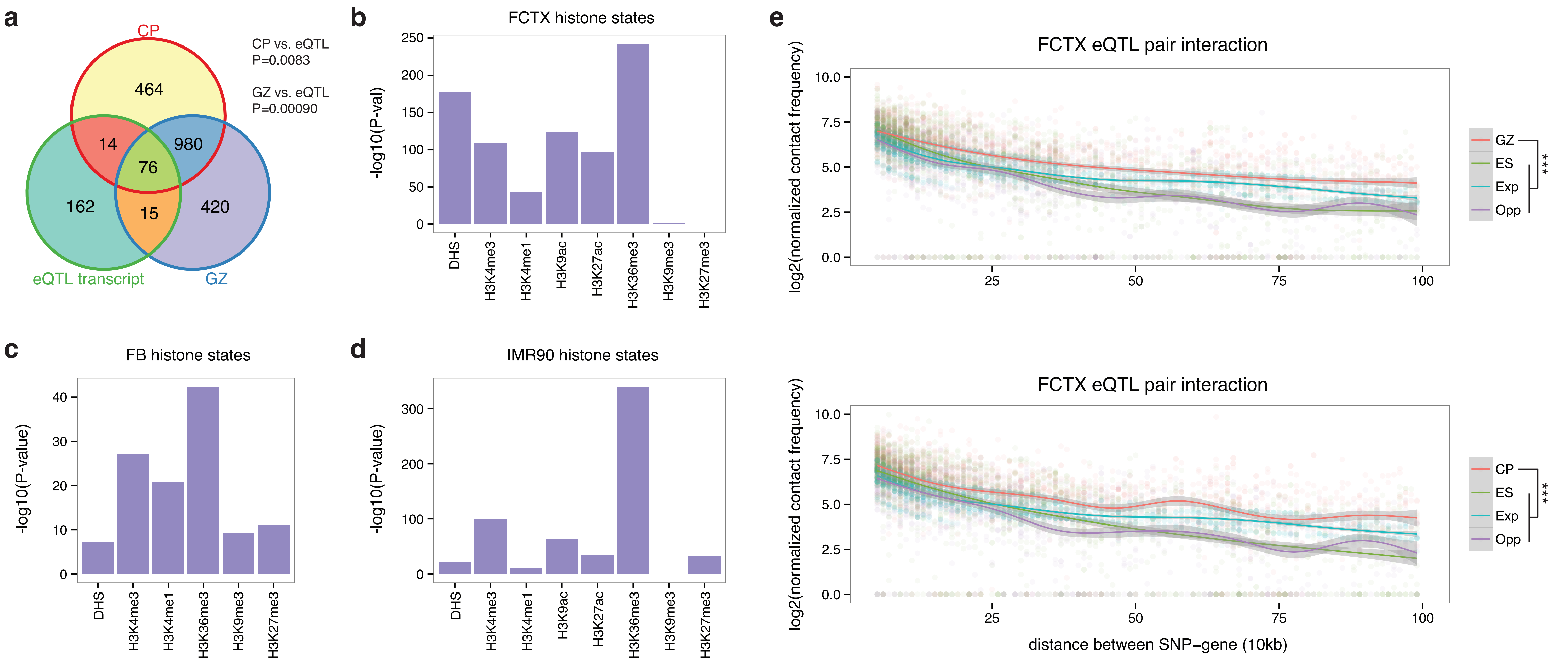

CP: interchromosomal
TssA
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
Odds ratio
2
1.5
1
0.5
0

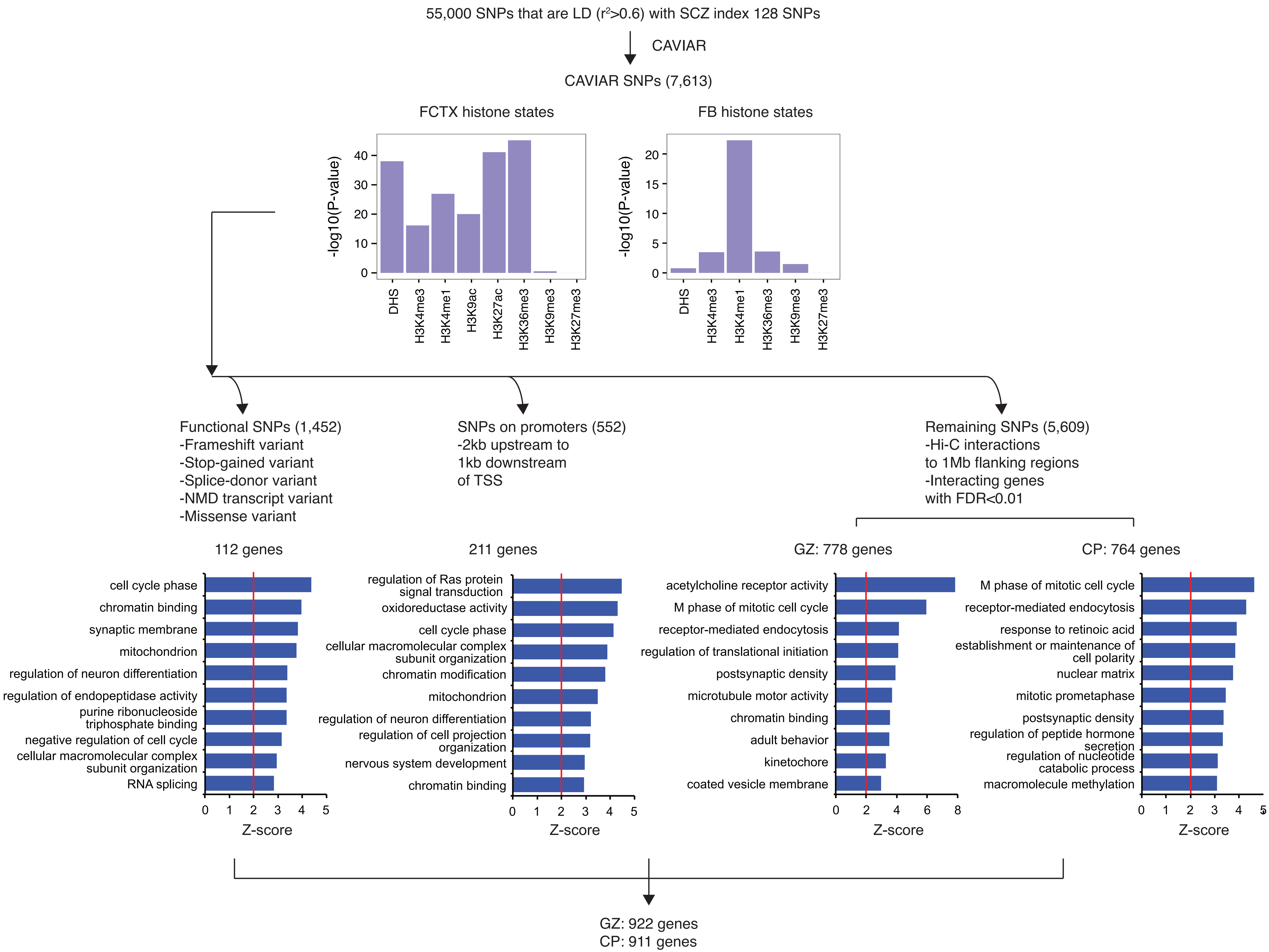

GZ: intrachromosomal
TssA
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
Odds ratio
1.4
1.2
1
0.8
0.6

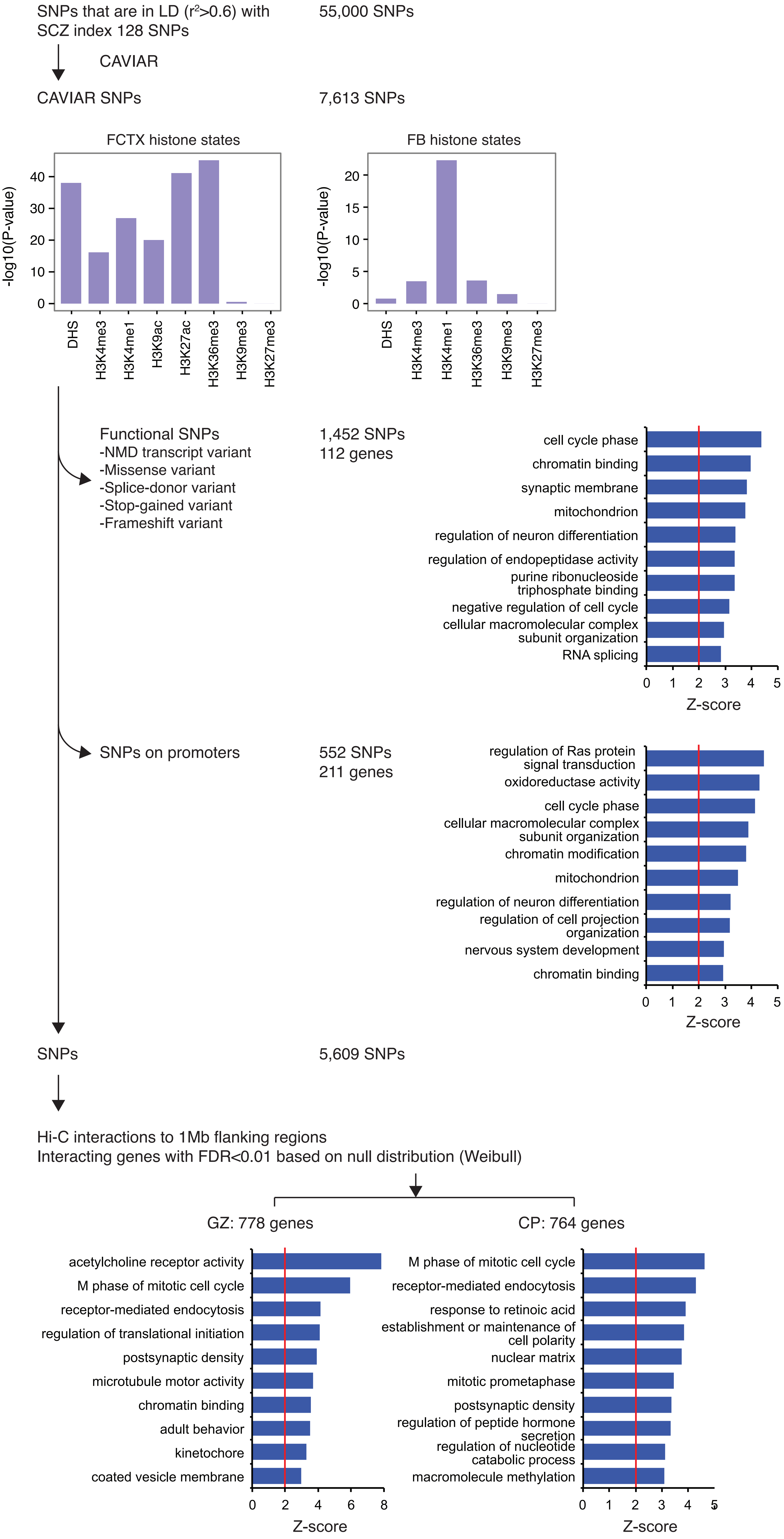

CP: intrachromosomal
TssA
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
Odds ratio
1.4
1.2
1
0.8
0.6

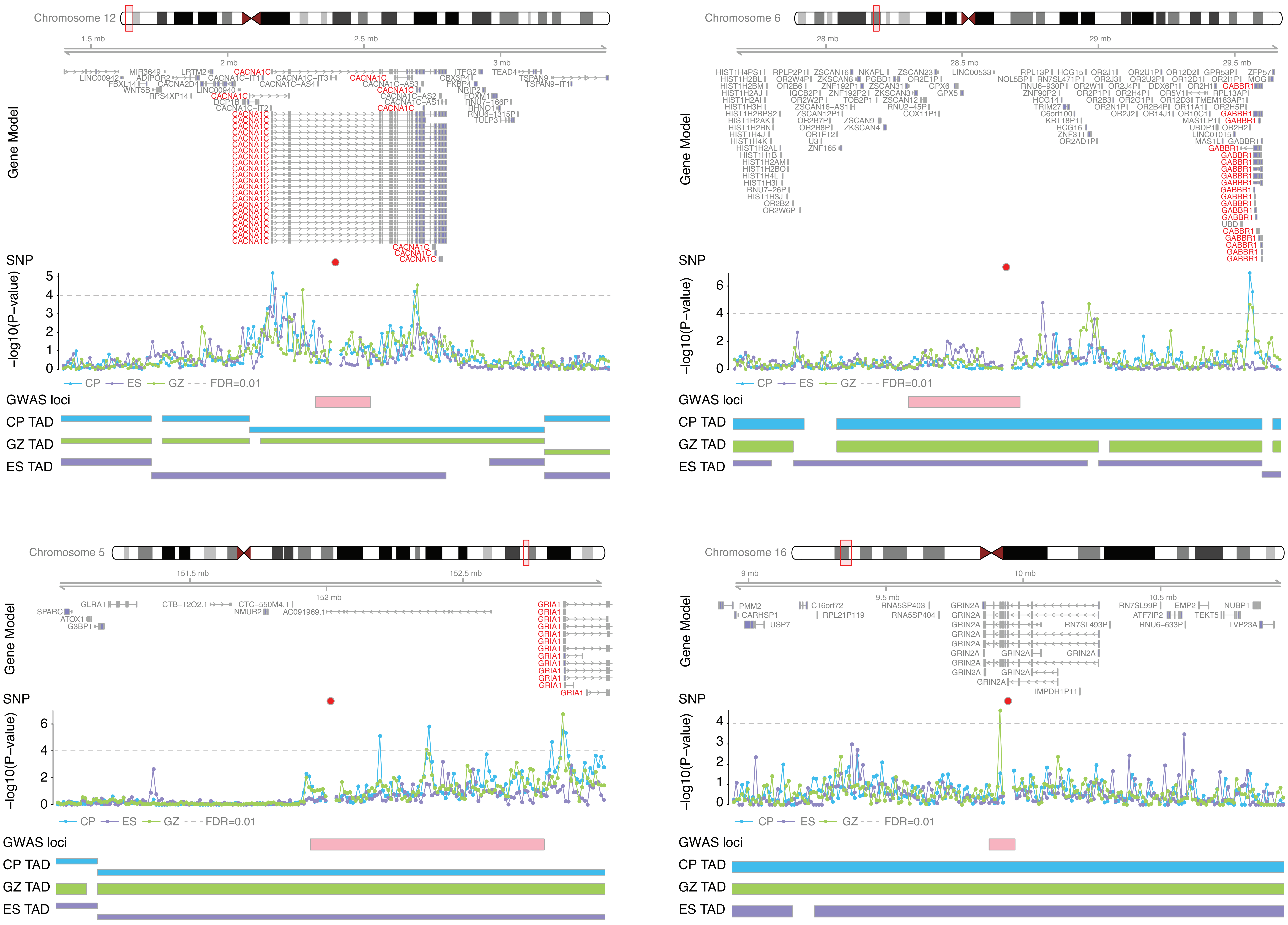

a
Histogram and theoretical densities
1Mb upstream
Density
weibull
lognormal
gamma
normal
0.0 0.1 0.2 0.3 0.4 0.5 0.6
0 1 2 3 4 5
Normalized interaction frequency
Empirical and theoretical CDFs
CDF
0.0 0.2 0.4 0.6 0.8 1.0
weibull
lognormal
gamma
normal
0 1 2 3 4 5
Normalized interaction frequency
Histogram and theoretical densities
100kb upstream
Density
weibull
lognormal
gamma
normal
0.0 0.1 0.2 0.3 0.4 0.5
0 10 20 30 40
Normalized interaction frequency
Empirical and theoretical CDFs
CDF
0.0 0.2 0.4 0.6 0.8 1.0
weibull
lognormal
gamma
normal
0 10 20 30 40
Normalized interaction frequency
b
GZ
Number of FDR<0.01 loci
0 20 40 60 80
−100 −80 −60 −40 −20 0 20 40 60 80 100
Bins from the evolutionary locus (10kb)
CP
Number of FDR<0.01 loci
0 10 30 50 70
−100 −80 −60 −40 −20 0 20 40 60 80 100
Bins from the evolutionary locus (10kb)
ES
Number of FDR<0.01 loci
0 20 40 60 80
−100 −80 −60 −40 −20 0 20 40 60 80 100
Bins from the evolutionary locus (10kb)
c
Fraction
0% 25% 50% 75% 100%
CP GZ
H3K4me3
H3K4me1
H3K36me3
others
GZ
-log10(P-value)
0 500 1000
TssA TssAFlnk TxFlnk Tx TxWk EnhG Enh ZNF/Rpts Het TssBiv BivFlnk EnhBiv ReprPC ReprPCWk Quies
Epigenetic marks
CP
-log10(P-value)
0 500 1000
TssA TssAFlnk TxFlnk Tx TxWk EnhG Enh ZNF/Rpts Het TssBiv BivFlnk EnhBiv ReprPC ReprPCWk Quies
Epigenetic marks

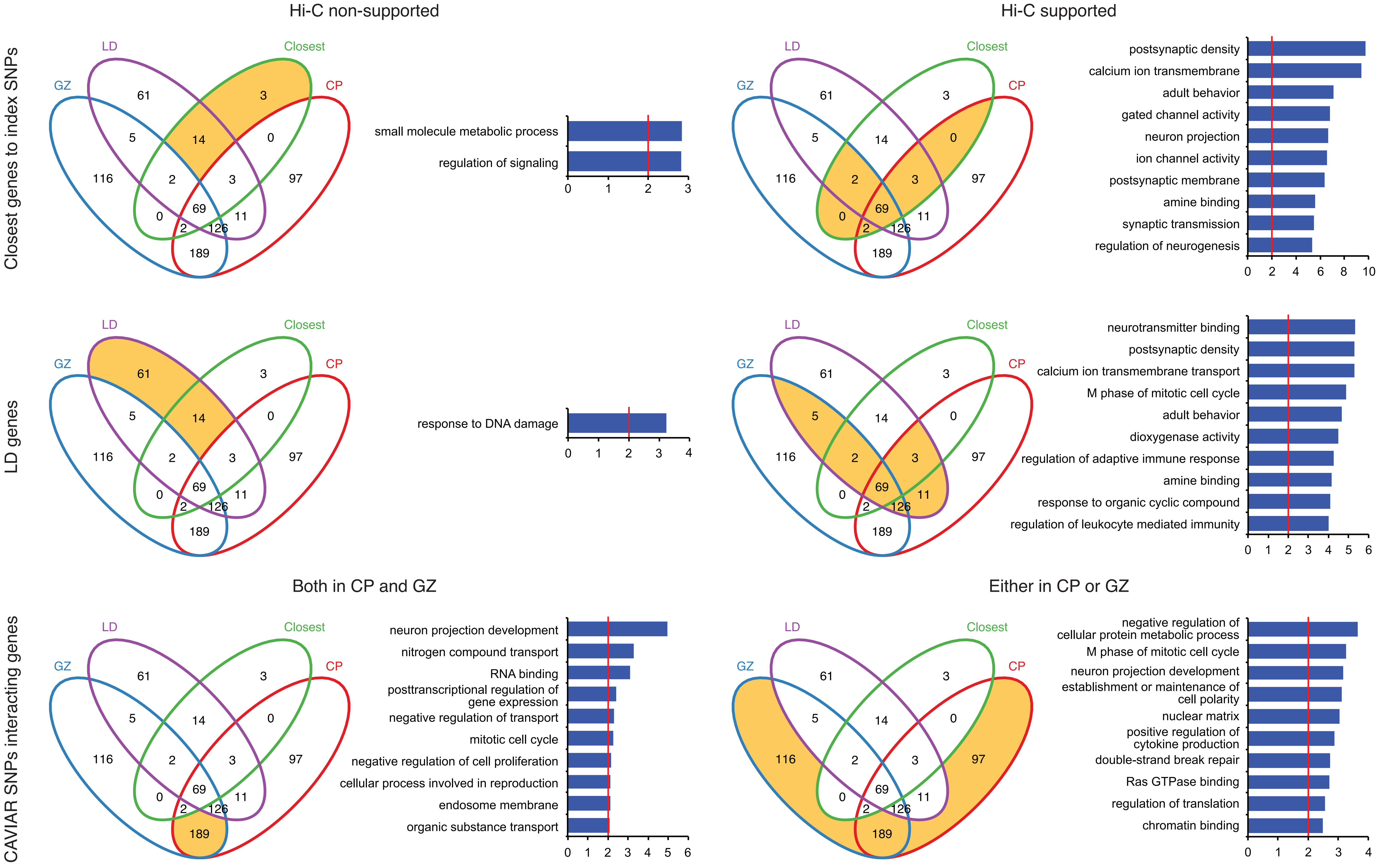

GZ
Tetrapods
P=0.0030
FDR=0.015
Amniotes
P=0.41
FDR=0.46
Mammals
P=0.52
FDR=0.52
Therians
P=0.16
FDR=0.27
Eutherians
P=0.40
FDR=0.46
Primates
P=0.00
FDR=0.00
GreatApes
P=0.0053
FDR=0.018
AfricanApes
P=0.026
FDR=0.065
Hominini
P=0.35
FDR=0.46
Human
P=0.11
FDR=0.22
CP
Tetrapods
P=0.0037
FDR=0.012
Amniotes
P=0.11
FDR=0.18
Mammals
P=0.51
FDR=0.51
Therians
P=0.22
FDR=0.27
Eutherians
P=0.094
FDR=0.18
Primates
P=0.00033
FDR=0.0033
GreatApes
P=0.0033
FDR=0.012
AfricanApes
P=0.43
FDR=0.48
Hominini
P=0.036
FDR=0.090
Human
P=0.13
FDR=0.19
frequency
lncRNA number

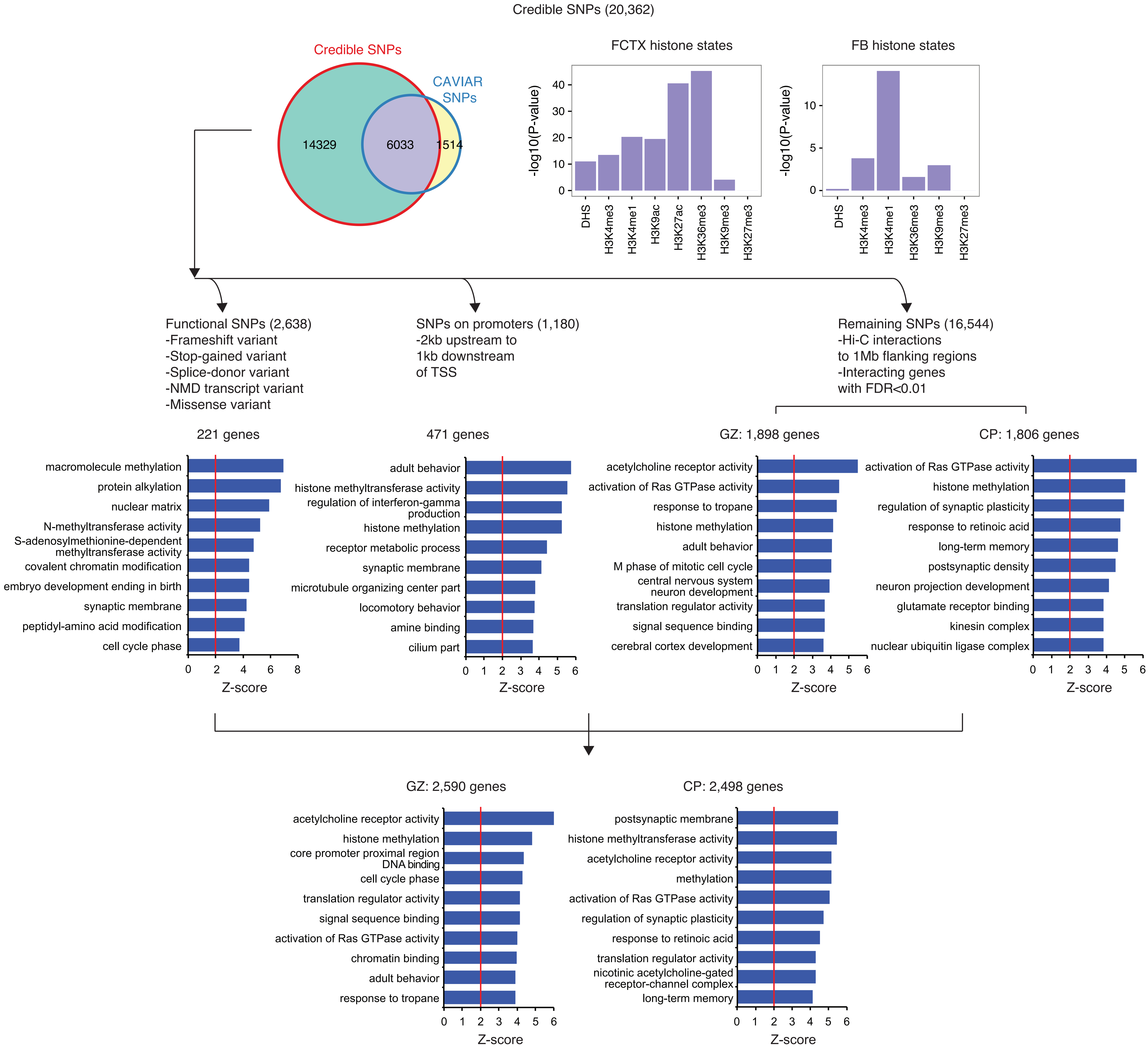

a
CP
464
14
980
76
162
15
420
eQTL transcript
GZ
CP vs. eQTL
P=0.0083
GZ vs. eQTL
P=0.00090
b
FCTX histone states
-log10(P-val)
250
200
150
100
50
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
c
FB histone states
-log10(P-value)
40
30
20
10
0
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
d
IMR90 histone states
-log10(P-value)
300
200
100
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
e
FCTX eQTL pair interaction
log2(normalized contact frequency)
10.0
7.5
5.0
2.5
0.0
25
50
75
100
GZ
ES
Exp
Opp
***
FCTX eQTL pair interaction
log2(normalized contact frequency)
10.0
7.5
5.0
2.5
0.0
25
50
75
100
distance between SNP–gene (10kb)
CP
ES
Exp
Opp
***


55,000 SNPs that are LD ($r^2$>0.6) with SCZ index 128 SNPs
CAVIAR
CAVIAR SNPs (7,613)
FCTX histone states
FB histone states
-log10(P-value)
40
30
20
10
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
-log10(P-value)
20
15
10
5
0
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs (1,452)
-Frameshift variant
-Stop-gained variant
-Splice-donor variant
-NMD transcript variant
-Missense variant
SNPs on promoters (552)
-2kb upstream to
1kb downstream
of TSS
Remaining SNPs (5,609)
-Hi-C interactions
to 1Mb flanking regions
-Interacting genes
with FDR<0.01
112 genes
cell cycle phase
chromatin binding
synaptic membrane
mitochondrion
regulation of neuron differentiation
regulation of endopeptidase activity
purine ribonucleoside triphosphate binding
negative regulation of cell cycle
cellular macromolecular complex subunit organization
RNA splicing
0 1 2 3 4 5
Z-score
211 genes
regulation of Ras protein signal transduction
oxidoreductase activity
cell cycle phase
cellular macromolecular complex subunit organization
chromatin modification
mitochondrion
regulation of neuron differentiation
regulation of cell projection organization
nervous system development
chromatin binding
0 1 2 3 4 5
Z-score
GZ: 778 genes
acetylcholine receptor activity
M phase of mitotic cell cycle
receptor-mediated endocytosis
regulation of translational initiation
postsynaptic density
microtubule motor activity
chromatin binding
adult behavior
kinetochore
coated vesicle membrane
0 2 4 6 8
Z-score
CP: 764 genes
M phase of mitotic cell cycle
receptor-mediated endocytosis
response to retinoic acid
establishment or maintenance of cell polarity
nuclear matrix
mitotic prometaphase
postsynaptic density
regulation of peptide hormone secretion
regulation of nucleotide catabolic process
macromolecule methylation
0 1 2 3 4 5
Z-score
GZ: 922 genes
CP: 911 genes


SNPs that are in LD ($r^2$>0.6) with SCZ index 128 SNPs
55,000 SNPs
CAVIAR
CAVIAR SNPs
7,613 SNPs
FCTX histone states
-log10(P-value)
40
30
20
10
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
FB histone states
-log10(P-value)
20
15
10
5
0
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs
-NMD transcript variant
-Missense variant
-Splice-donor variant
-Stop-gained variant
-Frameshift variant
1,452 SNPs
112 genes
cell cycle phase
chromatin binding
synaptic membrane
mitochondrion
regulation of neuron differentiation
regulation of endopeptidase activity
purine ribonucleoside triphosphate binding
negative regulation of cell cycle
cellular macromolecular complex subunit organization
RNA splicing
0
1
2
3
4
5
Z-score
SNPs on promoters
552 SNPs
211 genes
regulation of Ras protein signal transduction
oxidoreductase activity
cell cycle phase
cellular macromolecular complex subunit organization
chromatin modification
mitochondrion
regulation of neuron differentiation
regulation of cell projection organization
nervous system development
chromatin binding
0
1
2
3
4
5
Z-score
SNPs
5,609 SNPs
Hi-C interactions to 1Mb flanking regions
Interacting genes with FDR<0.01 based on null distribution (Weibull)
GZ: 778 genes
acetylcholine receptor activity
M phase of mitotic cell cycle
receptor-mediated endocytosis
regulation of translational initiation
postsynaptic density
microtubule motor activity
chromatin binding
adult behavior
kinetochore
coated vesicle membrane
0
2
4
6
8
Z-score
CP: 764 genes
M phase of mitotic cell cycle
receptor-mediated endocytosis
response to retinoic acid
establishment or maintenance of cell polarity
nuclear matrix
mitotic prometaphase
postsynaptic density
regulation of peptide hormone secretion
regulation of nucleotide catabolic process
macromolecule methylation
0
1
2
3
4
5
Z-score


Chromosome 12
1.5 mb
2 mb
2.5 mb
3 mb
Gene Model
CACNA1C
SNP
−log10(P−value)
CP
ES
GZ
FDR=0.01
GWAS loci
CP TAD
GZ TAD
ES TAD
Chromosome 6
28 mb
28.5 mb
29 mb
29.5 mb
GABBR1
Chromosome 5
151.5 mb
152 mb
152.5 mb
GRIA1
Chromosome 16
9 mb
9.5 mb
10 mb
10.5 mb
GRIN2A


Hi-C non-supported
Hi-C supported
Closest genes to index SNPs
LD genes
CAVIAR SNPs interacting genes
Both in CP and GZ
Either in CP or GZ
LD
Closest
GZ
CP
61
3
5
14
0
116
2
3
97
69
0
11
2
126
189
small molecule metabolic process
regulation of signaling
0
1
2
3
postsynaptic density
calcium ion transmembrane
adult behavior
gated channel activity
neuron projection
ion channel activity
postsynaptic membrane
amine binding
synaptic transmission
regulation of neurogenesis
0
2
4
6
8
10
response to DNA damage
0
1
2
3
4
neurotransmitter binding
postsynaptic density
calcium ion transmembrane transport
M phase of mitotic cell cycle
adult behavior
dioxygenase activity
regulation of adaptive immune response
amine binding
response to organic cyclic compound
regulation of leukocyte mediated immunity
0
1
2
3
4
5
6
neuron projection development
nitrogen compound transport
RNA binding
posttranscriptional regulation of gene expression
negative regulation of transport
mitotic cell cycle
negative regulation of cell proliferation
cellular process involved in reproduction
endosome membrane
organic substance transport
0
1
2
3
4
5
6
negative regulation of cellular protein metabolic process
M phase of mitotic cell cycle
neuron projection development
establishment or maintenance of cell polarity
nuclear matrix
positive regulation of cytokine production
double-strand break repair
Ras GTPase binding
regulation of translation
chromatin binding
0
1
2
3
4


Credible SNPs (20,362)
Credible SNPs
CAVIAR SNPs
14329
6033
1514
FCTX histone states
-log10(P-value)
0
10
20
30
40
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
FB histone states
-log10(P-value)
0
5
10
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs (2,638)
-Frameshift variant
-Stop-gained variant
-Splice-donor variant
-NMD transcript variant
-Missense variant
SNPs on promoters (1,180)
-2kb upstream to
1kb downstream
of TSS
Remaining SNPs (16,544)
-Hi-C interactions
to 1Mb flanking regions
-Interacting genes
with FDR<0.01
221 genes
macromolecule methylation
protein alkylation
nuclear matrix
N-methyltransferase activity
S-adenosylmethionine-dependent methyltransferase activity
covalent chromatin modification
embryo development ending in birth
synaptic membrane
peptidyl-amino acid modification
cell cycle phase
0 2 4 6 8
Z-score
471 genes
adult behavior
histone methyltransferase activity
regulation of interferon-gamma production
histone methylation
receptor metabolic process
synaptic membrane
microtubule organizing center part
locomotory behavior
amine binding
cilium part
0 1 2 3 4 5 6
Z-score
GZ: 1,898 genes
acetylcholine receptor activity
activation of Ras GTPase activity
response to tropane
histone methylation
adult behavior
M phase of mitotic cell cycle
central nervous system neuron development
translation regulator activity
signal sequence binding
cerebral cortex development
0 1 2 3 4 5 6
Z-score
CP: 1,806 genes
activation of Ras GTPase activity
histone methylation
regulation of synaptic plasticity
response to retinoic acid
long-term memory
postsynaptic density
neuron projection development
glutamate receptor binding
kinesin complex
nuclear ubiquitin ligase complex
0 1 2 3 4 5 6
Z-score
GZ: 2,590 genes
acetylcholine receptor activity
histone methylation
core promoter proximal region DNA binding
cell cycle phase
translation regulator activity
signal sequence binding
activation of Ras GTPase activity
chromatin binding
adult behavior
response to tropane
0 1 2 3 4 5 6
Z-score
CP: 2,498 genes
postsynaptic membrane
histone methyltransferase activity
acetylcholine receptor activity
methylation
activation of Ras GTPase activity
regulation of synaptic plasticity
response to retinoic acid
translation regulator activity
nicotinic acetylcholine-gated receptor-channel complex
long-term memory
0 1 2 3 4 5 6
Z-score

## APPENDIX

**Content:**

- Preprint of manuscript by Parikshak, N. N. et al. Global changes in patterning, splicing and primate specific lncRNAs in autism brain.
- Preprint of manuscript by Won, H. et al. Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development.

# Global changes in patterning, splicing and primate specific lncRNAs in autism brain

Neelroop N. Parikshak[1,2,]*, Vivek Swarup[1,]*, T. Grant Belgard[1,2, †,]*, Michael Gandal[1,2], Manuel Irimia[5,6], Virpi Leppa[1], Jennifer K. Lowe[1], Robert Johnson[7], Benjamin J. Blencowe[6], Steve Horvath[3-4], Daniel H. Geschwind[1-3]

1. Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
2. Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.
3. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
4. Department of Biostatistics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
5. EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, Barcelona 08003, Spain.
6. Donnelly Centre, University of Toronto, 160 College Street, Toronto, ON M5S 3E1, Canada; Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada.
7. NICHD Brain and Tissue Bank for Developmental Disorders, University of Maryland Medical School, Baltimore, Maryland 21201, USA.

*These authors contributed equally to this study.

†Current address: MRC Functional Genomics Unit, Department of Physiology, Anatomy & Genetics, University of Oxford, South Parks Road, Oxford, OX1 3PT, United Kingdom.

## Summary

We apply transcriptome-wide RNA sequencing in postmortem autism spectrum disorder (ASD) brain and controls and identify convergent alterations in the noncoding transcriptome, including primate specific lncRNA, and transcript splicing in ASD cerebral cortex, but not cerebellum. We characterize an attenuation of patterning between frontal and temporal cortex in ASD and identify *SOX5*, a transcription factor involved in cortical neuron fate specification, as a likely driver of this pattern. We further show that a genetically defined subtype of ASD, Duplication 15q Syndrome, shares the core transcriptomic signature of idiopathic ASD, indicating that observed molecular convergence in autism brain is the likely consequence of manifold genetic alterations. Using co-expression network analysis, we show that diverse forms of genetic risk for ASD affect convergent, independently replicated, biological pathways and provide an unprecedented resource for understanding the molecular alterations associated with ASD in humans.

Autism spectrum disorder (ASD) is a neurodevelopmental syndrome characterized by deficits in social communication and mental flexibility[1]. Genetic risk factors contribute substantially to ASD risk, and recent studies support the potential contribution of more than a thousand genes to ASD risk[2-4]. However, given the shared cognitive and behavioral features across the autism spectrum, one hypothesis is that diverse risk factors may converge on common molecular, cellular, and circuit level pathways to result in the shared phenotype[5,6]. Analysis of the transcriptome has been used to identify common molecular pathways in the cerebral cortex (CTX) from postmortem human brain tissue in individuals with ASD[7-11]. However, all transcriptomic studies in ASD to date have been limited to evaluating highly expressed mRNAs corresponding to protein coding genes. Moreover, most lack rigorous replication and do not assess gene expression patterns across brain regions.

We used rRNA-depleted RNA-seq (Methods) to evaluate transcriptomes from a large set of ASD and control (CTL) brain samples including neocortex (frontal and temporal) and cerebellum across 79 individuals (46 ASD, 33 CTL, 205 samples, Extended Data Fig. 1a-e, Supplementary Table 1). We first compared differential gene expression (DGE) between ASD and CTL individuals in CTX from a previously published[7] microarray study against new, independent gene expression profiles from RNA-seq to evaluate global reproducibility of DGE in ASD. We found a high degree of replication of DGE fold changes between the sample sets, despite evaluation on different gene expression platforms (fold changes at $P < 0.05$ in previously evaluated data correlate with new data with $R^2 = 0.60$, Extended Data Fig. 1f). We observed a much weaker overall signal and replication in cerebellum ($R^2 = 0.033$, Extended Data Fig. 1g). These analyses confirm the existence of a reproducible DGE signature in ASD CTX across different platforms and in independent samples.

We next combined samples from all individuals with idiopathic ASD into a covariate-matched "ASD Discovery Set" (Extended Data Fig. 1h) for CTX (106 samples, 26 ASD, 33 CTL individuals) and held out remaining samples for replication ("ASD Replication Set", Methods). For DGE analysis, we used a linear mixed effects model that accounts for biological and technical covariates (Methods) to identify 1156 genes differentially expressed in ASD CTX, 582 increased and 574 decreased (Benjamini-Hochberg FDR ≤ 0.05, Supplementary Table 2). Importantly, DGE analysis with additional covariates or different assumptions about the distribution of the data and test statistics yielded similar results (Extended Data Fig. 2a). Additionally this DGE signature clusters over two-thirds of ASD samples together and this clustering is not related to confounding factors such as cortical region, age, sex, and RNA quality (Figure 1a, Extended Data Fig. 2b). The most significantly down-regulated gene was *PVALB* (fold change = 0.53, FDR ≤ 0.05), a marker for GABAergic interneurons. *SST*, a marker for a different subpopulation of GABAergic interneurons, is also among the most downregulated (fold change = 0.61, FDR ≤ 0.05). Other down-regulated genes at FDR ≤ 0.05 include *NEUROD6*, involved in neuronal differentiation (fold change = 0.60), multiple ion channels, and *KDM5D*, a lysine demethylase (fold change = 0.66). In contrast, members of the complement cascade implicated in microglial-neuronal interactions (*C4A*, fold change = 1.94; *C1QB*, fold change = 1.65; both FDR ≤ 0.05) are upregulated in ASD CTX. Gene Ontology (GO) term enrichment analysis further supports the involvement of pathways implicated by these genes (Figure 1b), confirming previous findings[7]. Moreover, the upregulated set is enriched for astrocyte and microglia enriched genes, and the down-regulated set is enriched for synaptic genes (Extended Data Fig. 2c), consistent with previous observations[7,11].

We next sought to evaluate whether the transcriptional signature identified in the ASD Discovery Set generalizes to the ASD Replication set by assessing the 1st principal component of the DGE set, which summarizes the DGE expression pattern across all cortical samples. The ASD Discovery Set and ASD

Replication Set share this pattern, which is significantly different for both sets compared to CTL (Figure 1c). Moreover, this pattern is highly associated with ASD diagnosis, but not other biological factors, technical factors, or scores on sub-domains of an ASD diagnostic tool (Figure 1d). These analyses demonstrate that convergent differences in ASD CTX are reproducible in independent samples and are not related to confounding factors.

We also detected 2715 lncRNAs expressed in cerebral cortex (after careful filtering for high-confidence transcripts, Supplementary Information), of which 62 were significantly dysregulated between ASD and CTL (33 long intergenic RNAs, lincRNAs; 19 antisense transcripts; and 10 processed transcripts at FDR ≤ 0.05). Similar to the protein coding genes, these transcripts' expression patterns cluster ASD and CTL samples (Figure 1e). Most of these lncRNAs are developmentally regulated[12], have chromatin states indicative of transcription start sites (TSSs) near their 5´ end in brain[13], and are identified in other datasets[12,14] consistent with being valid, functional lncRNAs. Moreover, most (81%) exhibit primate-specific expression patterns in brain[15] (Supplementary Information). For example, Figure 1f depicts two lincRNAs, *LINC00693* and *LINC00689*, which are typically downregulated during development, yet are upregulated in ASD CTX relative to controls (Figure 1g), which we validated by RT-PCR (Extended Data Fig. 2d). *LINC00693* sequence is present, but poorly conserved in mouse, while *LINC00689* is primate-specific (present in macaque and other primates but not in any other species, Supplementary Information, Extended Data Fig. 3 for additional examples). These data indicate that dysregulation of lncRNAs, many of which are primate-specific and involved in brain development, is an important component of transcriptome dysregulation observed in ASD.

Previous work suggested that alterations in transcript splicing may contribute to transcriptomic changes in ASD[7,16,17] by evaluating splicing in a targeted manner and pooling samples across individuals[7,16,17]. Given the increased sequencing depth and reduced sequencing bias across transcript length in our dataset, we were able to perform an unbiased genome-wide analysis of differential alternative splicing (AS). We evaluated the percent spliced in (PSI, Extended Data Fig. 4a) for 34,025 AS events in CTX across the ASD Discovery Set, encompassing skipped exons (SE), alternative 5´ splice sites (A5SS), alternative 3´ splice sites (A3SS), and mutually exclusive exons (MXE) using the MATS pipeline[18] (Supplementary Information). We first asked whether there was a global signal, finding significant enrichment over background (Extended Data Fig. 4b). We identified 1127 events in 833 genes at FDR ≤ 0.5 in CTX (similar to the number of events at uncorrected P < 0.005). Importantly, we obtained similar results with a different splice junction mapping and quantification approach (Extended Data Fig. 4c).

We performed PCR validations with nine AS events from the differential splicing set (*ASTN2, MEF2D, ERC2, MED31, SMARCC2, SYNE1, NRCAM, GRIN1, NCAM*) and found that validated changes in splicing patterns were concordant with RNA-seq (Extended Data Fig. 4d-e), demonstrating that our approach identifies alterations in AS with high specificity. Similar to our observations with lncRNA and DGE, AS changes clustered the samples by diagnosis (Figure 2a). The most significantly different event was the inclusion of an exon in *ASTN2* (ΔPSI = 5.8 indicating a mean of 5.8% difference in inclusion in ASD vs CTL; $P = 7.8\times10^{-6}$), a gene implicated by copy number variation (CNV) in ASD and other developmental disorders[19]. GO term analysis of the genes implicated by these pathways indicates involvement of biological processes related to neuronal projection, biological adhesion, and morphogenesis (Figure 2b), pathways where alternative isoforms are critical to specifying interactions between protein products. Moreover, the 1st principal component of the cortex differential splicing signature replicates in the ASD Replication Set and is not associated with other biological or technical factors (Figures 2c-d, Extended Data Fig. 5a). Importantly, many splicing alterations occur in genes that are not differentially expressed

between ASD and CTL; removing AS events on genes exhibiting even nominal DGE ($P < 0.05$), still identified a strong difference between ASD and CTL CTX (Extended Data Fig. 5b).

A parallel analysis in cerebellum evaluating 32,954 AS events found no differentially regulated events significant at any multiple comparison correction thresholds (Extended Data Fig. 5c, Supplementary Table 3). There was no detectable global overlap between cerebellum and CTX above chance for events significant at $P < 0.05$ in both comparisons (fold enrichment = 1.1, P = 0.21). This suggests that AS alterations in ASD are largely confined to CTX cell types, consistent with the stronger overall DGE patterns observed in CTX versus cerebellum.

To further explore the underlying biology of AS dysregulation, we tested whether the shared splicing signature in ASD might be a product of perturbations in AS factors known to be important to neural development or preferentially expressed in neural tissue. We found that the expression levels of *RBFOX1*, *RBFOX2*, *SRRM4*, *NOVA1*, and *PTBP1* all had high correlations ($R^2 > 0.35$, FDR $\leq 0.05$) to AS alterations in CTX (Figure 2e), but not in cerebellum (Figure 2f). Furthermore, enrichment analysis revealed that most changes in cortical AS occur in neuron-specific exons that are excluded in ASD (exons with $\Delta PSI > 50\%$ in neurons overlap with exons excluded in ASD CTX, fold enrichment = 4.1, $P = 1.8 \times 10^{-7}$, Extended Data Fig. 5d).

To validate a regulatory relationship between splicing factors and these events, we evaluated experimental data from knockout, overexpression, and knockdown experiments for Rbfox1[20], SRRM4[21], and PTBP1[22], respectively . We found that exons regulated by each of these splicing factors were significantly enriched in the set of exons excluded in ASD (Figure 2g), while in contrast, there was no enrichment for targets of ESRP[23], a splicing factor involved in epithelial cell differentiation but not expressed in CTX. This shows that alterations in three splicing factors dysregulated in ASD regulate AS of the neuron-specific exons whose inclusion is dysregulated in ASD in CTX and not cerebellum, indicating selective alteration of neuronal splicing in ASD CTX. Remarkably, the expression patterns of these three splicing factors (and others for which appropriate validation experiments were unavailable) results in distinct clusters (Extended Data Fig. 5e), suggesting that subsets of splicing factors act in different individuals to mediate a common downstream AS alteration.

Taken together these results indicate global transcriptional alterations in ASD cerebral cortex, but not cerebellum at the level of protein coding transcripts, lncRNA and AS. Therefore, to determine how these different transcriptomic subcategories relate to each other in ASD, we compared the 1st PC for each type of transcriptomic alteration across individuals (Figure 2h). Remarkably, the PCs are highly correlated ($R^2 > 0.8$) indicating that the transcriptomic alteration is a unitary phenomenon across protein coding, noncoding, and splicing levels, rather than distinct forms of molecular alteration.

Previous analysis with gene expression microarrays in a small cohort suggested that the typical pattern of transcriptional differences between the frontal and temporal cortex may be attenuated in ASD[7]. To further test this possibility, we evaluated DGE between CTX regions (Supplementary Information) in 16 matched frontal and temporal CTX sample pairs from ASD and CTL subjects and found 551 genes differentially expressed between regions in controls, but only 51 in ASD (FDR $\leq 0.05$; Figure 3a). We refer to the set of 523 genes with this pattern in CTL, but not ASD as the "Attenuated Cortical Patterning" set. The attenuation of patterning is evident from the global distribution of test statistics between frontal and temporal CTX in ASD and CTL and genes in this set do not show a greater difference in variability in ASD versus controls compared to other genes (Kolmogorov-Smirnov test, two-tailed P = 0.11, Extended Data Fig. 6a).

We complemented this analysis with a machine learning approach using all 123 cortical samples, training a regularized regression model[24] to classify frontal versus temporal CTX with independent gene expression data from BrainSpan[25] (Extended Data Fig. 6b, Supplementary Information). Multiple approaches to training the classifier with BrainSpan can differentiate between frontal and temporal CTX in both CTL and ASD (Extended Data Fig. 6c-e), demonstrating that dissection and sample quality in our samples are of high quality. Loss of classification accuracy in ASD compared to CTL was observed when restricting the model to the genes with the most attenuated patterning in ASD (Extended Data Fig. 6f), demonstrating that attenuation of patterning generalizes across all samples. The Attenuated Cortical Patterning set includes multiple genes known to be involved in cell-cell communication and cortical patterning, such as *PCDH10*, *PCDH17*, *CDH12*, *MET*, and *PDGFD*, which was recently shown to mediate human specific aspects of cerebral cortical development[26]. GO term enrichment analysis of the Attenuated Cortical Patterning set identified enrichment for G protein coupled signaling, Wnt receptor signaling, and calcium binding, among several developmental processes (Figure 3b), and cell type enrichment analysis did not identify a strong preference for a particular cell type (Extended Data Fig. 6g).

To identify potential drivers of the alteration in cortical patterning, we evaluated transcription factor binding site enrichment upstream of genes in the Attenuated Cortical Patterning set (Supplementary Information), and found an enrichment of *SOX5* binding motifs (upstream of 364/523 genes, Figure 3c). Remarkably, *SOX5* itself belongs to the Attenuated Cortical Patterning set: while *SOX5* is differentially expressed between frontal and temporal CTX in CTL, it is not in ASD (Figure 3d). We thus predicted that if *SOX5* regulates cortically patterned genes, its expression should correlate with target gene transcript levels. Consistent with this prediction, we found that genes in the Attenuated Cortical Patterning set are anti-correlated with *SOX5* in CTL CTX, but not in ASD CTX (Figure 3e, top left; Wilcoxon rank sum test of R values, P = 0.01), suggesting that the normal role of SOX5 as a transcriptional repressor may be disrupted in ASD. We reasoned that a true loss of SOX5-mediated cortical patterning would be specific to the predicted SOX5 targets. Consistent with this, we find a loss of correlations between *SOX5* and predicted targets, but no difference in correlations between *SOX5* and non-targets in the Attenuated Cortical Patterning set (Figure 3e). Taken together, these findings show that a loss of regional patterning downstream of the transcriptional repressor *SOX5*, which plays a crucial role in glutamatergic neuron development[27,28], contributes to the loss of regional identity in ASD.

Gene expression changes in postmortem brain may be a consequence of genetic factors, environmental factors, or both. Brain tissue from individuals with ASD that harbor known, penetrant genetic causes are very rare. However, we were able to identify postmortem brain tissue from 8 subjects with one of the more common recurrent forms of ASD, Duplication 15q Syndrome (dup15q, which is present in about 0.5-1% of ASD cases, see Extended Data Fig. 7a for characterization of duplications). We performed RNA-seq across frontal and temporal cortex and compared DGE changes in dup15q with those observed in individuals with idiopathic ASD to better understand the extent to which the observed molecular pathology overlaps. As expected, most genes in the 15q11.1-13.2 duplicated region have higher expression in dup15q CTX compared to CTL (Figure 4a), although *SNRPN* and *SNURF* were notably downregulated. Conversely, no significant upregulation of genes in this region were identified in idiopathic ASD or controls. Strikingly, when we assessed genome-wide expression changes, we observed a strong signal of DGE in dup15q that widely overlaps with that of idiopathic ASD (fold changes at FDR $\leq 0.05$ in dup15q correlate with idiopathic ASD with $R^2 = 0.79$, Figure 4b). Moreover, the slope of the best-fit line through these changes is 2.0, indicating that on average, the transcriptional changes in dup15q CTX are highly similar, but twice the magnitude of those observed in ASD CTX.

Next, we sought to evaluate AS changes in dup15q. There is only one significant splicing change in the dup15q region (Supplementary Table 3), consistent with the idea that duplication in this region duplicates all isoforms of the genes, resulting in minimal alteration of transcript structure. Similar to DGE, global AS analysis in dup15q CTX vs to CTL CTX revealed a stronger, but highly overlapping signature with idiopathic ASD CTX (fold changes at FDR $\leq 0.2$ in dup15q agree correlate with idiopathic ASD with $R^2 = 0.66$) indicating that splicing changes in dup15q syndrome recapitulate those of idiopathic ASD (Figure 4c). The slope of the best-fit line through the PSI for spliced exons in dup15q CTX compared to those in ASD CTX is 2.5 similar to DGE. Notably, both gene expression and AS changes in dup15q implicated similar pathways as those found in idiopathic ASD (Extended Data Fig. 7c-d). Clustering dup15q samples and CTL samples using both the DGE set and the differential AS set showed that all dup15q samples cluster together (Figure 4d), as opposed to the more variable clustering of idiopathic ASD, supporting the hypothesis that this shared genetic abnormality leads to a more homogeneous molecular phenotype.

Next, to test whether this molecular ASD signature may be due to independent of postmortem or reactive effects (Supplementary Information), we compared our data with gene expression profiles from a iPSC-derived neurons (nIPSCs)[29] from dup15q were available, we could use these data to definitively reveal which changes in dup15q CTX are independent of postmortem or reactive effects (Supplementary Information), since such effects are not present *in vitro*. We observe that DGE in the 15q region is concordant with that seen in the nIPSCs (Figure 4e), even though the sample size is small and the analysis is likely underpowered. Upregulated changes in dup15q are also seen in nIPSCs (Figure 4f), consistent with our other statistical analyses showing limited effects of potential confounders. The very immature, fetal state of the nIPSCs[30] likely explains the absence of an enrichment signal for genes downregulated in postnatal ASD brain, which are enriched for genes involved in neurons with more mature synapses.

We next applied gene network analysis to construct an organizing framework to understand shared biological functions across idiopathic ASD and dup15q (combining the ASD Discovery Set, ASD Replication Set, and dup15q set). We utilized Weighted Gene Co-expression Network Analysis (WGCNA), which identifies groups of genes with shared expression patterns across samples (modules) from which shared biological function is inferred.  Modules identified via WGCNA can than be related to a range of relevant phenotypes and potential confounders[31,32]. We applied signed co-expression analysis and used bootstrapping to ensure the network was robust, and not dependent on any subset of samples (Supplementary Information), while controlling for technical factors and RNA quality ("Adjusted FPKM" levels, Methods). WGCNA identified 16 co-expression modules (Extended Data Fig. 8a, Supplementary Table 2), which are further characterized by their association to ASD (Extended Data Fig. 8b), enrichment for cell-type specific genes (Extended Data Fig. 8c), and enrichment for GO terms (Extended Data Fig. 9). Of the downregulated modules, three are associated with ASD and dup15q (M1/10/17) and one with dup15q only (M11). Five of the upregulated modules are associated with ASD and dup15q (M4/5/6/9/12) and one is specific to dup15q (M13) (Figure 5a, top). Additionally, we identified a module strongly enriched for genes from the Attenuated Cortical Patterning set and *Wnt* signaling that contains *SOX5* (M12; fold enrichment = 3.0, $P = 3 \times 10^{-8}$), verifying the strong relationship observed between the *Wnt* pathway regulating TF *SOX5* and attenuation of cortical patterning[33].

Notably, the modules identified here significantly overlap with previous patterns identified in ASD (asdM12$_{array}$ and asdM16$_{array}$[7]; Figure 5a, middle). We found that the ASD-associated modules identified by our larger sample size and RNA-seq provide significant refinement of previous observations by identifying more discrete biological processes related to cortical development[34], the post-synaptic density[35], and

lncRNAs (Figure 5a, bottom). For example, M1 overlaps a subset of asdM12$_{array}$ (fold enrichment = 5.7) and developmental modules (devM16 fold enrichment = 3.7), and is enriched for proteins found in the PSD and genes involved in calcium signaling and gated ion channel signaling. Another subset of asdM12$_{array}$, M10 (fold enrichment = 11) overlaps more with a mid-fetal upregulated cortical development module (devM13 fold enrichment = 4.0), and genes involved in secretory pathways and intracellular signaling. A third module, M17 shows the least overlap with asdM12$_{array}$ (fold enrichment = 2.2) and is related to energy metabolism. Notably, these three modules are enriched for neuron-specific genes (Extended Data Fig. 8c), but not all neuronal modules are down regulated in ASD (M3 is not altered in ASD CTX). Taken together, specific neurobiological processes are affected in individuals with ASD related to developmentally regulated neurodevelopmental processes.

The most upregulated modules, M5 and M9, both strongly overlap (fold enrichments > 20) with previously identified upregulated co-expression module asdM16$_{array}$. M5 is enriched for microglial cell markers and immune response pathways, whereas M9 is enriched for astrocyte markers and immune-mediated signaling and immune cell activation (Extended Data Fig. 8c, Extended Data Fig. 9). This analysis clearly separates the contributions of the coordinated biological processes of microglial activation and reactive astrocytosis, which were previously not distinguishable as separate modules[7]. Thus, our analysis pinpoints more specific biological pathways in idiopathic ASD than those previously identified and reveals that similar changes occur downstream of the genetic perturbation in dup15q.

We evaluated the relationship between the five modules most strongly associated with ASD (M1/5/9/10/17, which are supported by module-trait association analysis and gene set enrichment analysis, Supplementary Information), and found that there was a remarkably high anti-correlation between the eigengene of M5 and downregulated modules, particularly M1 ($R^2$ = 0.76) (Figure 5b). M1 (Figure 5c) is downregulated in ASD and enriched for genes at the PSD and genes involved in synaptic transmission, while M5 (Figure 5d) is enriched for microglial genes and cytokine activation. This strong anti-correlation between microglial signaling and synaptic signaling in ASD and dup15q provides evidence in humans for dysregulation of microglia-mediated synaptic pruning, as previously suggested[36].

Next, to determine the role of causal genetic variation, we evaluated enrichment of both rare genetic variants, focusing on genes affected by ASD associated gene disrupting (LGD) *de novo* mutations[37], and common variants[38,39]. Genes within three modules, M1, M3, and M12, show enrichment for common variation signal for ASD (Figure 5e, Methods). Remarkably, M12 (Figure 5f), which is related to cortical patterning and Wnt signaling, also exhibit GWAS signal enrichment, providing the first evidence that risk conferred by common variation in ASD may affect regionalization of the cortex. Interestingly, M3 is significantly enriched for both schizophrenia (SCZ) and ASD common variants, is related to synaptic transmission, nervous system development, and regulation of ion channel activity (Extended Data Fig. 9), consistent with the notion that ASD and SCZ share common and rare genetic risk[1,40-43].

We only identified one module, M2 (Figure 5g), as significantly enriched in protein disrupting (nonsense, splice site, or frameshift) rare *de novo* variants previously associated with SCZ and ASD. M2 overlaps with a cortical developmental module implicated in ASD[34] (devM2 fold enrichment = 5.1). Notably, M2 is not differential between ASD and CTL in our dataset, consistent with the observation that these genes are primarily expressed during early neuronal development in fetal brain[34]. Remarkably, M2 contains an unusually large fraction of lncRNAs (15% of the genes in M2 are classified as lncRNAs, while other modules are 1-5% lncRNA). We hypothesize that, in addition to protein coding genes involved in transcriptional and chromatin regulation, rare *de novo* variants may also affect lncRNAs in ASD, a prediction that will be testable once large sets of whole genome sequences are available.

These combined transcriptomic and genetic analyses reveal that different forms of genetic variation affect biological processes involved in multiple stages of cortical development. Common genetic risk is enriched in M3, M1, and M12, which reflect early glutamatergic neurogenesis, later neuronal function, and cortical patterning, respectively. We also observe that rare *de novo* variation, which is enriched in M2, affects distinct biology related to transcriptional regulation and chromatin modification. These findings are consistent with transcriptomic analyses of early prenatal brain development and ASD risk mutations that implicate chromatin regulation and glutamatergic neuron development[34,44].

We provide the first comprehensive picture of largely unexplored aspects of transcription in ASD, lncRNA and alternative splicing, and identify a strong convergent signal in these, as well as protein coding genes[7]. These results will aid in interpreting genetic variation outside of the known exome, as whole genome sequencing supplants current methods. A role of lncRNAs has been previously explored in ASD[45], but only two individuals were evaluated with targeted microarrays. We evaluate lncRNAs in an unbiased manner across many individuals, notably identifying an enrichment of lncRNAs in M2, most of which are uncharacterized in brain and arose on the primate lineage. The involvement of lncRNAs in this early developmental program that is enriched for *de novo* mutations implicated in ASD suggests their study will be particularly relevant to understanding the emergence of primate higher cognition on the mammalian lineage, and by extension human brain evolution[15,46,47].

We also provide the first confirmation of an attenuation of genes that typically show differential expression between frontal and temporal lobe in ASD CTX and further identified *SOX5*, known to regulate cortical laminar development[50,51], as a putative regulator of this disruption. That M12, which is enriched for genes exhibiting cortical regionalization and is also enriched in ASD GWAS signal, supports the prediction that attenuation of patterning may be mediated by common genetic variation in or near the *SOX5* target genes. Disruption of cortical lamination by direct effects on glutamatergic neurogenesis and function has been predicted by independent data, including network analyses of rare ASD associated variants identified in exome sequencing studies[34,44].

These data, in conjunction with previous studies, reveal a consistent picture of the ASD's emerging postnatal and adult pathology. Specific neuronal signaling and synaptic molecules are downregulated and astrocyte and microglial genes are upregulated in over 2/3 of cases. Microglial infiltration has been observed in ASD cortex with independent methods[52], and normal microglial pruning has been shown to be necessary for brain development[36]. Our findings further suggest that aberrant microglial-neuronal interactions may be pervasive in ASD and related to the gene expression signature seen in a majority of individuals. In our comprehensive AS analysis, we identify three splicing factors upstream of the altered splicing signature observed in ASD CTX. These factors are known to be involved in coordinating sequential processes in neuronal development[17,21] and maintaining neuronal function[48,49]. It may therefore be sufficient to disrupt any one of these factors to induce a similar outcome during brain development, which would be consistent with the shared downstream perturbation observed here.

Finally, evaluation of the transcriptome in dup15q supports the enormous value of the "genotype first" approach of studying syndromic forms of ASD, with known penetrant genetic lesions[53]. It is highly unlikely that the shared transcriptional dysregulation in dup15q is due to a shared environmental insult. Thus, the most parsimonious explanation for the convergent transcriptomic pathology seen in all dup15q and over 2/3 of the cases of idiopathic ASD is that it represents an adaptive or maladaptive response to a primary genetic insult, which in most cases of ASD will be genetic[2,54]. As future investigations pursue the full range of causal genetic variation contributing to ASD risk, these analyses and data will be valuable for interpreting genetic and epigenetic studies of ASD as well as those of other neuropsychiatric disorders.

Figure Legends

Figure 1 | Transcriptome-wide differential gene expression in ASD. a, Average linkage hierarchical clustering of samples in the ASD Discovery Set using the top 100 upregulated and top 100 downregulated protein coding genes. b, Gene Ontology (GO) term enrichment analysis of upregulated and downregulated genes in ASD. *FDR $\leq$ 0.05 across all GO terms and gene sets. c, 1st principal component of the CTX DGE set (CTX DGE PC1) is able to distinguish ASD and CTL samples, including independent samples from the ASD Replication Set. d, CTX DGE PC1 is primarily associated with diagnosis, and not other factors. e, Average linkage hierarchical clustering of ASD Discovery Set using all lncRNAs in the DGE set. f, UCSC genome browser track displaying reads per million (RPM) in a representative ASD and CTL sample, superimposed over the gene models and sequence conservation for genomic regions including *LINC00693* and *LINC00689*. g, *LINC00693* and *LINC00689* are upregulated across ASD samples and downregulated during frontal cortex development. Abbreviations: FC, frontal cortex; TC, temporal cortex; RIN, RNA integrity number; ADI-R score, Autism Diagnostic Interview Revised score; FPKM, fragments per kilobase million mapped reads.

Figure 2 | Alteration of alternative splicing in ASD. a, Average linkage hierarchical clustering of ASD discovery set using top 100 differentially included and top 100 differentially excluded exons from the differential splicing (DS) set across the ASD Discovery Set. b, Gene Ontology term enrichment analysis of genes with DS in ASD. c, 1st principal component 1 of the CTX differential alternative splicing set (CTX DS PC1) is able to distinguish ASD and CTL samples using independent samples from the ASD Replication Set. d, CTX DS PC1 is primarily associated with diagnosis, and not other factors. e, Correlation between CTX DS PC1 and gene expression of neuronal splicing factors in CTX. f, Correlation between 1st principal component of cerebellum differential splicing (CB DS PC1) and gene expression of neuronal splicing factors in cerebellum. g, Overlap between DS set and splicing events regulated by splicing factors where experimental data was available. h, Scatterplots and correlations between the 1st principal component across the ASD versus CTL DGE sets for different transcriptome subcategories. Abbreviations: FC, frontal cortex; TC, temporal cortex; RIN, RNA integrity number; ADI-R score, Autism Diagnostic Interview Revised score; FPKM, fragments per kilobase million mapped reads.

Figure 3 | Attenuation of cortical patterning in ASD cortex. a, Heatmap of 551 genes exhibiting cortical patterning between frontal cortex (FC) and temporal cortex (TC) in ASD, with samples sorted by diagnostic status and brain region. b, Gene ontology term enrichment analysis of genes exhibiting attenuated cortical patterning (ACP). c, Schematic of transcription factor motif enrichment upstream of genes in the ACP set, with the *SOX5* motif sequence logo. d, The *SOX5* gene exhibits attenuated cortical patterning in ASD CTX compared to CTLs. Lines connect FC-TC pairs that are from the same individual. e, Correlation between *SOX5* gene expression and predicted targets in CTL and ASD, with all ACP genes (top left), SOX5 targets from the ACP set (top right), SOX5 non-targets from the ACP set (bottom left), and all genes not in the ACP set (bottom right). Plots show the difference in correlation between *SOX5* and other genes in ASD and CTL ($\Delta$R).

Figure 4 | Duplication 15q Syndrome recapitulates transcriptomic changes in idiopathic ASD. a, DGE changes across the 15q11-13.2 region for ASD and dup15q compared to CTL, error bars are +/- 95% confidence intervals for the fold changes. b, Comparison of effect sizes in dup15q vs CTL and ASD vs

CTL, with changes in dup15q at FDR ≤ 0.05 highlighted. c, Comparison of differential splicing (DS) changes in dup15q vs CTL and ASD vs CTL, highlighting 402 events at FDR ≤ 0.2 in dup15q. d, Average linkage hierarchical clustering of dup15q samples and controls using the DGE and DS gene sets. e, Plot of fold changes between induced pluripotent stem cells differentiated into neurons (nIPSCs) from dup15q vs CTL and postmortem CTX DGE from dup15q vs CTL in the 15q region. f, Heatmap overlapping the top 1000 genes up- and down- regulated in the nIPSC comparison to the up- and down- regulated genes in dup15q and idiopathic ASD CTX.

Figure 5 | Co-expression network analysis across all ASD and CTL samples in CTX. a, Gene set enrichment analyses comparing the 16 co-expression modules with multiple gene sets from this RNA-seq study, from postmortem ASD CTX microarray, from human brain development, from the postsynaptic density and set of all brain-expressed lncRNAs. b, Comparison of five ASD-associated modules against each other by correlating module eigengenes. c, Module plot of M1 displaying the top 25 hub genes along with the module's Gene Ontology term enrichment. d, similar to c, but for M5. e, Gene set enrichment analysis with genome-wide whole-exome sequencing data (Rare *de novo* hit genes) and genome-wide association study (GWAS) results in ASD, schizophrenia (SCZ), and intellectual disability (ID). Boxes are filled if the odds ratio is greater than 0, and the enrichment $P < 0.05$. Asterisks* indicate FDR ≤ 0.05 across all comparisons in a and e. f,g, similar to c, but for M12 and M2, respectively. Abbreviations: LGD, likely gene disrupting, genes affected by nonsense, nonsynonymous, or splice-site mutations or frame-shift indels; AGRE, AGP/CHOP, and PGC refer to consortia that collect genetic data (Supplementary Information for details).

## Methods

Sample description: Brain tissue for ASD and control individuals was acquired from the Autism Tissue Program (ATP) brain bank at the Harvard Brain and Tissue Bank and the University of Maryland Brain and Tissue Bank (a Brain and Tissue Repository of the NIH NeuroBioBank). Sample acquisition protocols were followed for each brain bank, and samples were de-identified prior to acquisition. Brain sample and individual level metadata is available in Supplementary Table 1.

RNA-seq methodology: Starting with 1ug of total RNA, samples were rRNA depleted (RiboZero Gold, Illumina) and libraries were prepared using the TruSeq v2 kit (Illumina) to construct unstranded libraries with a mean fragment size of 150bp (range 100-300bp) that underwent 50bp paired end sequencing on an Illumina HiSeq 2000 or 2500 machine. Paired-end reads were mapped to hg19 using Gencode v18 annotations[55] via Tophat2[56]. Gene expression levels were quantified using union exon models with HTSeq[57]. For additional and information on sequencing and read alignment parameters, please see Supplementary Information.

Sample sets for analysis: For differential gene expression and splicing analysis, we defined an age matched set, referred to as the ASD Discovery Set (106 samples in CTX, 51 in cerebellum) of idiopathic ASD and control samples for the discovery set, and held out younger or unmatched samples as the ASD Discovery Set (17 in CTX, 8 in cerebellum). Dup15q individuals were analysed separately, utilizing the full set of controls from the ASD Discovery Set. For co-expression network analysis, we combined the discovery set, replication set, and dup15q individuals for a total of 137 CTX samples and 59 cerebellum samples.

Differential Gene Expression (DGE): DGE analysis was performed with expression levels adjusted for gene length, library size, and G+C content (referred to as "Normalized FPKM") Supplementary Information.

CTX samples (frontal and temporal) were analyzed separately from cerebellum samples. A linear mixed effects model framework was used to assess differential expression in log2(Normalized FPKM) values for each gene for cortical regions (as multiple brain regions were available from the same individuals) and a linear model was used for cerebellum (where one brain region was available in each individual, with a handful of technical replicates removed). Individual brain ID was treated as a random effect, while age, sex, brain region (except in the case of cerebellum, where there is only one region), and diagnoses were treated as fixed effects. We also used technical covariates accounting for RNA quality, library preparation, and batch effects as fixed effects into this model (Supplementary Information).

Reproducibility analyses: We assessed replication between datasets by evaluating the concordance between independent sample sets by comparing the squared correlation ($R^2$) of fold changes of genes in each sample set at a non-stringent P value threshold. This general approach has been shown to be effective for identifying reproducible gene expression patterns[58], and we modify it such that the P value threshold is set in one sample set (the *x* axis in the scatterplots), and the $R^2$ with fold changes in these genes are evaluated in an independent sample set (the *y* axis in the scatterplots).

Differential Splicing Analysis: Alternative splicing was quantified using the percent spliced in (PSI) metric using Multivariate Analysis of Transcript Splicing (MATS, v3.08)[18]. For each event, MATS reports counts supporting the inclusion (I) or exclusion (E) of a splicing event. To reduce spurious events due to low counts, we required at least 80% of samples to have I + S >= 10. For these events, the percent spliced in is calculated as PSI = I / (I + S) (Extended Data Fig. 4a). Statistical analysis for differential splicing was performed utilizing the linear mixed effects model regression framework as described above for DGE. This approach is advantageous over existing methods as it allows modeling of covariates and takes into consideration the variability in PSI across samples when assessing event significance with ASD (Supplementary Information).

Genotyping dup15q: For Dup15q samples, the type of duplication and copy number in the breakpoint 2-3 region were available for these brains[59]. To expand this to the regions between each of the recurrent breakpoint in these samples, 7/8 dup15q brains were genotyped (one was not genotyped due to limitations in tissue availability). The number of copies between each of the breakpoints is reported in Extended Data Fig. 7a.

Co-expression network analysis: The R package weighted gene co-expression network analysis (WGCNA) was used to construct co-expression networks using the technical variation normalized data[31,60] (referred to as “Adjusted FPKM”). We used the biweight midcorrelation to assess correlations between log2(Normalized FPKM) and parameters for network analysis are described in Supplementary Information. Notably, we utilized a modified version of WGCNA that involves bootstrapping the underlying dataset 100 times and constructing 100 networks. The consensus of these networks (50th percentile across all edges) was then used as the final network [32], ensuring that a handful of samples do not determine the network structure. For module-trait analyses, 1st principal component of each module (eigengene) was related to ASD diagnosis, age, sex, and brain region in a linear mixed effects framework as above, only replacing the expression values of each gene with the eigengene.

Enrichment analysis of gene sets and GWAS: Enrichment analyses were performed either with Fisher’s exact test (cell type and splicing factor enrichments) or logistic regression (all enrichment analyses in Figure 5). We used logistic regression in the latter case to control for gene length or other biases that may influence enrichment analysis (Supplementary Information). All GO term enrichment analysis was performed using GO Elite[61] with 10,000 permutations. We focused on molecular function and biological process terms for display purposes.

## Extended Data Figure Legends

Extended Data Figure 1 | Methodology, quality control, and differential expression replication analysis. a, RNA-seq workflow, including RNA extraction, library preparation, sequencing, read alignment, and quality control. b, RNA-seq quality and alignment statistics from this study, including RNA integrity number (RIN), number of aligned reads, proportion of reads mapping to different genomic features (mRNA, intronic, intergenic), and bias in coverage from the 5' to the 3' end of the top 1000 expressed transcripts (statistics compiled using PicardTools). c, Similar statistics as in b for another RNA-seq study that utilized polyA tail selection mRNA-seq to evaluate the transcriptome in ASD cortex[11] (primarily BA19, visual cortex, but also including some BA10/44 samples, frontal cortex). d, RNA-seq read coverage relative to normalized gene length across transcripts from the 5' to the 3' end in this study. e, Dependence between coverage and RIN across gene body (correlation between RIN and coverage in d across samples). f, Correlation of ASD vs CTL fold changes between previously evaluated and new ASD samples in CTX by microarray (left) and RNA-seq (right) using genes that were at $P < 0.05$ the samples from Voineagu et al., 2011. g, Correlation between effect sizes as in f, but for cerebellum (CB) samples. h,i, Correlation between covariates and ASD vs CTL status in CTX (h) and CB (i) in the ASD Discovery Set.

Extended Data Figure 2 | Transcriptome-wide differential gene expression (DGE) analysis in CTX. a, Comparison of P value rankings across different methods for DGE with Spearman's correlation. From left to right: removal of three additional principal components of sequencing statistics (Supplementary Information) related to RNA-sequencing quality, application of a permutation analysis for DGE P value computation, application of variance-weighted linear regression for DGE[62], and using surrogate variable analysis for DGE[63]. b, Average linkage hierarchical clustering heatmap using all genes DGE in the ASD Discovery Set, but including all idiopathic ASD frontal cortex (FC) and temporal cortex (TC) samples across 123 samples, combining the ASD Discovery set and the ASD Replication set. Bolded samples in the dendrogram are used for validation in d. c, Enrichment analysis of cell-type specific gene sets (5-fold enriched in the cell type compared to all other cells) with genes decreased and increased in ASD. d, RT-PCR validation of the two lincRNAs shown in Figure 1f-g, P values are computed with the Wilcoxon rank-sum test.

Extended Data Figure 3 | Gene browser tracks for selected primate-specific lncRNAs. For each lncRNA, expression for representative samples for ASD vs CTL (top) in human, macaque (middle), and mouse (bottom) are shown. The genome location for macaque and mouse displayed is syntenic to the human region, with the expected location of the lncRNA highlighted.

Extended Data Figure 4 | Splicing analyses and validation in ASD. a, Schematic describing how the percent spliced in (PSI) metric is computed. b, Distribution of *P* values for changes in the PSI between ASD and CTL in CTX for all events (left) and event subtypes (SE, spiced exon; A5SS, alternative 5' splice site; A3SS, alternative 3' splice site; MXE, mutually exclusive exons). c, Comparison of the CTX splicing analyses in when using PSI values obtained via read alignment by TopHat2[64] followed by the MATS[18] pipeline (used throughout this study) against read alignment by OLego followed by Quantas[65]. d, Comparison of ΔPSI values in nine splicing events between PCR and RNA-seq. e, PCR validation and sashimi plots for the nine splicing events delineated in d, from the samples highlighted in Extended Data Fig. 5a.

Extended Data Figure 5 | Additional splicing analyses in ASD. a, Average linkage hierarchical clustering heatmap using all differentially spiced (DS) events from the ASD Discovery Set, but including all idiopathic ASD neocortical samples (FC and TC) across 123 samples, combining the ASD Discovery set and the ASD Replication set. Bolded samples in the dendrogram were used for PCR validation in Extended Data Fig. 4. b, Top: difference between ASD and CTL in the DS set based on PC1 of the DS set at the PSI level, and PC1 of the gene expression levels of genes in the DS set. Bottom: Same comparison after differentially expressed genes ($p < 0.05$) are removed. c, Distribution of P values for changes in the PSI between ASD and CTL in cerebellum. d, Cell-type enrichment analysis of splicing events from CTX. e, Average-linkage hierarchical clustering using 1-(Pearson's correlation) to compare the gene expression patterns of the splicing factors investigated in Figure 2.

Extended Data Figure 6 | Attenuation of cortical patterning in ASD. a, Histograms of P values from paired Wilcoxon rank-sum test differential gene expression between 16 frontal cortex (FC) and 16 temporal cortex (TC) in CTL and ASD and a histogram of Bartlett's test P values for differences in gene expression variance between ASD and CTL for all genes (white) and genes in the Attenuated Cortical Patterning (ACP) set (red). c, Approach to training the elastic net model on BrainSpan and application of the model on 123 cortical samples in this study. c-e, Results of learned cortical region classifications with different starting gene sets, with the BrainSpan training set (left), CTL samples (middle), and ASD samples (right) in each panel and the Wilcoxon rank-sum test P value of FC vs TC difference for each comparison. f, Summary of results form c-e. g, Cell type enrichment analysis for genes in the ACP set. Abbreviations: A1C, primary auditory cortex; DFC, dorsolateral prefrontal cortex; MFC, medial prefrontal cortex; STC, superior temporal cortex; FC, frontal cortex; TC, temporal cortex; AUROC, area under the receiver-operator characteristic curve.

Extended Data Figure 7 | Dup15q syndrome analyses. a, Copy number between breakpoints (BP) in the 15q region. Genome-wide CNV analysis allowed evaluation of copy number in additional regions from previous studies[59,66]. b, Differential expression across the 15q region of interest in dup15q vs CTL and ASD vs CTL cerebellum, note only 3 samples were available for dup15q cerebellum so additional analyses were not pursued. c, Gene Ontology term enrichment analysis for the dup15q CTX differential expression set. d, Gene Ontology term enrichment analysis for the dup15q CTX differential splicing (DS) set. e, Hierarchical clustering of iPSC-derived neurons from dup15q, Angelman syndrome, and a control[29].

Extended Data Figure 8 | Co-expression network analysis in ASD CTX. a, Modules identified from a dendrogram constructed from a consensus of 100 bootstrapped datasets using the 137 CTX samples. Correlations for each gene to each measured factor are delineated below the dendrogram (blue = negative, red = positive correlation). b, Module-trait associations as computed by a linear mixed effects model with all factors on the x-axis used as covariates. All P values are displayed where the coefficient passed $p < 0.01$. Note that this alternative approach to module-trait association agrees with the Fisher's exact test used in Figure 5a when the fold enrichment for module overlap with DGE sets is > 2.8, and we use an intersection of both methods for the modules focused on in Figure 5b. c, Module enrichments for cell type specific gene expression patterns.

Extended Data Figure 9 | GO term enrichments for all modules. *FDR  < 0.05 across all GO enrichments across all modules.

# References

1. Geschwind, D. H. Genetics of autism spectrum disorders. *Trends Cogn. Sci. (Regul. Ed.)* 15, 409–416 (2011).
2. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* 46, 881–885 (2014).
3. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* (2015). doi:10.1016/S1474-4422(15)00044-7
4. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* 17, 782–790 (2014).
5. Chen, J. A., Peñagarikano, O., Belgard, T. G., Swarup, V. & Geschwind, D. H. The emerging picture of autism spectrum disorder: genetics and pathology. *Annu Rev Pathol* 10, 111–144 (2015).
6. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 9, 341–355 (2008).
7. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384 (2011).
8. Purcell, A. E., Jeon, O. H., Zimmerman, A. W., Blue, M. E. & Pevsner, J. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* 57, 1618–1628 (2001).
9. Garbett, K. *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiology of Disease* 30, 303–311 (2008).
10. Chow, M. L. *et al.* Age-Dependent Brain Gene Expression and Copy Number Anomalies in Autism Suggest Distinct Pathological Processes at Young Versus Mature Ages. *PLoS Genet.* 8, e1002592 (2012).
11. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Comms* 5, 5748 (2014).
12. Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Publishing Group* 18, 154–161 (2015).
13. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).
14. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* 47, 199–208 (2015).
15. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640 (2014).
16. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Reports* 6, 1139–1152 (2014).
17. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159, 1511–1523 (2014).
18. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* 40, e61–e61 (2012).
19. Lionel, A. C. *et al.* Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. *Human Molecular Genetics* 23, 2752–2768 (2014).
20. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* 20, 1434–1442 (2013).
21. Raj, B. *et al.* A Global Regulatory Mechanism for Activating an Exon Network Required for Neurogenesis. *Molecular Cell* 56, 90–103 (2014).
22. Gueroussov, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* 349, 868–873 (2015).
23. Dittmar, K. A. *et al.* Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Molecular and Cellular Biology* 32, 1468–1482 (2012).

24. Tibshirani, R., Johnstone, I., Hastie, T. & Efron, B. Least angle regression. *The Annals of Statistics* 32, 407–499 (2004).

25. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* 41, D996–D1008 (2013).

26. Lui, J. H. *et al.* Radial glia require PDGFD–PDGFRβ signalling in human but not mouse neocortex. *Nature* 515, 264–268 (2014).

27. Lai, T. *et al.* SOX5 Controls the Sequential Generation of Distinct Corticofugal Neuron Subtypes. *Neuron* 57, 232–247 (2008).

28. Kwan, K. Y. *et al.* SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16021–16026 (2008).

29. Germain, N. D. *et al.* Gene expression analysis of human induced pluripotent stem cell-derived neurons carrying copy number variants of chromosome 15q11-q13.1. *Mol Autism* 5, 44 (2014).

30. Stein, J. L. *et al.* A Quantitative Framework to Evaluate Modeling of Cortical Development by Neural Stem Cells. *Neuron* 83, 69–86 (2014).

31. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* 4, Article17 (2005).

32. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 1, 54 (2007).

33. Morales, P. L. M., Quiroga, A. C., Barbas, J. A. & Morales, A. V. SOX5 controls cell cycle progression in neural progenitors by interfering with the WNT–β-catenin pathway. *EMBO reports* 11, 466–472 (2010).

34. Parikshak, N. N. *et al.* Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* 155, 1008–1021 (2013).

35. Bayés, À. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* 14, 19–21 (2010).

36. Schafer, D. P. *et al.* Microglia Sculpt Postnatal Neural Circuits in an Activity and Complement-Dependent Manner. *Neuron* 74, 691–705 (2012).

37. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221 (2014).

38. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Human Molecular Genetics* 21, 4781–4792 (2012).

39. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. 459, 528–533 (2009).

40. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* 45, 984–994 (2013).

41. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506, 185–190 (2014).

42. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res* 25, 142–154 (2015).

43. Gilman, S. R. *et al.* Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nat. Neurosci.* 15, 1723–1728 (2012).

44. Willsey, A. J. *et al.* Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell* 155, 997–1007 (2013).

45. Ziats, M. N. & Rennert, O. M. Aberrant Expression of Long Noncoding RNAs in Autistic Brain. *J Mol Neurosci* 49, 589–593 (2012).

46. Geschwind, D. H. & Rakic, P. Cortical Evolution: Judge the Brain by Its Cover. *Neuron* 80, 633–647 (2013).

47. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated Recruitment of New Brain Development Genes into the Human Genome. *PLoS Biol* 9, e1001179 (2011).

48. Fogel, B. L. *et al.* RBFOX1 regulates both splicing and transcriptional networks in human neuronal

development. *Human Molecular Genetics* 21, 4171–4186 (2012).

49. Gehman, L. T. *et al.* The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nat Genet* 43, 706–711 (2011).

50. Greig, L. C., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat Rev Neurosci* 14, 755–769 (2013).

51. Srinivasan, K. *et al.* A network of genetic repression and derepression specifies projection fates in the developing neocortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19071–19078 (2012).

52. Morgan, J. T. *et al.* Abnormal microglial–neuronal spatial organization in the dorsolateral prefrontal cortex in autism. *Brain Research* 1456, 72–81 (2012).

53. Stessman, H. A., Bernier, R. & Eichler, E. E. A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell* 156, 872–877 (2014).

54. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet* 15, 133–141 (2014).

55. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7, S4–9 (2006).

56. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31, 46–53 (2012).

57. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).

58. Shi, L. *et al.* The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics* 9, S10 (2008).

59. Scoles, H. A., Urraca, N., Chadwick, S. W., Reiter, L. T. & LaSalle, J. M. Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. *Mol Autism* 2, 19 (2011).

60. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559 (2008).

61. Zambon, A. C. *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* 28, 2209–2210 (2012).

62. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15, R29 (2014).

63. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* 3, e161 (2007).

64. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578 (2012).

65. Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res* 41, 5149–5163 (2013).

66. Wintle, R. F. *et al.* A genotype resource for postmortem brain samples from the Autism Tissue Program. *Autism Res* 4, 89–97 (2011).

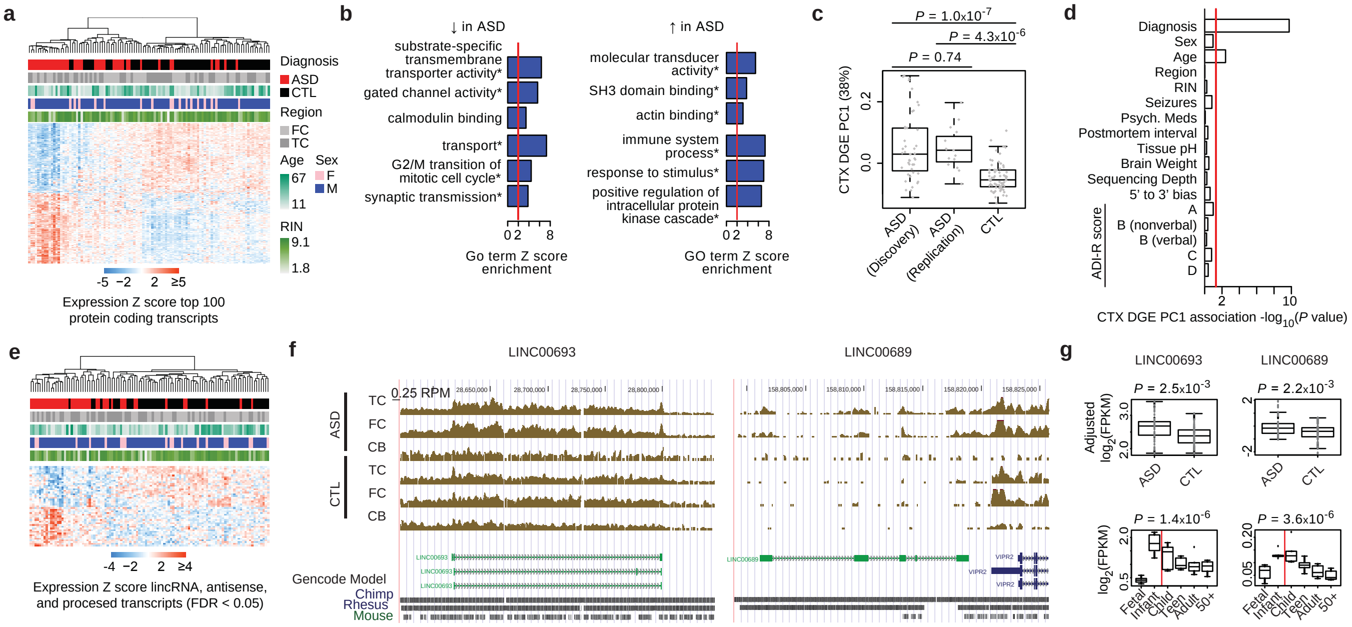
End Notes

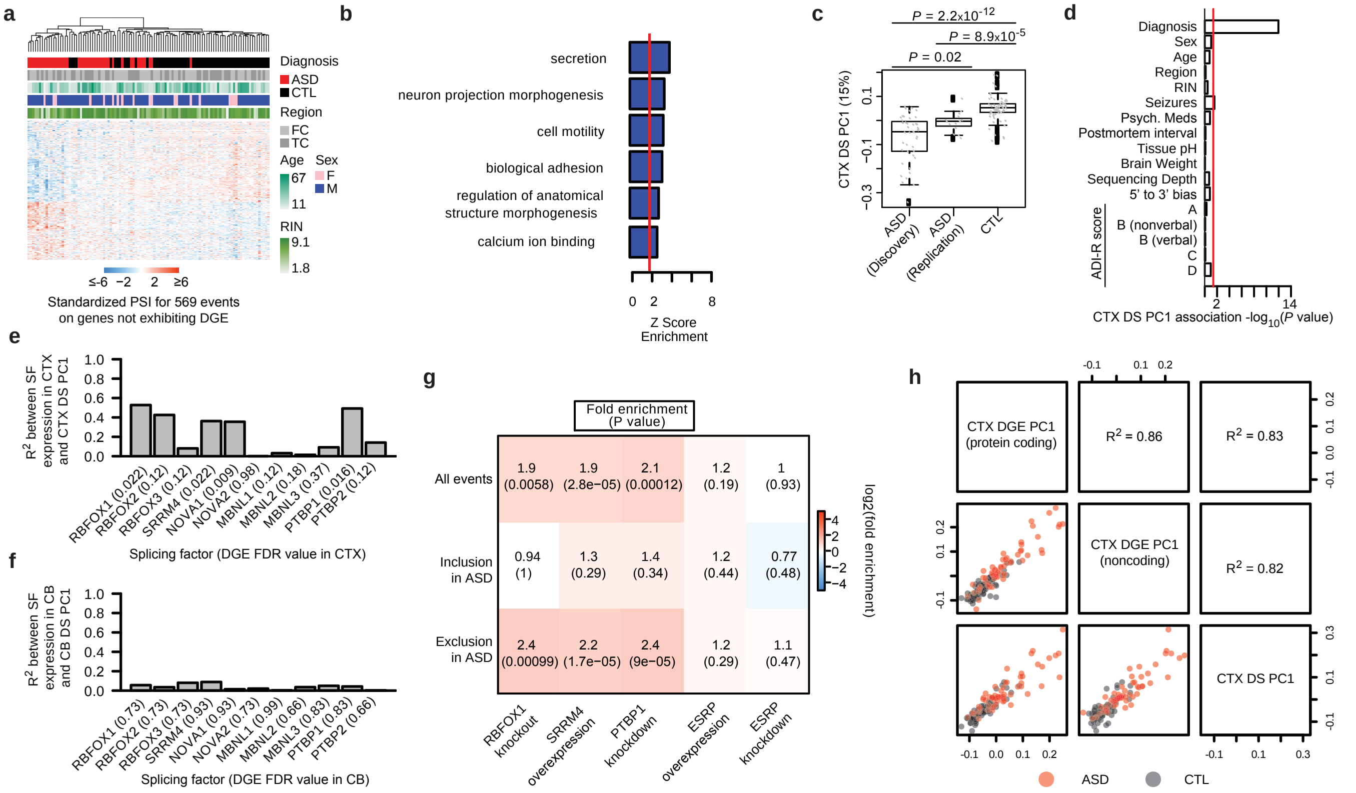
None.

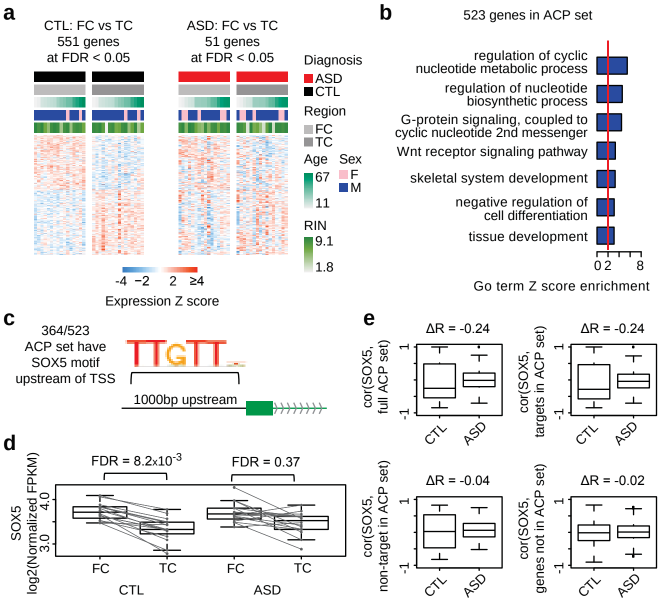
Author Contributions

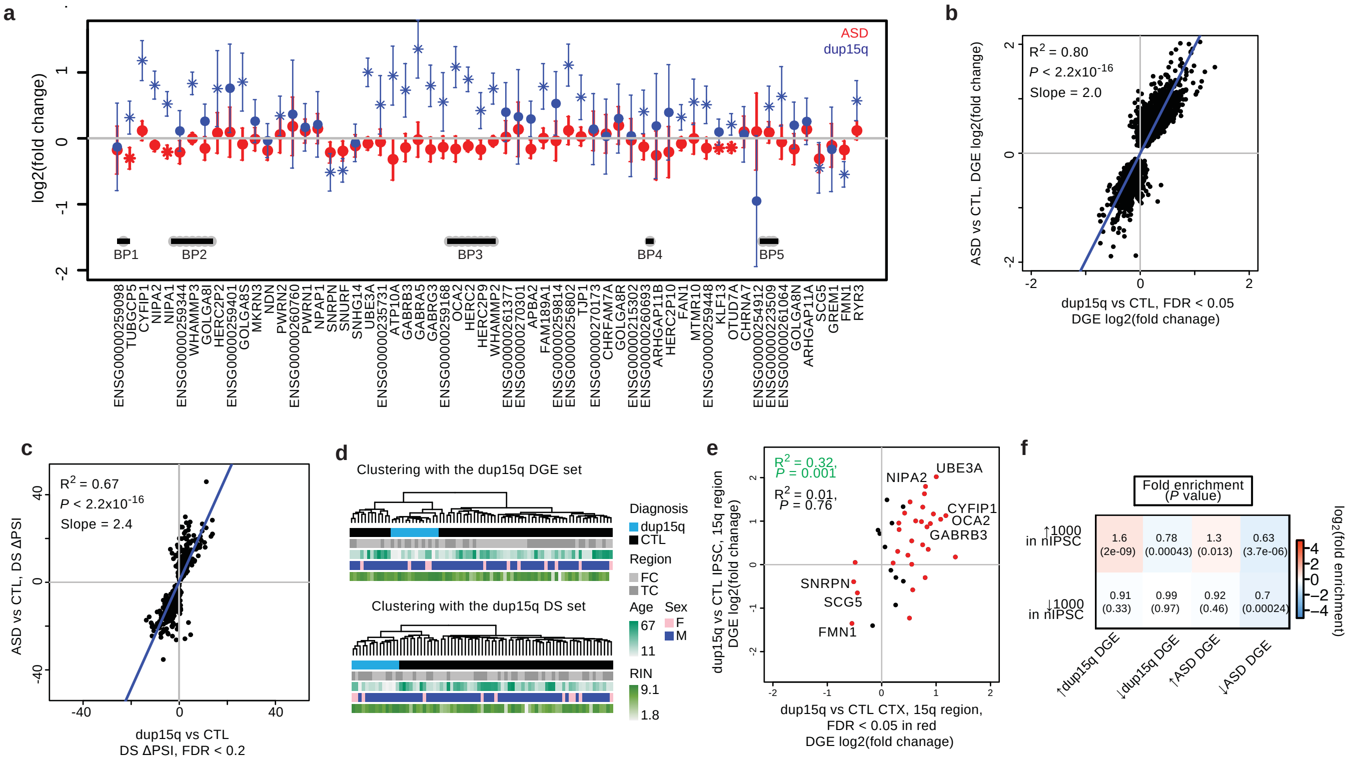
NNP, VS, and TGB performed dissections and RNA-seq analyses and differential gene expression analysis. NNP and VS performed splicing and co-expression network analysis. NNP and TGB performed analyses with Duplication 15q Syndrome individuals. NNP and MG reviewed clinical information and performed meta-analysis of ASD gene expression studies. VL and JKL performed

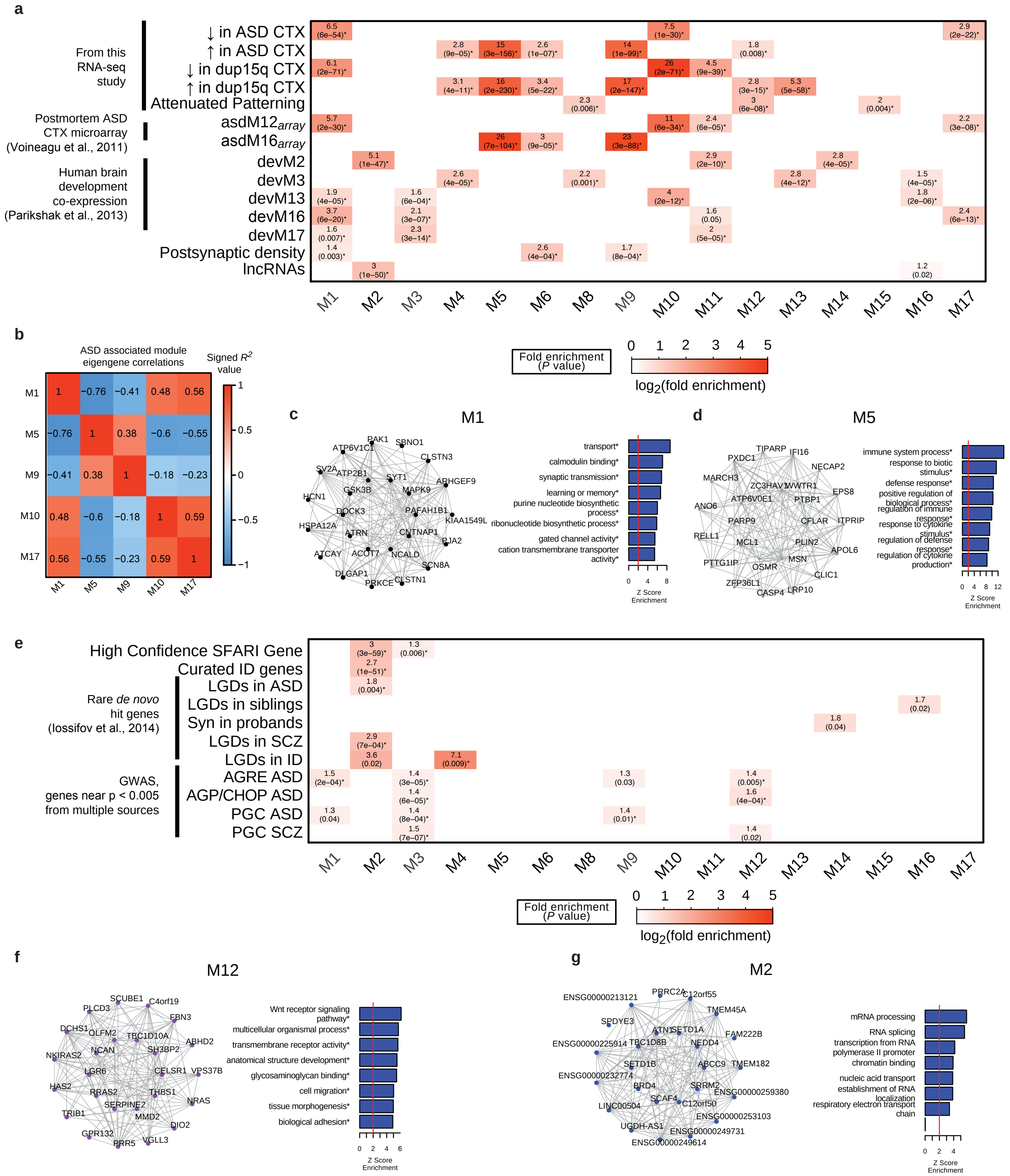
genotyping and CNV analysis on dup15q samples. VS performed validation experiments for gene and splicing level alterations in ASD. MI and BJB assisted with splicing analyses.  RJ performed dissections. SH provided guidance on differential gene expression and co-expression analyses. DHG provided guidance on all experiments and analyses. NNP and DHG wrote the manuscript. All authors contributed to revising and finalizing the manuscript.

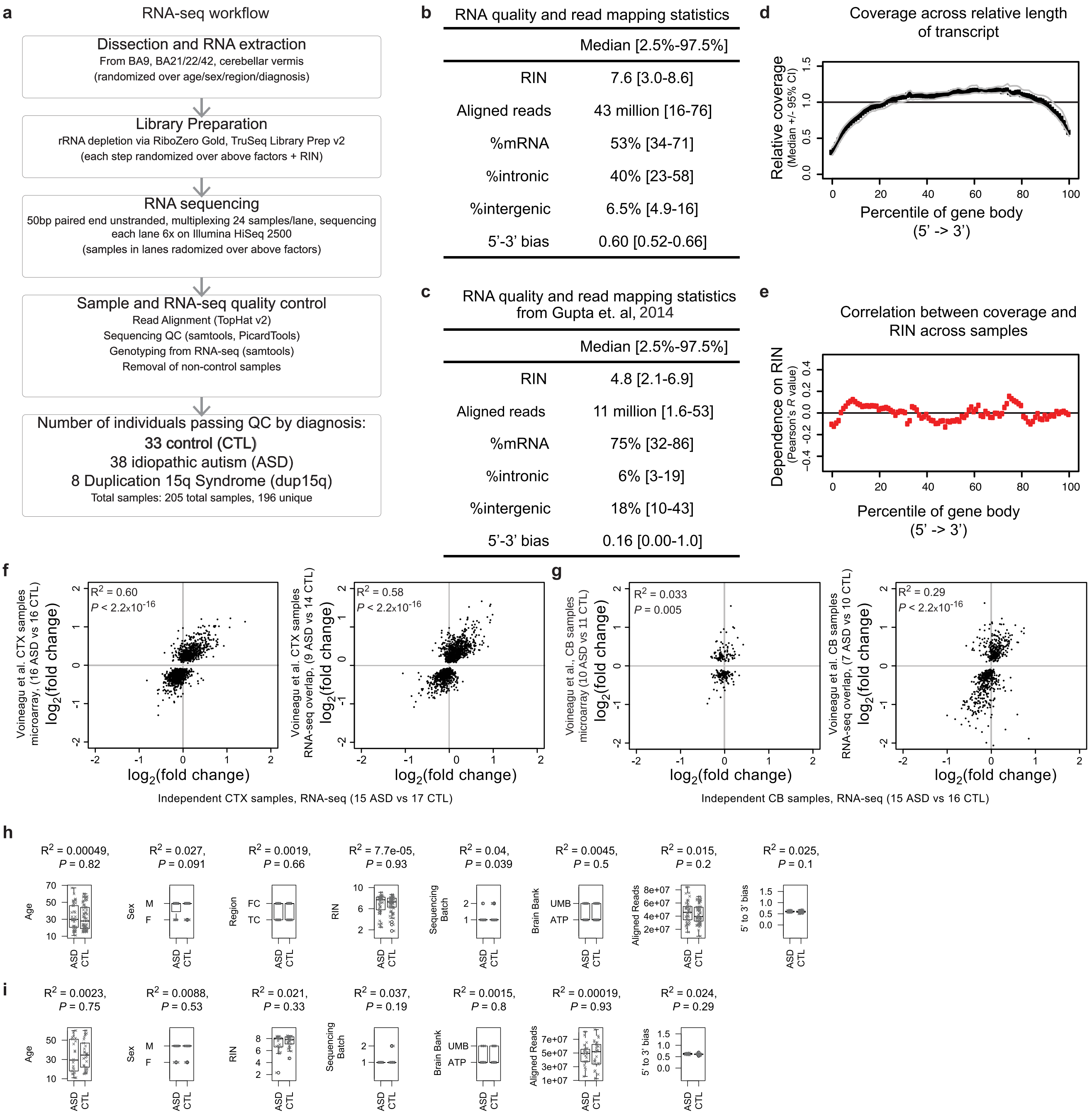

a
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
-5 −2 2 ≥5
Expression Z score top 100
protein coding transcripts
b
↓ in ASD
substrate-specific transmembrane transporter activity*
gated channel activity*
calmodulin binding
transport*
G2/M transition of mitotic cell cycle*
synaptic transmission*
0 2 8
Go term Z score enrichment
↑ in ASD
molecular transducer activity*
SH3 domain binding*
actin binding*
immune system process*
response to stimulus*
positive regulation of intracellular protein kinase cascade*
0 2 8
GO term Z score enrichment
c
P = 1.0x10-7
P = 4.3x10-6
P = 0.74
CTX DGE PC1 (38%)
0.2
0.0
ASD (Discovery)
ASD (Replication)
CTL
d
Diagnosis
Sex
Age
Region
RIN
Seizures
Psych. Meds
Postmortem interval
Tissue pH
Brain Weight
Sequencing Depth
5' to 3' bias
ADI-R score
A
B (nonverbal)
B (verbal)
C
D
2
10
CTX DGE PC1 association -log10(P value)
e
-4 −2 2 ≥4
Expression Z score lincRNA, antisense, and processed transcripts (FDR < 0.05)
f
LINC00693
LINC00689
0.25 RPM
ASD
TC
FC
CB
CTL
TC
FC
CB
28,650,000
28,700,000
28,750,000
28,800,000
158,805,000
158,810,000
158,815,000
158,820,000
158,825,000
Gencode Model
Chimp
Rhesus
Mouse
LINC00693
LINC00689
VIPR2
g
LINC00693
LINC00689
P = 2.5x10-3
P = 2.2x10-3
Adjusted log2(FPKM)
3.0
2.0
2
-2
ASD
CTL
P = 1.4x10-6
P = 3.6x10-6
log2(FPKM)
2.0
0.5
0.20
0.05
Fetal
Infant
Child
Teen
Adult
50+


a
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
≤-6 −2 2 ≥6
Standardized PSI for 569 events
on genes not exhibiting DGE
b
secretion
neuron projection morphogenesis
cell motility
biological adhesion
regulation of anatomical
structure morphogenesis
calcium ion binding
0 2 8
Z Score
Enrichment
c
P = 2.2x10^-12
P = 8.9x10^-5
P = 0.02
CTX DS PC1 (15%)
0.1
−0.1
−0.3
ASD
(Discovery)
ASD
(Replication)
CTL
d
Diagnosis
Sex
Age
Region
RIN
Seizures
Psych. Meds
Postmortem interval
Tissue pH
Brain Weight
Sequencing Depth
5' to 3' bias
ADI-R score
A
B (nonverbal)
B (verbal)
C
D
2
14
CTX DS PC1 association -log10(P value)
e
R2 between SF
expression in CTX
and CTX DS PC1
1.0
0.8
0.6
0.4
0.2
0.0
RBFOX1 (0.022)
RBFOX2 (0.12)
RBFOX3 (0.12)
SRRM4 (0.022)
NOVA1 (0.009)
NOVA2 (0.98)
MBNL1 (0.12)
MBNL2 (0.18)
MBNL3 (0.37)
PTBP1 (0.016)
PTBP2 (0.12)
Splicing factor (DGE FDR value in CTX)
f
R2 between SF
expression in CB
and CB DS PC1
1.0
0.8
0.6
0.4
0.2
0.0
RBFOX1 (0.73)
RBFOX2 (0.73)
RBFOX3 (0.73)
SRRM4 (0.93)
NOVA1 (0.93)
NOVA2 (0.73)
MBNL1 (0.99)
MBNL2 (0.66)
MBNL3 (0.83)
PTBP1 (0.83)
PTBP2 (0.66)
Splicing factor (DGE FDR value in CB)
g
Fold enrichment
(P value)
All events
Inclusion
in ASD
Exclusion
in ASD
1.9 (0.0058)
1.9 (2.8e−05)
2.1 (0.00012)
1.2 (0.19)
1 (0.93)
0.94 (1)
1.3 (0.29)
1.4 (0.34)
1.2 (0.44)
0.77 (0.48)
2.4 (0.00099)
2.2 (1.7e−05)
2.4 (9e−05)
1.2 (0.29)
1.1 (0.47)
RBFOX1
knockout
SRRM4
overexpression
PTBP1
knockdown
ESRP
overexpression
ESRP
knockdown
log2(fold enrichment)
4
2
0
−2
−4
h
CTX DGE PC1
(protein coding)
R2 = 0.86
R2 = 0.83
CTX DGE PC1
(noncoding)
R2 = 0.82
CTX DS PC1
-0.1 0.1 0.2
-0.1 0.0 0.1 0.2
-0.1 0.1 0.3
ASD
CTL


a
CTL: FC vs TC
551 genes
at FDR < 0.05
ASD: FC vs TC
51 genes
at FDR < 0.05
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
-4 −2 2 ≥4
Expression Z score
b
523 genes in ACP set
regulation of cyclic nucleotide metabolic process
regulation of nucleotide biosynthetic process
G-protein signaling, coupled to cyclic nucleotide 2nd messenger
Wnt receptor signaling pathway
skeletal system development
negative regulation of cell differentiation
tissue development
0 2 8
Go term Z score enrichment
c
364/523
ACP set have
SOX5 motif
upstream of TSS
TTGTT
1000bp upstream
d
FDR = 8.2x10^-3
FDR = 0.37
SOX5
log2(Normalized FPKM)
4.0
3.0
FC
TC
FC
TC
CTL
ASD
e
ΔR = -0.24
cor(SOX5, full ACP set)
ΔR = -0.24
cor(SOX5, targets in ACP set)
ΔR = -0.04
cor(SOX5, non-target in ACP set)
ΔR = -0.02
cor(SOX5, genes not in ACP set)
1
-1
CTL
ASD


a
ASD
dup15q
log2(fold change)
1
0
-1
-2
BP1
BP2
BP3
BP4
BP5
ENSG00000259098
TUBGCP5
CYFIP1
NIPA2
NIPA1
ENSG00000259344
WHAMMP3
GOLGA8I
HERC2P2
ENSG00000259401
GOLGA8S
MKRN3
NDN
PWRN2
ENSG00000260760
PWRN1
NPAP1
SNRPN
SNURF
SNHG14
UBE3A
ENSG00000235731
ATP10A
GABRB3
GABRA5
GABRG3
ENSG00000259168
OCA2
HERC2
HERC2P9
WHAMMP2
ENSG00000261377
ENSG00000270301
APBA2
FAM189A1
ENSG00000259814
ENSG00000256802
TJP1
ENSG00000270173
CHRFAM7A
GOLGA8R
ENSG00000215302
ENSG00000260693
ARHGAP11B
HERC2P10
FAN1
MTMR10
ENSG00000259448
KLF13
OTUD7A
CHRNA7
ENSG00000254912
ENSG00000223509
ENSG00000261064
GOLGA8N
ARHGAP11A
SCG5
GREM1
FMN1
RYR3
b
R$^2$ = 0.80
$P$ < 2.2x10$^{-16}$
Slope = 2.0
ASD vs CTL, DGE log2(fold change)
2
0
-2
-2
0
2
dup15q vs CTL, FDR < 0.05
DGE log2(fold chanage)
c
R$^2$ = 0.67
$P$ < 2.2x10$^{-16}$
Slope = 2.4
ASD vs CTL, DS ΔPSI
40
0
-40
-40
0
40
dup15q vs CTL
DS ΔPSI, FDR < 0.2
d
Clustering with the dup15q DGE set
Clustering with the dup15q DS set
Diagnosis
dup15q
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
e
R$^2$ = 0.32,
$P$ = 0.001
R$^2$ = 0.01,
$P$ = 0.76
UBE3A
NIPA2
CYFIP1
OCA2
GABRB3
SNRPN
SCG5
FMN1
dup15q vs CTL IPSC, 15q region
DGE log2(fold chanage)
2
1
0
-1
-2
-2
-1
0
1
2
dup15q vs CTL CTX, 15q region,
FDR < 0.05 in red
DGE log2(fold chanage)
f
Fold enrichment
($P$ value)
↑1000
in nIPSC
↓1000
in nIPSC
1.6
(2e-09)
0.78
(0.00043)
1.3
(0.013)
0.63
(3.7e-06)
0.91
(0.33)
0.99
(0.97)
0.92
(0.46)
0.7
(0.00024)
↑dup15q DGE
↓dup15q DGE
↑ASD DGE
↓ASD DGE
log$_2$(fold enrichment)
4
2
0
−2
−4


a
From this RNA-seq study
↓ in ASD CTX
↑ in ASD CTX
↓ in dup15q CTX
↑ in dup15q CTX
Attenuated Patterning
Postmortem ASD CTX microarray (Voineagu et al., 2011)
asdM12array
asdM16array
Human brain development co-expression (Parikshak et al., 2013)
devM2
devM3
devM13
devM16
devM17
Postsynaptic density
lncRNAs
↓ in ASD CTX: M1 6.5 (6e−54)*; M10 7.5 (1e−30)*; M17 2.9 (2e−22)*
↑ in ASD CTX: M4 2.8 (9e−05)*; M5 15 (3e−156)*; M6 2.6 (1e−07)*; M9 14 (1e−99)*; M12 1.8 (0.008)*
↓ in dup15q CTX: M1 6.1 (2e−71)*; M10 26 (2e−71)*; M11 4.5 (9e−39)*
↑ in dup15q CTX: M4 3.1 (4e−11)*; M5 16 (2e−230)*; M6 3.4 (5e−22)*; M9 17 (2e−147)*; M12 2.8 (3e−15)*; M13 5.3 (5e−58)*
Attenuated Patterning: M8 2.3 (0.006)*; M12 3 (6e−08)*; M15 2 (0.004)*
asdM12array: M1 5.7 (2e−30)*; M10 11 (6e−34)*; M11 2.4 (6e−05)*; M17 2.2 (3e−08)*
asdM16array: M5 26 (7e−104)*; M6 3 (9e−05)*; M9 23 (3e−88)*
devM2: M2 5.1 (1e−47)*; M11 2.9 (2e−10)*; M14 2.8 (4e−05)*
devM3: M4 2.6 (4e−05)*; M8 2.2 (0.001)*; M13 2.8 (4e−12)*; M16 1.5 (4e−05)*
devM13: M1 1.9 (4e−05)*; M3 1.6 (6e−04)*; M10 4 (2e−12)*; M16 1.8 (2e−06)*
devM16: M1 3.7 (6e−20)*; M3 2.1 (3e−07)*; M11 1.6 (0.05); M17 2.4 (6e−13)*
devM17: M1 1.6 (0.007)*; M3 2.3 (3e−14)*; M11 2 (5e−05)*
Postsynaptic density: M1 1.4 (0.003)*; M6 2.6 (4e−04)*; M9 1.7 (8e−04)*
lncRNAs: M2 3 (1e−50)*; M16 1.2 (0.02)
M1 M2 M3 M4 M5 M6 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17
Fold enrichment (P value)
0 1 2 3 4 5
log2(fold enrichment)
b
ASD associated module eigengene correlations
Signed R2 value
1 0.5 0 −0.5 −1
M1: 1 −0.76 −0.41 0.48 0.56
M5: −0.76 1 0.38 −0.6 −0.55
M9: −0.41 0.38 1 −0.18 −0.23
M10: 0.48 −0.6 −0.18 1 0.59
M17: 0.56 −0.55 −0.23 0.59 1
M1 M5 M9 M10 M17
c
M1
PAK1 SBNO1 ATP6V1C1 CLSTN3 SV2A ATP2B1 SYT1 ARHGEF9 HCN1 GSK3B MAPK9 DOCK3 PAFAH1B1 KIAA1549L HSPA12A ATRN CNTNAP1 PJA2 ATCAY ACOT7 NCALD SCN8A DLGAP1 PRKCE CLSTN1
transport*
calmodulin binding*
synaptic transmission*
learning or memory*
purine nucleotide biosynthetic process*
ribonucleotide biosynthetic process*
gated channel activity*
cation transmembrane transporter activity*
0 4 8
Z Score Enrichment
d
M5
TIPARP IFI16 PXDC1 NECAP2 MARCH3 ZC3HAV1 WWTR1 EPS8 ATP6V0E1 PTBP1 ANO6 PARP9 CFLAR ITPRIP RELL1 MCL1 PLIN2 APOL6 PTTG1IP OSMR MSN CLIC1 ZFP36L1 CASP4 LRP10
immune system process*
response to biotic stimulus*
defense response*
positive regulation of biological process*
regulation of immune response*
response to cytokine stimulus*
regulation of defense response*
regulation of cytokine production*
0 4 8 12
Z Score Enrichment
e
High Confidence SFARI Gene
Curated ID genes
Rare de novo hit genes (Iossifov et al, 2014)
LGDs in ASD
LGDs in siblings
Syn in probands
LGDs in SCZ
LGDs in ID
GWAS, genes near p < 0.005 from multiple sources
AGRE ASD
AGP/CHOP ASD
PGC ASD
PGC SCZ
High Confidence SFARI Gene: M2 3 (3e−59)*; M3 1.3 (0.006)*
Curated ID genes: M2 2.7 (1e−51)*
LGDs in ASD: M2 1.8 (0.004)*
LGDs in siblings: M16 1.7 (0.02)
Syn in probands: M14 1.8 (0.04)
LGDs in SCZ: M2 2.9 (7e−04)*
LGDs in ID: M2 3.6 (0.02); M4 7.1 (0.009)*
AGRE ASD: M1 1.5 (2e−04)*; M3 1.4 (3e−05)*; M9 1.3 (0.03); M12 1.4 (0.005)*
AGP/CHOP ASD: M3 1.4 (6e−05)*; M12 1.6 (4e−04)*
PGC ASD: M1 1.3 (0.04); M3 1.4 (8e−04)*; M9 1.4 (0.01)*
PGC SCZ: M3 1.5 (7e−07)*; M12 1.4 (0.02)
M1 M2 M3 M4 M5 M6 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17
Fold enrichment (P value)
0 1 2 3 4 5
log2(fold enrichment)
f
M12
SCUBE1 C4orf19 PLCD3 FBN3 DCHS1 OLFM2 TBC1D10A ABHD2 SH3BP2 NKIRAS2 NCAN LGR6 CELSR1 VPS37B HAS2 RRAS2 THBS1 SERPINE2 NRAS TRIB1 MMD2 DIO2 GPR132 PRR5 VGLL3
Wnt receptor signaling pathway*
multicellular organismal process*
transmembrane receptor activity*
anatomical structure development*
glycosaminoglycan binding*
cell migration*
tissue morphogenesis*
biological adhesion*
0 2 4 6
Z Score Enrichment
g
M2
ENSG00000213121 PRRC2A C12orf55 TMEM45A SPDYE3 ATN1 SETD1A FAM222B ENSG00000225914 TBC1D8B NEDD4 SETD1B ABCC9 TMEM182 ENSG00000232774 BRD4 SRRM2 ENSG00000259380 SCAF4 C12orf50 LINC00504 ENSG00000253103 UGDH-AS1 ENSG00000249731 ENSG00000249614
mRNA processing
RNA splicing
transcription from RNA polymerase II promoter
chromatin binding
nucleic acid transport
establishment of RNA localization
respiratory electron transport chain
0 2 4
Z Score Enrichment



a RNA-seq workflow

Dissection and RNA extraction
From BA9, BA21/22/42, cerebellar vermis
(randomized over age/sex/region/diagnosis)

Library Preparation
rRNA depletion via RiboZero Gold, TruSeq Library Prep v2
(each step randomized over above factors + RIN)

RNA sequencing
50bp paired end unstranded, multiplexing 24 samples/lane, sequencing each lane 6x on Illumina HiSeq 2500
(samples in lanes radomized over above factors)

Sample and RNA-seq quality control
Read Alignment (TopHat v2)
Sequencing QC (samtools, PicardTools)
Genotyping from RNA-seq (samtools)
Removal of non-control samples

Number of individuals passing QC by diagnosis:
33 control (CTL)
38 idiopathic autism (ASD)
8 Duplication 15q Syndrome (dup15q)
Total samples: 205 total samples, 196 unique

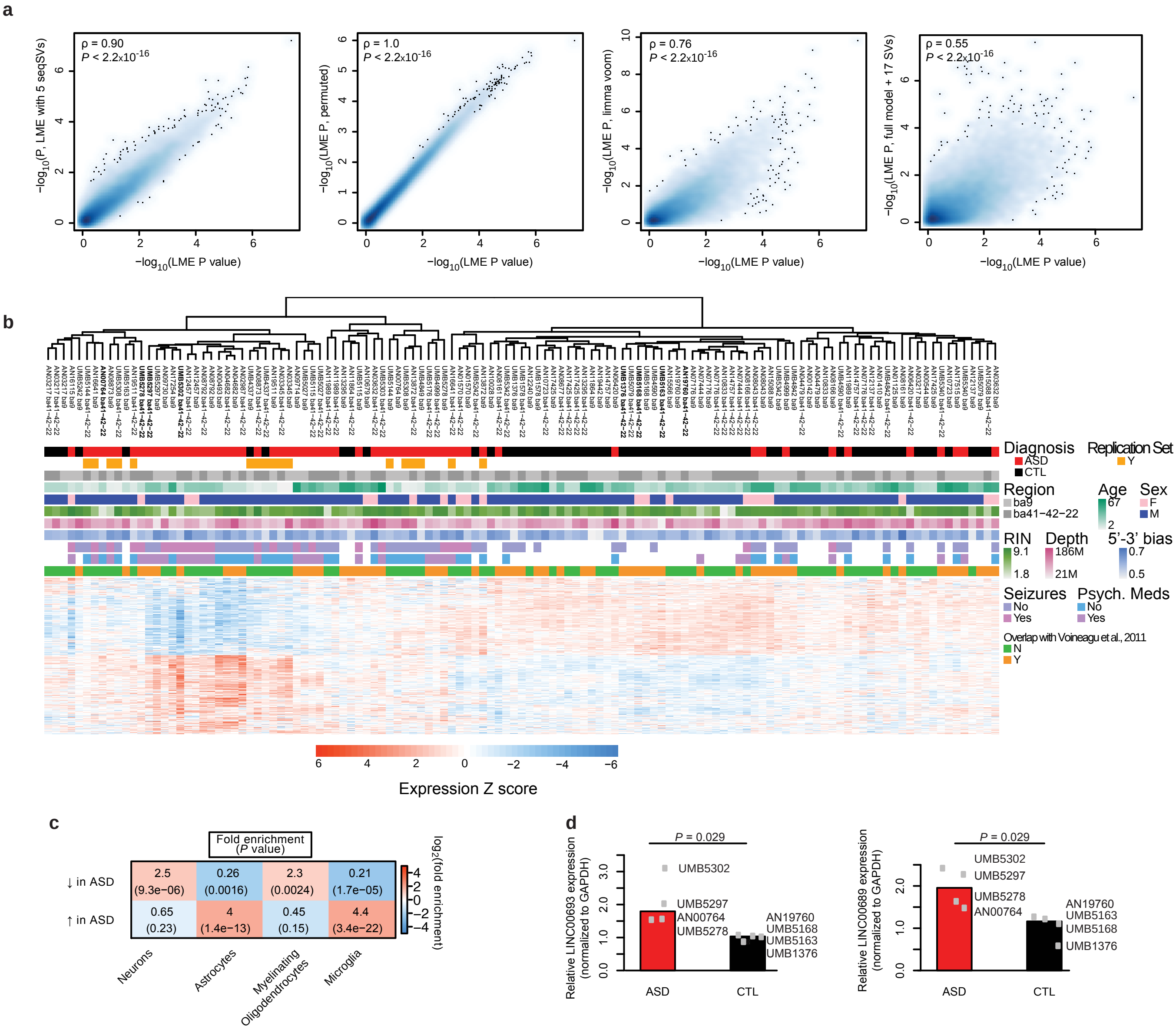
b RNA quality and read mapping statistics

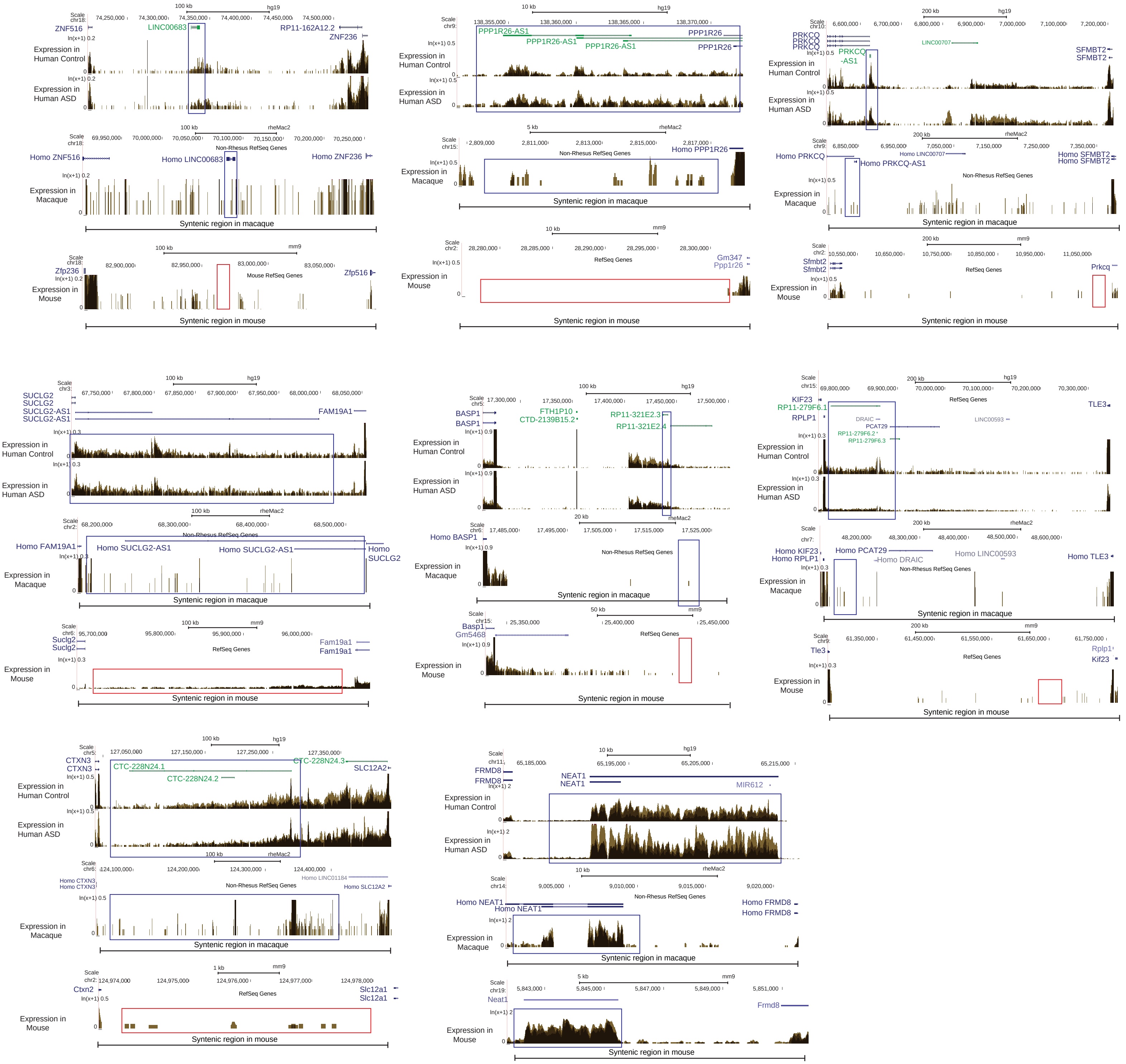
| | Median [2.5%-97.5%] |
|---|---|
| RIN | 7.6 [3.0-8.6] |
| Aligned reads | 43 million [16-76] |
| %mRNA | 53% [34-71] |
| %intronic | 40% [23-58] |
| %intergenic | 6.5% [4.9-16] |
| 5'-3' bias | 0.60 [0.52-0.66] |

c RNA quality and read mapping statistics from Gupta et. al, 2014

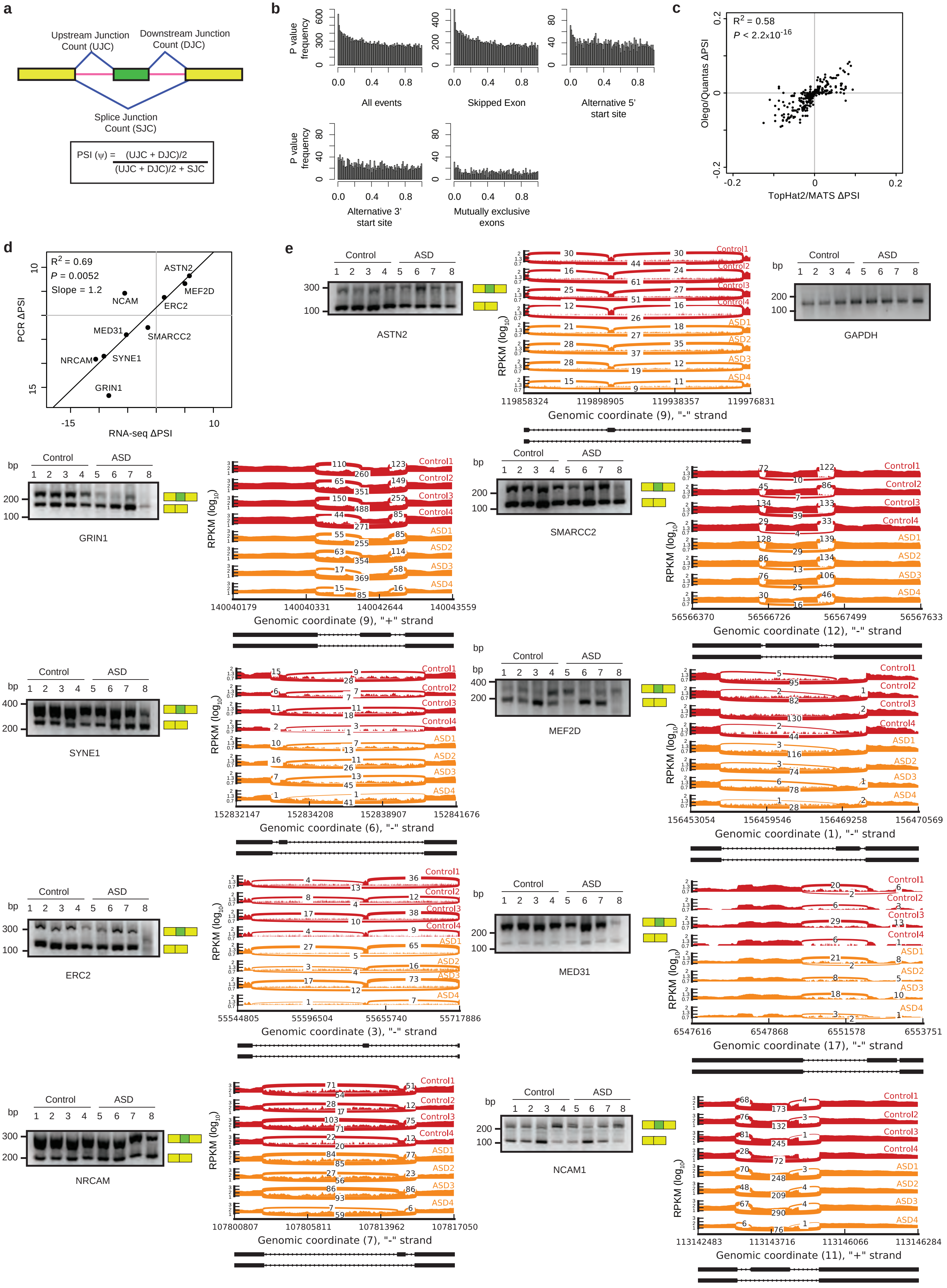
| | Median [2.5%-97.5%] |
|---|---|
| RIN | 4.8 [2.1-6.9] |
| Aligned reads | 11 million [1.6-53] |
| %mRNA | 75% [32-86] |
| %intronic | 6% [3-19] |
| %intergenic | 18% [10-43] |
| 5'-3' bias | 0.16 [0.00-1.0] |

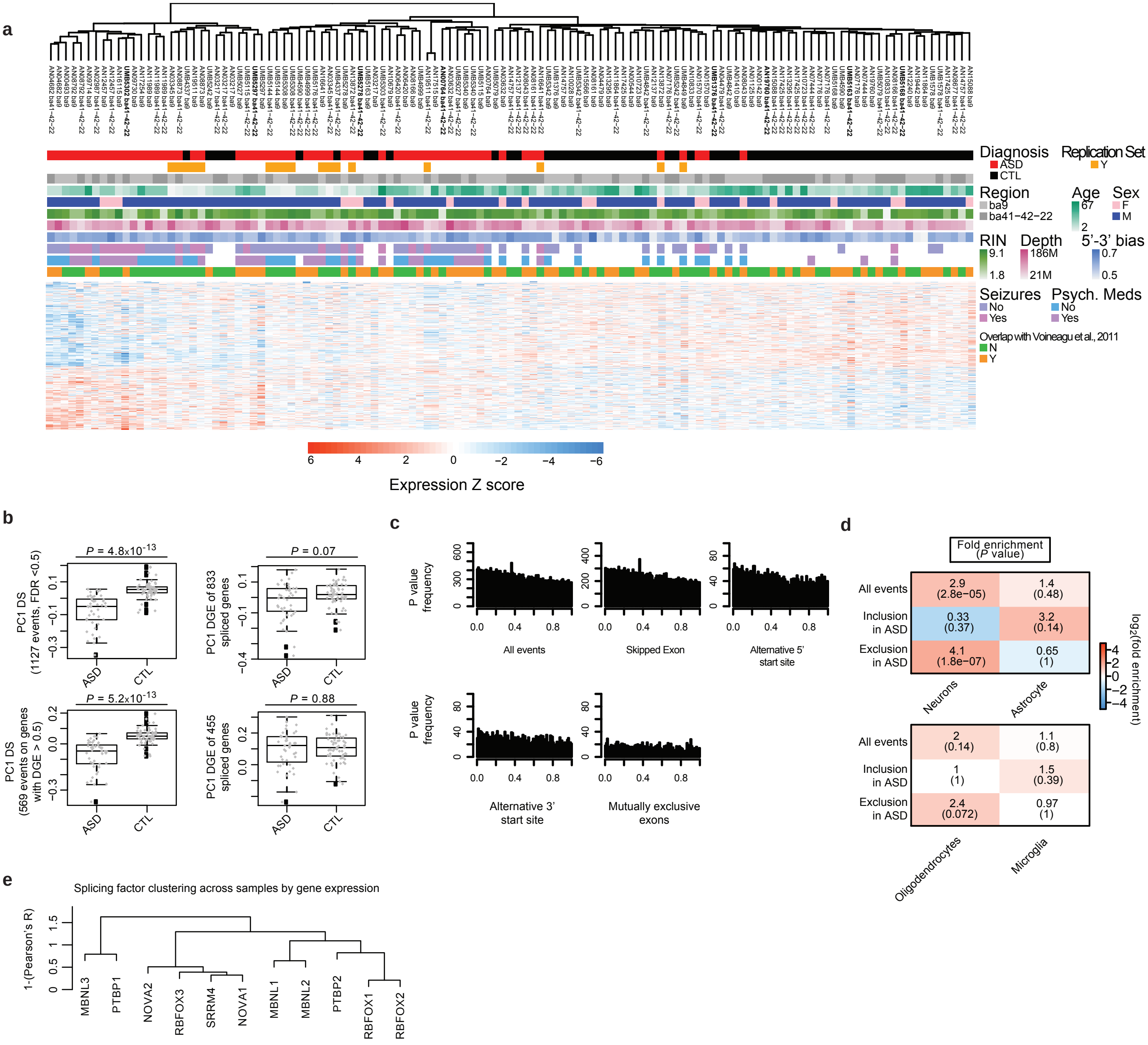
d
Coverage across relative length of transcript
Relative coverage (Median +/- 95% CI)
0.0
0.5
1.0
1.5
0
20
40
60
80
100
Percentile of gene body (5' -> 3')
e
Correlation between coverage and RIN across samples
Dependence on RIN (Pearson's R value)
−0.4
−0.2
0.0
0.2
0.4
0
20
40
60
80
100
Percentile of gene body (5' -> 3')
f
Voineagu et al. CTX samples microarray, (16 ASD vs 16 CTL) $\log_2$(fold change)
$R^2$ = 0.60
$P < 2.2 \times 10^{-16}$
Voineagu et al. CTX samples RNA-seq overlap, (9 ASD vs 14 CTL) $\log_2$(fold change)
$R^2$ = 0.58
$P < 2.2 \times 10^{-16}$
-2
-1
0
1
2
$\log_2$(fold change)
Independent CTX samples, RNA-seq (15 ASD vs 17 CTL)
g
Voineagu et al., CB samples microarray (10 ASD vs 11 CTL) $\log_2$(fold change)
$R^2$ = 0.033
$P$ = 0.005
Voineagu et al. CB samples RNA-seq overlap, (7 ASD vs 10 CTL) $\log_2$(fold change)
$R^2$ = 0.29
$P < 2.2 \times 10^{-16}$
$\log_2$(fold change)
Independent CB samples, RNA-seq (15 ASD vs 16 CTL)
h
$R^2$ = 0.00049, $P$ = 0.82
Age
10
30
50
70
$R^2$ = 0.027, $P$ = 0.091
Sex
M
F
$R^2$ = 0.0019, $P$ = 0.66
Region
FC
TC
$R^2$ = 7.7e-05, $P$ = 0.93
RIN
2
6
10
$R^2$ = 0.04, $P$ = 0.039
Sequencing Batch
1
2
$R^2$ = 0.0045, $P$ = 0.5
Brain Bank
UMB
ATP
$R^2$ = 0.015, $P$ = 0.2
Aligned Reads
2e+07
4e+07
6e+07
8e+07
$R^2$ = 0.025, $P$ = 0.1
5' to 3' bias
0.0
0.5
1.0
1.5
ASD
CTL
i
$R^2$ = 0.0023, $P$ = 0.75
Age
10
30
50
$R^2$ = 0.0088, $P$ = 0.53
Sex
M
F
$R^2$ = 0.021, $P$ = 0.33
RIN
2
4
6
8
$R^2$ = 0.037, $P$ = 0.19
Sequencing Batch
1
2
$R^2$ = 0.0015, $P$ = 0.8
Brain Bank
UMB
ATP
$R^2$ = 0.00019, $P$ = 0.93
Aligned Reads
1e+07
3e+07
5e+07
7e+07
$R^2$ = 0.024, $P$ = 0.29
5' to 3' bias
0.0
0.5
1.0
1.5
ASD
CTL

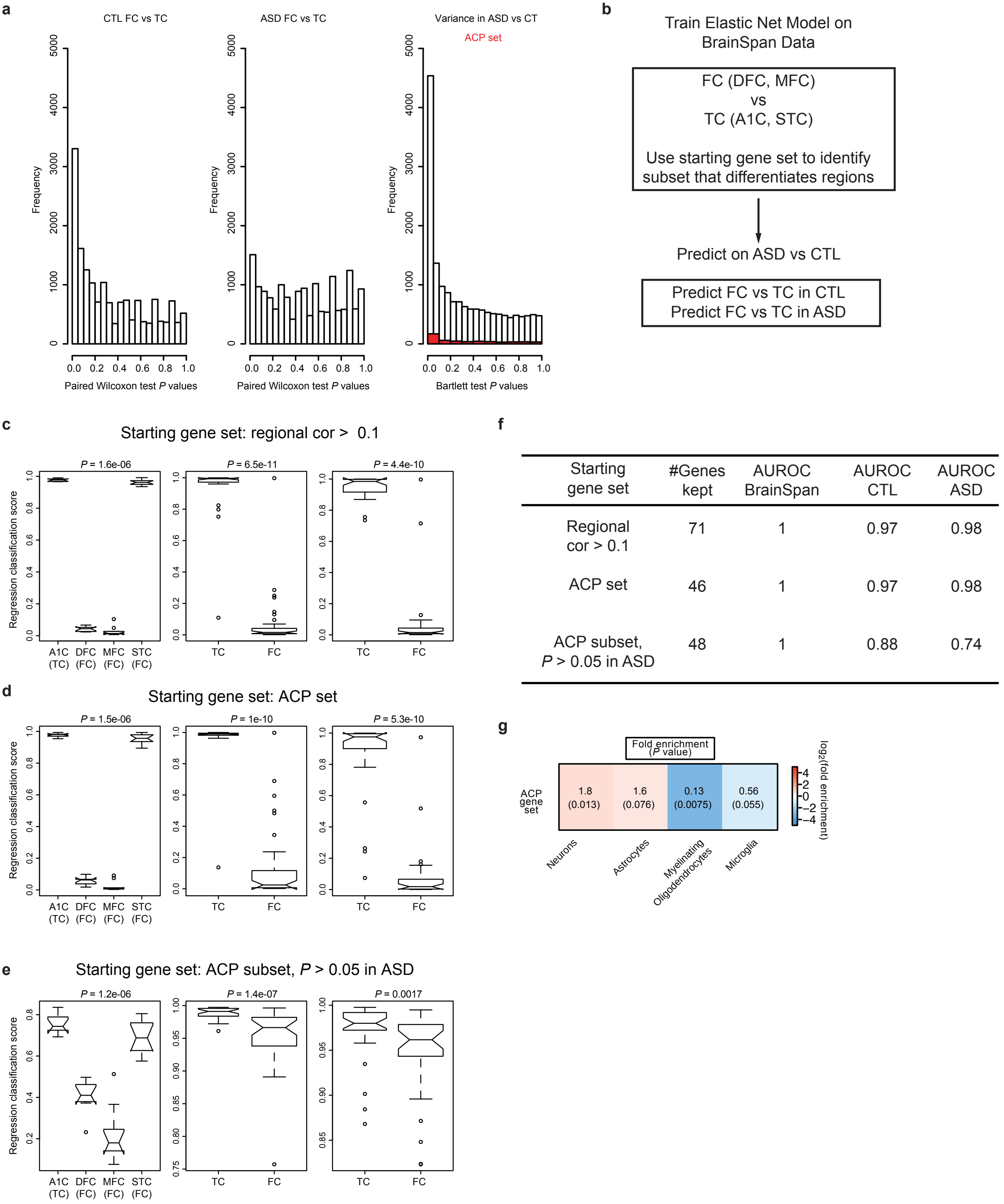

a
ρ = 0.90
P < 2.2x10^-16
−log10(P, LME with 5 seqSVs)
−log10(LME P value)
ρ = 1.0
P < 2.2x10^-16
−log10(LME P, permuted)
−log10(LME P value)
ρ = 0.76
P < 2.2x10^-16
−log10(LME P, limma voom)
−log10(LME P value)
ρ = 0.55
P < 2.2x10^-16
−log10(LME P, full model + 17 SVs)
−log10(LME P value)
b
Diagnosis
ASD
CTL
Replication Set
Y
Region
ba9
ba41−42−22
Age
67
2
Sex
F
M
RIN
9.1
1.8
Depth
186M
21M
5'-3' bias
0.7
0.5
Seizures
No
Yes
Psych. Meds
No
Yes
Overlap with Voineagu et al., 2011
N
Y
Expression Z score
6
4
2
0
−2
−4
−6
c
Fold enrichment
(P value)
↓ in ASD
2.5
(9.3e−06)
0.26
(0.0016)
2.3
(0.0024)
0.21
(1.7e−05)
↑ in ASD
0.65
(0.23)
4
(1.4e−13)
0.45
(0.15)
4.4
(3.4e−22)
Neurons
Astrocytes
Myelinating Oligodendrocytes
Microglia
log2(fold enrichment)
4
2
0
−2
−4
d
P = 0.029
Relative LINC00693 expression
(normalized to GAPDH)
UMB5302
UMB5297
AN00764
UMB5278
AN19760
UMB5168
UMB5163
UMB1376
ASD
CTL
P = 0.029
Relative LINC00689 expression
(normalized to GAPDH)
UMB5302
UMB5297
UMB5278
AN00764
AN19760
UMB5163
UMB5168
UMB1376
ASD
CTL



Scale
chr18:
100 kb
hg19
74,250,000
74,300,000
74,350,000
74,400,000
74,450,000
74,500,000
ZNF516
LINC00683
RP11-162A12.2
ZNF236
Expression in Human Control
Expression in Human ASD
ln(x+1) 0.2
0
100 kb
rheMac2
69,950,000
70,000,000
70,050,000
70,100,000
70,150,000
70,200,000
70,250,000
Non-Rhesus RefSeq Genes
Homo ZNF516
Homo LINC00683
Homo ZNF236
Expression in Macaque
Syntenic region in macaque
100 kb
mm9
82,900,000
82,950,000
83,000,000
83,050,000
Mouse RefSeq Genes
Zfp236
Zfp516
Expression in Mouse
Syntenic region in mouse
Scale
chr9:
10 kb
hg19
138,355,000
138,360,000
138,365,000
138,370,000
PPP1R26-AS1
PPP1R26
PPP1R26-AS1
PPP1R26-AS1
PPP1R26
ln(x+1) 0.5
Expression in Human Control
Expression in Human ASD
chr15:
5 kb
rheMac2
2,809,000
2,811,000
2,813,000
2,815,000
2,817,000
Non-Rhesus RefSeq Genes
Homo PPP1R26
Expression in Macaque
Syntenic region in macaque
chr2:
10 kb
mm9
28,280,000
28,285,000
28,290,000
28,295,000
28,300,000
RefSeq Genes
Gm347
Ppp1r26
Expression in Mouse
Syntenic region in mouse
Scale
chr10:
200 kb
hg19
6,600,000
6,700,000
6,800,000
6,900,000
7,000,000
7,100,000
7,200,000
PRKCQ
PRKCQ
PRKCQ
PRKCQ-AS1
LINC00707
SFMBT2
SFMBT2
Expression in Human Control
Expression in Human ASD
chr9:
200 kb
rheMac2
6,850,000
6,950,000
7,050,000
7,150,000
7,250,000
7,350,000
Homo PRKCQ
Homo LINC00707
Homo PRKCQ-AS1
Homo SFMBT2
Homo SFMBT2
Non-Rhesus RefSeq Genes
Expression in Macaque
Syntenic region in macaque
chr2:
200 kb
mm9
10,550,000
10,650,000
10,750,000
10,850,000
10,950,000
11,050,000
Sfmbt2
Sfmbt2
RefSeq Genes
Prkcq
Expression in Mouse
Syntenic region in mouse
Scale
chr3:
100 kb
hg19
67,750,000
67,800,000
67,850,000
67,900,000
67,950,000
68,000,000
68,050,000
SUCLG2
SUCLG2
SUCLG2-AS1
SUCLG2-AS1
FAM19A1
ln(x+1) 0.3
Expression in Human Control
Expression in Human ASD
chr2:
100 kb
rheMac2
68,200,000
68,300,000
68,400,000
68,500,000
Non-Rhesus RefSeq Genes
Homo FAM19A1
Homo SUCLG2-AS1
Homo SUCLG2-AS1
Homo SUCLG2
Expression in Macaque
Syntenic region in macaque
chr6:
100 kb
mm9
95,700,000
95,800,000
95,900,000
96,000,000
Suclg2
Suclg2
RefSeq Genes
Fam19a1
Fam19a1
Expression in Mouse
Syntenic region in mouse
Scale
chr5:
100 kb
hg19
17,300,000
17,350,000
17,400,000
17,450,000
17,500,000
BASP1
BASP1
FTH1P10
CTD-2139B15.2
RP11-321E2.3
RP11-321E2.4
ln(x+1) 0.9
Expression in Human Control
Expression in Human ASD
chr6:
20 kb
rheMac2
17,485,000
17,495,000
17,505,000
17,515,000
17,525,000
Homo BASP1
Non-Rhesus RefSeq Genes
Expression in Macaque
Syntenic region in macaque
chr15:
50 kb
mm9
25,350,000
25,400,000
25,450,000
Basp1
Gm5468
RefSeq Genes
Expression in Mouse
Syntenic region in mouse
Scale
chr15:
200 kb
hg19
69,800,000
69,900,000
70,000,000
70,100,000
70,200,000
70,300,000
KIF23
RP11-279F6.1
RPLP1
RefSeq Genes
DRAIC
PCAT29
RP11-279F6.2
RP11-279F6.3
LINC00593
TLE3
ln(x+1) 0.3
Expression in Human Control
Expression in Human ASD
chr7:
200 kb
rheMac2
48,200,000
48,300,000
48,400,000
48,500,000
48,600,000
Homo KIF23
Homo RPLP1
Homo PCAT29
Homo DRAIC
Homo LINC00593
Homo TLE3
Non-Rhesus RefSeq Genes
Expression in Macaque
Syntenic region in macaque
chr9:
200 kb
mm9
61,350,000
61,450,000
61,550,000
61,650,000
61,750,000
Tle3
RefSeq Genes
Rplp1
Kif23
Expression in Mouse
Syntenic region in mouse
Scale
chr5:
100 kb
hg19
127,050,000
127,150,000
127,250,000
127,350,000
CTXN3
CTXN3
CTC-228N24.1
CTC-228N24.2
CTC-228N24.3
SLC12A2
ln(x+1) 0.5
Expression in Human Control
Expression in Human ASD
chr6:
100 kb
rheMac2
124,100,000
124,200,000
124,300,000
124,400,000
Homo LINC01184
Homo CTXN3
Homo CTXN3
Non-Rhesus RefSeq Genes
Homo SLC12A2
Expression in Macaque
Syntenic region in macaque
chr2:
1 kb
mm9
124,974,000
124,975,000
124,976,000
124,977,000
124,978,000
Ctxn2
RefSeq Genes
Slc12a1
Slc12a1
Expression in Mouse
Syntenic region in mouse
Scale
chr11:
10 kb
hg19
65,185,000
65,195,000
65,205,000
65,215,000
FRMD8
FRMD8
NEAT1
NEAT1
MIR612
ln(x+1) 2
Expression in Human Control
Expression in Human ASD
chr14:
10 kb
rheMac2
9,005,000
9,010,000
9,015,000
9,020,000
Homo NEAT1
Homo NEAT1
Non-Rhesus RefSeq Genes
Homo FRMD8
Homo FRMD8
Expression in Macaque
Syntenic region in macaque
chr19:
5 kb
mm9
5,843,000
5,845,000
5,847,000
5,849,000
5,851,000
Neat1
Frmd8
Expression in Mouse
Syntenic region in mouse


a
Upstream Junction Count (UJC)
Downstream Junction Count (DJC)
Splice Junction Count (SJC)
PSI (ψ) = (UJC + DJC)/2 / ((UJC + DJC)/2 + SJC)
b
P value frequency
All events
Skipped Exon
Alternative 5' start site
Alternative 3' start site
Mutually exclusive exons
0.0 0.4 0.8
c
$R^2$ = 0.58
$P < 2.2 \times 10^{-16}$
Olego/Quantas ΔPSI
TopHat2/MATS ΔPSI
-0.2 0 0.2
d
$R^2$ = 0.69
$P$ = 0.0052
Slope = 1.2
PCR ΔPSI
RNA-seq ΔPSI
-15 10
ASTN2
MEF2D
ERC2
NCAM
MED31
SMARCC2
SYNE1
NRCAM
GRIN1
e
Control
ASD
bp
1 2 3 4 5 6 7 8
ASTN2
GAPDH
GRIN1
SMARCC2
SYNE1
MEF2D
ERC2
MED31
NRCAM
NCAM1
RPKM ($\log_{10}$)
Control1
Control2
Control3
Control4
ASD1
ASD2
ASD3
ASD4
119858324 119898905 119938357 119976831
Genomic coordinate (9), "-" strand
140040179 140040331 140042644 140043559
Genomic coordinate (9), "+" strand
56566370 56566726 56567499 56567633
Genomic coordinate (12), "-" strand
152832147 152834208 152838907 152841676
Genomic coordinate (6), "-" strand
156453054 156459546 156469258 156470569
Genomic coordinate (1), "-" strand
55544805 55596504 55655740 55717886
Genomic coordinate (3), "-" strand
6547616 6547868 6551578 6553751
Genomic coordinate (17), "-" strand
107800807 107805811 107813962 107817050
Genomic coordinate (7), "-" strand
113142483 113143716 113146066 113146284
Genomic coordinate (11), "+" strand


a
Diagnosis
ASD
CTL
Replication Set
Y
Region
ba9
ba41-42-22
Age
67
2
Sex
F
M
RIN
9.1
1.8
Depth
186M
21M
5'-3' bias
0.7
0.5
Seizures
No
Yes
Psych. Meds
No
Yes
Overlap with Voineagu et al., 2011
N
Y
6
4
2
0
−2
−4
−6
Expression Z score
b
P = 4.8x10$^{-13}$
PC1 DS
(1127 events, FDR <0.5)
P = 0.07
PC1 DGE of 833
spliced genes
P = 5.2x10$^{-13}$
PC1 DS
(569 events on genes
with DGE > 0.5)
P = 0.88
PC1 DGE of 455
spliced genes
ASD
CTL
c
P value
frequency
All events
Skipped Exon
Alternative 5'
start site
Alternative 3'
start site
Mutually exclusive
exons
d
Fold enrichment
(P value)
All events
Inclusion
in ASD
Exclusion
in ASD
2.9
(2.8e−05)
1.4
(0.48)
0.33
(0.37)
3.2
(0.14)
4.1
(1.8e−07)
0.65
(1)
Neurons
Astrocyte
2
(0.14)
1.1
(0.8)
1
(1)
1.5
(0.39)
2.4
(0.072)
0.97
(1)
Oligodendrocytes
Microglia
log2(fold enrichment)
e
Splicing factor clustering across samples by gene expression
1-(Pearson's R)
MBNL3
PTBP1
NOVA2
RBFOX3
SRRM4
NOVA1
MBNL1
MBNL2
PTBP2
RBFOX1
RBFOX2



Train Elastic Net Model on BrainSpan Data

FC (DFC, MFC)
vs
TC (A1C, STC)

Use starting gene set to identify subset that differentiates regions

Predict on ASD vs CTL

Predict FC vs TC in CTL
Predict FC vs TC in ASD

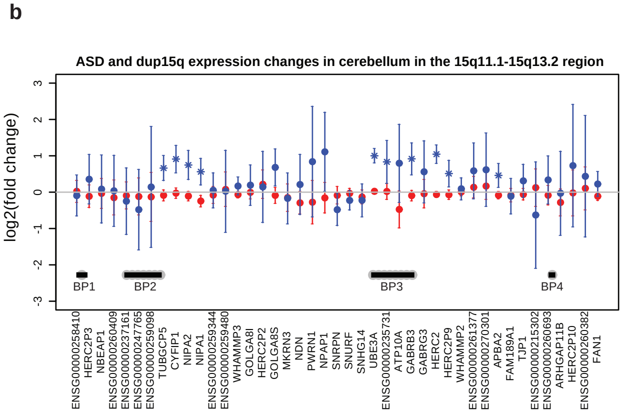
| Starting gene set | #Genes kept | AUROC BrainSpan | AUROC CTL | AUROC ASD |
|---|---|---|---|---|
| Regional cor > 0.1 | 71 | 1 | 0.97 | 0.98 |
| ACP set | 46 | 1 | 0.97 | 0.98 |
| ACP subset, $P > 0.05$ in ASD | 48 | 1 | 0.88 | 0.74 |

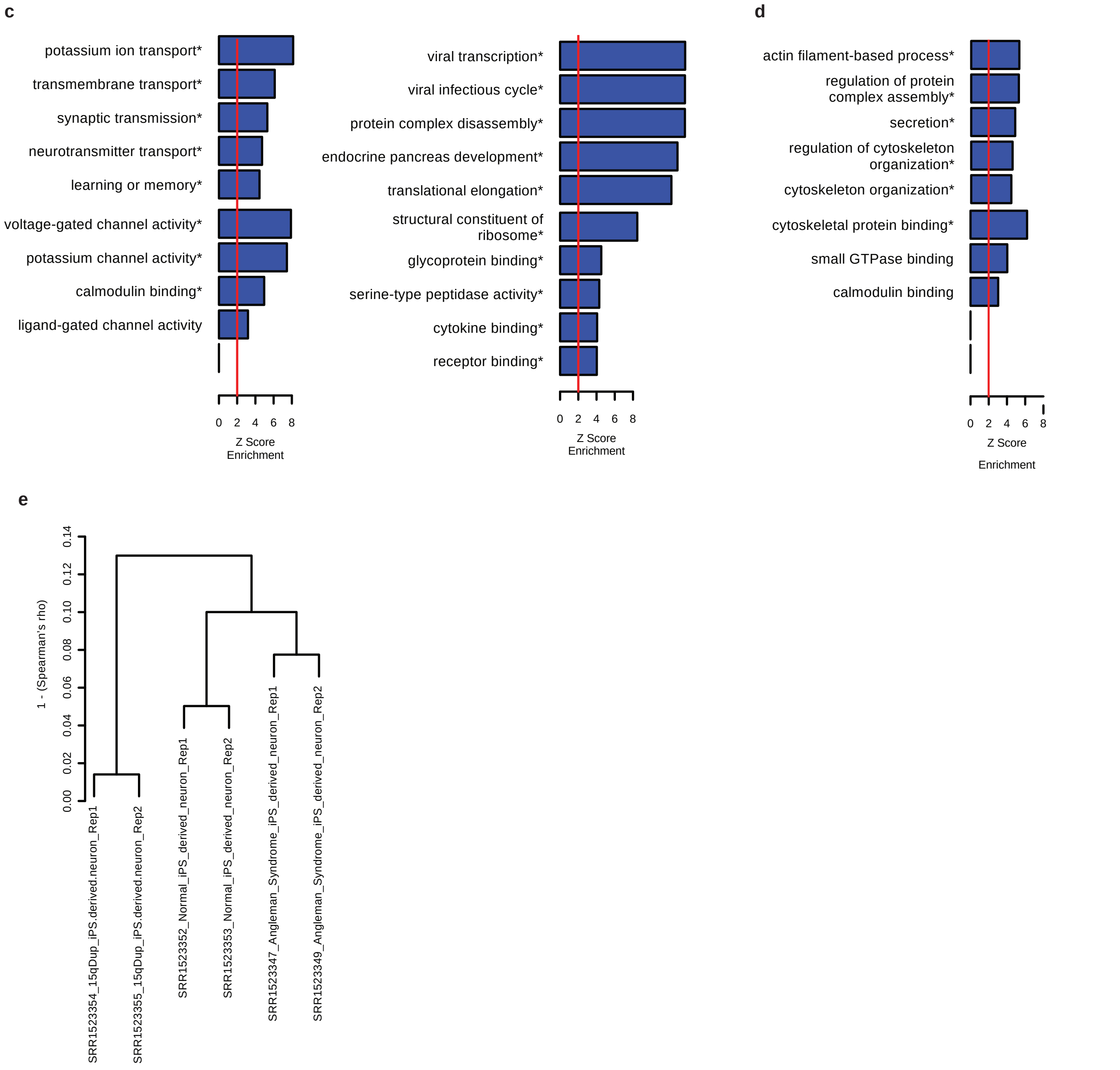
a

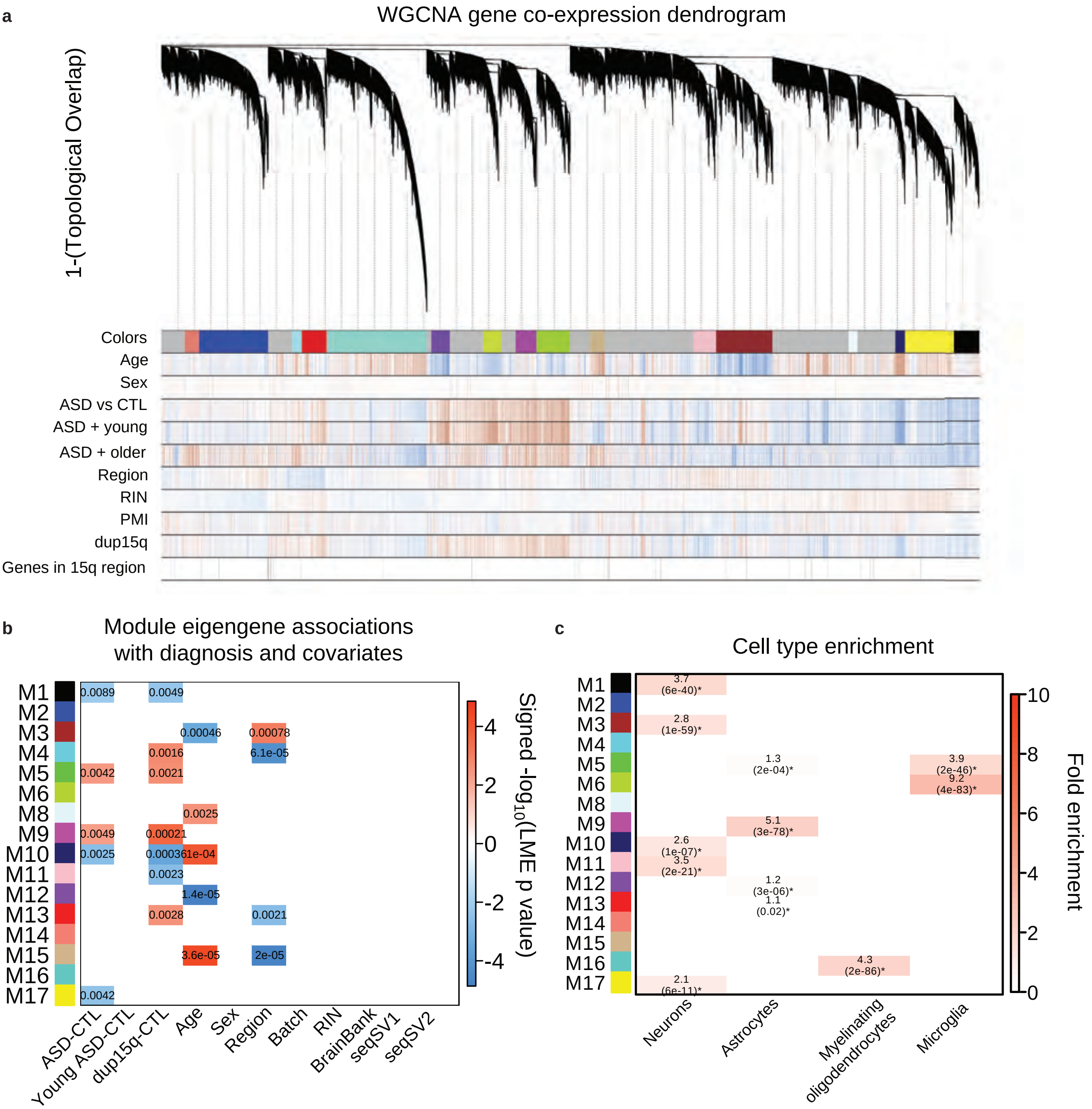
Duplication 15q breakpoints across individuals

| Sample | BP1-2 | BP2-3 | BP3-4 | BP4-5 |
|---|---|---|---|---|
| AN09402 | 4 | 4,b | 2 | 2 |
| AN14829 | 4 | 4 | 4 | 3 |
| AN17138 | 4 | 4 | 2 | 2 |
| AN03935 | 4 | 4 | 4 | 3 |
| AN05983 | 4 | 4 | 4 | 3 |
| AN06365 | 4 | 4 | 4 | 3 |
| AN11931 | 4 | 4 | 4 | 3 |
| AN14762 | - | 4,a | - | - |

a,Obtained from Scoles et al., 2011 who evaluated duplication in this region by RT-PCR of SNRPN/GABRB3/UBE3A vs B2M
b,Discrepancy with Scoles et al., who report 5 here






a
WGCNA gene co-expression dendrogram
1-(Topological Overlap)
Colors
Age
Sex
ASD vs CTL
ASD + young
ASD + older
Region
RIN
PMI
dup15q
Genes in 15q region
b
Module eigengene associations with diagnosis and covariates
M1
M2
M3
M4
M5
M6
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
0.0089
0.0049
0.00046
0.00078
0.0016
6.1e-05
0.0042
0.0021
0.0025
0.0049
0.00021
0.0025
0.00036
1e-04
0.0023
1.4e-05
0.0028
0.0021
3.6e-05
2e-05
0.0042
ASD-CTL
Young ASD-CTL
dup15q-CTL
Age
Sex
Region
Batch
RIN
BrainBank
seqSV1
seqSV2
Signed -log$_{10}$(LME p value)
4
2
0
-2
-4
c
Cell type enrichment
3.7 (6e-40)*
2.8 (1e-59)*
1.3 (2e-04)*
3.9 (2e-46)*
9.2 (4e-83)*
5.1 (3e-78)*
2.6 (1e-07)*
3.5 (2e-21)*
1.2 (3e-06)*
1.1 (0.02)*
4.3 (2e-86)*
2.1 (6e-11)*
Neurons
Astrocytes
Myelinating oligodendrocytes
Microglia
Fold enrichment
10
8
6
4
2
0

## Title

Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development

## Authors and affiliations

Hyejung Won[1], Luis de la Torre-Ubieta[1], Jason L. Stein[1], Neelroop N. Parikshak[1], Farhad Hormozdiari[3], Changhoon Lee[1], Eleazar Eskin[3,4], Jason Ernst[2,4], Daniel H. Geschwind[1,4*]

[1] Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles

[2] Department of Biological Chemistry, David Geffen School of Medicine, University of California Los Angeles

[3] Department of Computer Science, University of California Los Angeles

[4] Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles

[5]Department of Molecular, Cell and Developmental Biology, University of California Los Angeles, Los Angeles

[*] Correspondence: dhg@mednet.ucla.edu

## Introduction

The demonstration that chromatin exhibits a complex 3 dimensional organization, whereby short and long distance physical interactions correspond to complex gene regulatory processes has opened a new window on understanding the functional organization of the human genome[1-4]. Recently, chromatin remodeling has also been causally implicated in several neurodevelopmental disorders, including autism and schizophrenia[5-7]. However, it remains unclear whether knowledge of chromosome organization in a tissue specific manner might inform our understanding of gene regulation in brain development or disease. Here we determined the genome-wide landscape of chromosome conformation during early human cortical development by performing Hi-C analysis in the mitotically active and post mitotic laminae of human fetal brain. We integrate Hi-C data with transcriptomic and epigenomic data and utilize chromosome contact information to delineate physical gene-gene regulatory interactions for non-coding regulatory elements. We show how these data permit large-scale functional annotation of non-coding variants identified in schizophrenia GWAS and of human specific enhancers[8,9]. These data provide a rubric that illustrates the power of tissue-specific annotation of non-coding regulatory elements, as well as novel insights into the pathogenic mechanisms of neurodevelopmental disorders and the evolution of higher cognition.

Recent advances in high-throughput sequencing have unveiled the epigenomic landscape of multiple human cell types, as well as 3 dimensional folding principles of chromatin[10,11]. In particular, chromosome conformation capture experiments demonstrate that chromatin is organized into hierarchical structures, which include compartments (a few megabase (Mb))[1], topological associating domains (TADs, sub-Mb)[12], and loops (ranging from few kilobase (kb) to few hundred kb)[2,4]. These structures are thought to play a role in gene regulation and biological function by defining functional genomic units and mediating the effects of *cis*-regulatory elements via both short- and long-range physical interactions (e.g. promotor-enhancer interactions), relationships that cannot simply be predicted by linear adjacency in chromosomes. Coupled with epigenomic data, such higher order chromatin interactions should facilitate systemic annotation of *cis*-regulatory elements, as well as intergenic and intronic variants, which will further expand our understanding of tissue specific developmental programs, as well as disease pathogenesis.

We constructed multiple Hi-C libraries in mid gestation fetal cerebral cortex from three individuals during the peak of neurogenesis and migration (gestation week, GW17-18). We reasoned that it would be useful to analyze mitotically active neuronal precursors involved in neurogenesis separately from post-mitotic migrating and maturing neurons, so we dissected the cortical anlage into two major structures: the cortical and subcortical plate (CP), consisting primarily of post mitotic neurons and the germinal zone (GZ), containing primarily mitotically active neural progenitors (representative heatmap in **Fig. 1a**, **Extended Data Fig. 1a-b**). For comparison with non-neuronal cell types, we also used publicly available Hi-C data on human embryonic stem (ES) cells and IMR90 cells[11,12]. To provide grounding for our data and compare global chromosome architecture between different cell types, we performed principal component analysis (PCA)[13] on the genome-wide inter-chromosomal contact matrices of CP, GZ, ES, and IMR90. As previously demonstrated, global chromosome architecture does not change dramatically between different cell types[13]. However, the first principal components (PC1s) from neuronal tissues (CP and GZ) have significantly higher correlation than the PC1s between different cell types (**Fig. 1b**), consistent with the higher similarity between tissues from brain, versus the two other cell lines.

**3D chromatin structure reflects gene regulation during neural differentiation.**

Previous studies have shown that genome-wide chromosome conformation captures multiple levels of genomic features related to biological function, ranging from GC content and gene number to marks of open chromatin, such as DNase I hypersensitivity sites (DHS)[13]. Most human-relevant Hi-C has been conducted in cell lines[1,2,4,11,12,14] and not in complex tissue, such as developing brain. As an initial first step to insure the quality and validity of our data, we analyzed the relationships between the major component of the inter-chromosomal interaction matrix with these major genomic features, finding high correlation with GC content, gene number, DHS[10], and to a lesser extent, gene expression[15] (**Fig. 1c**, **Extended Data Fig. 2a**), as has been previously observed in non-neural cell lines[13].

To further explore the biological significance of chromosome contact changes during neural differentiation, we explored whether the genes in regions of dynamic chromatin structure were related to neural differentiation by comparing the inter-chromosome contact matrices (binned to 100kb) in different cell types and selecting bins with the highest chromatin contact count changes between two cell types (**Methods**). Genes located in the regions of highest inter-chromosomal interaction changes between CP and GZ were enriched for neuronal genes, represented by the gene ontology (GO) categories of neuron recognition, axon guidance, central nervous system (CNS)

development, and synapse (**Fig. 1d**, **Extended Data Fig. 2b**; **Methods**). Genes located in regions with highest inter-chromosomal interaction changes between CP and ES cells were enriched for developmental genes involved in forebrain development and chromatin organization (**Fig. 1d**, **Extended Data Fig. 2b**), indicating that these interactions reflect tissue relevant developmental gene regulation.

To further explore how these physical chromatin interactions relate to biological function, we hypothesized that highly interacting chromatin regions would be more likely to be co-regulated. To test this, we compared the distribution of correlation patterns for genes locating in (1) the regions of highest interaction values in both CP and GZ, (2) the lowest interacting regions in both CP and GZ, and (3) the regions of differential interaction values (the regions of highest interaction values in CP and lowest interaction values in GZ and vice versa). Highly interacting regions tend to be biased toward positive correlations, while there was no bias in correlation for low and differential interacting regions (**Fig. 1e**). Interestingly, the positive correlation for high interacting regions becomes even higher when more stringent cutoffs are used, supportive of the quantitative nature of interaction-driven co-expression, whereby the relationship between physical 3D chromatin interactions and expression is mostly driven by the top percentiles of interacting regions (**Extended Data Fig. 2c**). To further elucidate the epigenetic regulatory mechanisms behind the apparent interaction-mediated co-expression, we marked bins in which epigenetic marks from two loci appear together. By comparing the epigenetic mark combination matrix with the Hi-C contact matrix, we observed that interacting regions exhibit shared epigenetic patterns at the level of both inter- and intra-chromosomal interactions (**Fig. 1f**, **Extended Data Fig. 3**; **Methods**). In particular, regions associated with positive transcriptional regulation and enhancers are more likely to physically interact with each other, consistent with their co-regulation.

One of the core functional units of general genome organization recently uncovered by chromatin capture methods across a wide variety of cell types is the compartment, a relatively large, dynamic domain[1], which is comprised of smaller, sub-Mb regions of topologically associating domains (TADs)[12]. Compartments are divided into two types, type A compartments that consist primarily of euchromatin and actively transcribed genes and type B compartments, which are heterochromatic and repressed. TADs have been previously shown to be relatively stable, whereas compartments have been shown to change during lineage specification in stem cells[11]. Consistent with this, we observed dynamic compartment switching between CP and GZ, enriched for GO categories related to neuronal genes and phosphatase activity (**Fig. 2c**), as well as compartment switching between CP and ES (**Fig. 2a,d**). Genes that change compartments from ES to CP are decreased for A to B transitions across differentiation and increased for changes from the B to A compartments (**Fig. 2b**), as expected. Compartment changes are also accompanied by epigenetic changes, so that the B to A compartment shift is associated with increased DHS and active epigenetic marks indicative of open chromatin, whereas the A to B shift is associated with decreased DHS and increased repressive marks (**Fig. 2b,e**). The same pattern was observed for GZ vs. ES and CP vs. GZ (**Fig. 2b,e**, **Extended Data Fig. 2d**), demonstrating that gene expression changes across development are tightly linked to epigenetic changes coupled with compartment switching.

TADs are thought to mediate co-transcriptional regulation primarily within their boundaries (100kb-1Mb) through physical “looping” interactions of promotors and enhancers in co-regulated genes[4,16]. Since TAD boundaries are conserved across different cell types[12], we hypothesized that changes in epigenetic marks in TADs, rather than the boundaries of TADs, would be most associated with gene expression changes

across development. To test this, we divided genes based on their fold change in expression between ES and differentiated neurons[17] (both increased and decreased), and assessed changes in epigenetic marks within the TADs where these genes reside (**Extended Data Fig. 1c-e**, **Methods**). Notably, active marks including enhancers and elements related to transcribed regions are increased in TADs that contain upregulated genes, whereas repressive marks are increased in TADs that contain downregulated genes (**Fig. 2f**). Collectively, these results indicate that our Hi-C data reflects the major elements of global chromosome architecture in fetal brains, providing a framework for exploring gene regulatory mechanism related to human neural development and function.

Next, to demonstrate how knowledge of intra-chromosomal contacts could significantly advance understanding of important gene regulatory relationships in the nervous system, we performed two integrative experiments. In the first, we used these chromatin contact data to functionally annotate specific non-coding regulatory elements in the developing brain. We leveraged recent efforts that have identified >2000 developmental enhancers gained specifically in the human cerebral cortex, providing a remarkable resource for understanding the evolution of human cognition[8]. Usually, in the absence of such tissue specific data, regulatory elements are assigned to the closest gene[18,19], a convention that we compared with our Hi-C derived interactions. We reasoned that our Hi-C data from fetal brain could be used to identify the target genes for many of these enhancers, which based on previously chromatin looping analyses in cell lines are often not the closest gene[4,16,18,19].

We derived an interaction map of human-gained enhancers, defined as significant interacting regions (at a 1% false discovery rate, FDR) compared to the null distribution generated by fitting the contact frequencies of all fetal brain enhancers identified in the same study[8] (**Extended Data Fig. 4a**, **Methods**). We defined the search space as including the 1Mb flanking regions, since most enhancer-promoter interactions are within this range[4]. Although statistically significant interactions are increased upon proximity to the enhancer, the majority of interactions are at relatively long-ranges (>100kb, **Extended Data Fig. 4b**) and are not restricted to the adjacent genes. Indeed, ~65% of the closest genes to human-gained enhancers are not identified through fetal brain Hi-C interactions, revealing that the majority of enhancers are not interacting with the most adjacent gene (**Fig. 3c**). Compared to the original study[8], which used human-gained enhancer hotspot TADs in ES cells and IMR90 cells due to the lack of Hi-C data from relevant tissue, our approach provides genes of action with higher resolution in the matching tissue (fetal cortices) from which evolutionary enhancers were identified. Human-gained enhancer-interacting regions were enriched with enhancers, promoters, and transcription start sites (TSSs) (**Fig. 3a**, **Extended Data Fig. 4c**), consistent with the previous findings that enhancers interact with promoters, as well as other enhancers[16]. The majority of interactions (>75%) were in the same TADs (**Fig. 3b**), also consistent with observations in cell lines that most enhancer-promoter interactions are in the same TAD[16,19]. Human-gained enhancer interacting genes (Hi-$C_{evol}$ genes) are involved in GTPase regulation as well as G-protein coupled receptor (GPCR) and CREB signaling, and are enriched with GO terms representing synaptic and axon guidance genes (**Fig. 3e**, representative interactions in **Fig. 3d**). One striking example is a human-gained enhancer that interacts with *ARHGAP11B*, a human-specific gene implicated in the expansion of human neocortex[20] (**Fig. 3d**).

Given the high conservation of protein-coding genes across the vertebrate lineage, comparative genomics have suggested that human-specific traits most likely result from changes in regulatory elements[8,21]. Indeed, protein-coding Hi-$C_{evol}$ genes have a lower

non-synonymous substitution (dN)/synonymous substitution (dS) ratio compared to Hi-C non-interacting protein-coding genes in multiple lineages (**Extended Data Fig. 5**). These results indicate that human-gained enhancers are interacting with protein-coding genes that undergo purifying selection, further supporting the hypothesis that non-coding elements undergo evolutionary selection to induce species-specific changes in gene expression[8,21]. We also investigated whether human-gained enhancers are interacting with lineage-specific long non-coding RNAs (lncRNAs)[22]. We observed that lineage-specific interactions with human-gained enhancers were enriched for primate-specific lncRNAs, as well as evolutionary conserved lncRNAs (**Fig. 3f**, **Extended Data Fig. 5**). Thus, while human-gained enhancers interact and possibly regulate evolutionary conserved protein-coding genes, they are more likely to interact with primate-specific lncRNAs.

Since the development of human higher cognition is dependent on the development of the human cerebral cortex via elaboration of novel gene regulatory relationships[8,23], we reasoned, as have others[8] that the genes regulated by these human specific enhancers would be associated with intellectual functioning in humans. Remarkably, we found that the Hi-$C_{evol}$ genes in fetal brain, but not the genes defined by proximity to the enhancers are significantly enriched with intellectual disability (ID) risk genes[6] (**Fig. 3g**). This result provides experimental support for the contention that human-gained enhancers are associated with the evolution of human cognitive function[8]. This enrichment was tissue-specific, as Hi-$C_{evol}$ genes defined by Hi-C interactions in ES cells did not show enrichment for ID risk genes (**Fig. 3g**). Indeed, ~56% of the Hi-$C_{evol}$ genes in neuronal tissue were not identified through chromatin contacts in ES cells, emphasizing the importance of defining tissue-relevant chromatin contacts, as well as importance of using the relevant tissue for Hi-C analysis (**Fig. 4c**).

Since most disease related common genetic variation is located in non-protein coding regions, we next assessed the ability of Hi-C data for functional annotation of common single nucleotide polymorphisms (SNPs). As a first line verification that Hi-C data could identify known functional relationships between SNPs and gene expression we used *cis*-expression quantitative trait loci (eQTL) data from adult frontal cortex[24], since such data is not yet available from fetal brain. For each significant eQTL locus, we obtained a set of significant eQTL SNPs with >95% likelihood of containing the causal SNP from association statistics and linkage disequilibrium (LD; 1000 Genomes) structure using CAVIAR[25]. We then identified genes interacting to likely causal eQTL SNPs via the chromatin contact matrix (Hi-$C_{eQTL}$ genes, **Methods**), and compared Hi-$C_{eQTL}$ genes with the known associated gene from the eQTL study, finding that Hi-$C_{eQTL}$ genes significantly overlapped with eQTL transcripts (**Extended Data Fig. 6a**). There were many Hi-$C_{eQTL}$ genes that were not identified as eQTL transcripts, likely due to a combination of factors, including low power of the eQTL sample, limited resolution of Hi-C (SNP-transcript interactions within 20kb cannot be detected), and the difference in age of tissues used for each analysis. Indeed, eQTL SNPs identified by CAVIAR were highly enriched with adult frontal cortex, but not fetal brain, enhancers (**Extended Data Fig. 6b-d**). Despite this, eQTL SNP-transcript pairs exhibit higher chromatin contact frequency than expected by chance across all distance ranges (**Extended Data Fig. 6e**), further supporting the utility of Hi-C to infer the biological function of regulatory variation.

Next, we applied a similar logic to advance our understanding of 108 genome-wide significant schizophrenia-associated loci, most of which are in relatively uncharacterized non-coding regions of the genome[9]. We obtained credible SNPs using CAVIAR, and split SNPs into those without known function and likely functional SNPs (SNPs that cause missense, frameshift, and splice variants and SNPs that fall onto gene promoters;

Methods). Credible SNPs were enriched with enhancers in fetal brain and adult frontal cortex, confirming the likely regulatory role of these SNPs in the brain (**Extended Data Fig. 7**). SNPs defined as likely functional SNPs and promoter SNPs were directly assigned to their target genes. For the remaining intergenic and intronic SNPs that were un-annotated, and therefore without clear function, we used the chromatin contact matrix to find genes with which the regions where the SNPs are located are physically interacting (diagram in **Extended Data Fig. 7**).

Combining genes annotated as functional SNPs, promoter SNPs, and by Hi-C interactions, we obtained a total of ~900 genes (Hi-$C_{SCZ}$ genes) associated with schizophrenia risk variants. Hi-C contacts identified numerous genes that were neither adjacent to index SNPs nor in LD with them (**Fig. 4a-c**, **Extended Data Fig. 9**). While almost 70-80% of the LD genes and closest genes were identified as Hi-$C_{SCZ}$ genes, only half of them were identified by chromatin contacts, indicating that many of them were identified by functional SNPs residing in the genes. Moreover, 70-90% of the Hi-$C_{SCZ}$ genes were not identified by using LD genes or the closest genes to the association signal, consistent with observations that the linear organization of genes and regulatory elements on the chromosome does not reflect regulatory interactions[4,18,19].

Hi-C analysis showed that schizophrenia-associated common variants converge into specific molecular pathways related to neuronal function, including the postsynaptic density, acetylcholine receptors, cell cycle, and chromatin remodelers (**Fig. 4d-e**, **Extended Data Fig. 7-8**). To insure that this was not an artifact of the method used for credible SNP selection, we used a different method to define the set of credible SNPs[9] (**Extended Data Fig. 9**) and found the same enrichments, demonstrating the robustness of the genes identified through the Hi-C analysis. One notable example is illustrated by credible SNPs (rs4245150, rs17602038, rs4938021, rs4936275, rs4936276) that reside upstream of the *Dopamine D2 Receptor* (*DRD2*), the target of antipsychotic drugs. Although these SNPs are close to the *DRD2* TSS, they are not within the gene, which complicates interpretation of their biological function. Hi-C analysis demonstrates for the first time that indeed these SNPs are interacting with the TSS of *DRD2* (**Fig 4e**), providing biological insights into the function of these SNPs.

Another relevant example is an index SNP (rs79212538) interacting with *GRIA1*, an ionotropic glutamate receptor subunit, although *GRIA1* is neither the closest gene nor in LD with the index SNP (**Extended Data Fig. 8**). Additionally, Hi-C shows that schizophrenia associated non-coding SNPs interact with multiple genes involved in excitatory synaptic transmission, including *CACNA1C*, *GRIN2A*, and *NLGN4X*, further supporting glutamatergic transmission defects in schizophrenia pathophysiology (**Extended Data Fig. 8**). Interestingly, Hi-$C_{SCZ}$ genes significantly overlap with ASD *de novo* likely gene-disrupting (LGD) targets (CP: OR=2.4, P=$1.6x10^{-5}$, GZ: OR=1.8, P=0.006), consistent with a shared genetic etiology between ASD and schizophrenia[26]. The fact that genes with LGD mutations in ASD are associated with regulatory variants in schizophrenia suggests that complete abrogation of these genes may cause developmental defects as in ASD, while regulatory changes in these genes may cause later-onset of neuropsychiatric symptoms as in schizophrenia. Collectively, genes annotated by chromatin contact information provide novel insights into schizophrenia pathogenesis.

In conclusion, we demonstrate how a comprehensive analysis of genome-wide chromatin configuration during human neural development informs our view of gene regulation. This chromatin contact landscape provides important biological insights on gene regulatory mechanisms, such that co-expressed genes share epigenetic co-regulation of interacting regions, and that changes in functional epigenetic marks are tightly linked to TADs and compartment switching to induce changes in gene expression.

We also annotated non-coding regulatory elements in the genome based on long-range chromatin contacts to identify enhancer-promoter interactions during human brain development, as well as genes of actions for eQTL. In turn, we show how these interactions can be used to inform our biological interpretation of risk variants for schizophrenia, which serves as a template for understanding the role of non-coding variation more broadly in neuropsychiatric disorders.

## Methods

### Fetal brain layer dissection

Human fetal cortical tissues from three individuals were collected from frontoparietal cortex at gestation week (GW) 17-18 (one sample from GW17 and two samples from GW18). In cold DMEM/F-12 (ThermoFisher, 11320-033), frontoparietal cortex was first dissected to thin (~1mm) slices to visualize layers. Under the light field microscope, cortical slice was dissected to germinal zone (GZ) and cortical plates (CP). GZ contains ventricular zone and subventricular zone, and hence comprised of proliferating neurons. CP refers to intermediate zone, cortical plate, and marginal zone, which are mainly composed of differentiated and migrating neurons. By dissecting layers from same fetal cortices, we can compare progenitors to differentiated neurons with same genotype and minimize intersample heterogeneity.

### Hi-C

Collected tissue was dissociated with trypsin and cell number was counted. Ten million cells were fixed in 1% formaldehyde for 10 min. Cross-linked DNA was digested by restriction enzyme HindIII (NEB, R0104). Digested chromatin ends were filled and marked with biotin-14-dCTP (ThermoFisher, 19518-018). Resulting blunt-end fragments were ligated under dilute concentration to minimize random intermolecular ligations. DNA purified after crosslinking was reversed by proteinase K (NEB, P8107) treatment. Biotins from unligated ends were removed by exonuclease activity of T4 DNA polymerase (ThermoFisher, 18005). DNA was sheared by sonication (Covaris, M220) and 300-600bp fragments were selected. Biotin-tagged DNA, which is intermolecular ligation products, was pulled down with streptavidin beads (Invitrogen, 65001), and ligated with Illumina paired end adapters. Resulting Hi-C library was amplified by PCR (KAPA Biosystems HiFi HotStart PCR kit, KK2502) with the minimum number of cycle (typically 12-13 cycles), and sequenced by Illumina 50bp paired-end sequencing.

### Hi-C reads mapping and pre-processing

Note that mapping and filtering of the reads, as well as normalization of experimental and intrinsic biases of Hi-C contact matrices were conducted with the following method regardless of cell types to minimize potential variance in the data obtained from different platforms. We implemented *hiclib* (https://bitbucket.org/mirnylab/hiclib) to perform initial analysis on Hi-C data from mapping to filtering and bias correction. Briefly, quality analysis was performed using a phred score, and sequenced reads were mapped to hg19 human genome by *Bowtie2* (with increased stringency, *--score-min -L 0.6,0.2--very-sensitive*) through iterative mapping. Read pairs were then allocated to HindIII restriction enzyme fragments. Self-ligated and unligated fragments, fragments from repeated regions of the genome, PCR artifacts, and genome assembly errors were removed. Filtered reads were binned at 10kb, 40kb, and 100kb resolution to build a genome-wide contact matrix at a given bin size. This contact map depicts contact frequency between any two genomic loci. Biases can be introduced to contact matrices by experimental procedures and intrinsic properties of the genome. To decompose biases from the contact matrix and yield a true contact probability map, filtered bins were subjected to iterative correction[13], the basic assumption of which is that each locus has uniform coverage. Bias correction and normalization results in a corrected heatmap of bin-level resolution. 100kb resolution bins were assessed for inter-chromosomal interactions, 40kb for TAD analysis, and 10kb for gene loop detection.

### Inter-chromosomal principal component analysis

Principal component analysis (PCA) was conducted in a genome-wide inter-chromosome contact map (100kb binned) as described previously[13]. Since intra-

chromosome conformation may drive the PCA results, *cis* contacts were iteratively replaced to random *trans* counts. After removing diagonal and poorly covered regions, we performed PCA using *hiclib* command *doEig*.

Pearson's correlations between the first principal components (PC1) from different cell types (CP, GZ, ES, and IMR90[12]) were calculated to compare similarities in inter-chromosomal interactions between different cell types.

Spearman's correlations between PC1/PC2 and biological traits (GC content, gene density, DNase I hypersensitivity (DHS), gene expression) were calculated. GC content (%) for each 100kb bin was calculated by *gcContentCalc* command from R package *Repitools*. Gene density (number of genes in 100kb bin) was obtained based on longest isoforms from GENCODE19. DHS of fetal brains from Epigenomic roadmap[10] and gene expression level of prenatal cortical layers from Miller et al.[15] were used and average values per 100kb bin were calculated.

### Gene enrichment analysis

Gene ontology (GO) enrichment was performed by GO-Elite Pathway Analysis (http://www.genmapp.org/go_elite/). All genes in the genome except the ones located in the chromosome Y and mitochondrial DNA were used as a background gene list. Because Hi-C interaction is measured in bins, sometimes we cannot dissect the individual genes when they are clustered in the genome (i.e. PCDH locus). To prevent several gene clusters overriding entire GO terms, we removed GO mainly defined by gene clusters (for 100kb or 40kb binned data) or we randomly included one gene per cluster (e.g. PCDHA1 for PCDHA1-13 cluster) prior to GO analysis (for 10kb binned data).

Gene enrichment for the curated gene lists was performed using binomial generalized linear model to regress out exome length. Autism spectrum disorder (ASD) *de novo* gene list and intellectual disability (ID) curated gene list from Iossifov et al.[27] and Pariskshak et al.[6] were used for the enrichment test, respectively. Protein-coding genes based on biomaRt were used as a background gene list.

### Identification of the regions with largest inter-chromosomal conformation changes

Chromosome contact matrix was normalized with the total interaction counts between two cell types for comparison. Intra-chromosomal interactions were masked from the genome-wide contact matrix, and top 1000 bins with the largest interaction changes between different cell types (GZ vs. CP or ES vs. CP) were selected. As one bin is comprised of two loci that are interacting with each other, this would give ~2000 sites in the genome. Genes located in those ~2000 sites were combined to perform GO analysis.

### Co-expression of inter-chromosomal interacting regions

Using transcriptome from fetal cortical layers[28], average expression values per 100kb bin were calculated. Pearson correlation matrix was calculated from 100kb binned expression data from all layers to generate gene co-expression matrix. At this step, gene co-expression matrix has the same dimension as inter-chromosomal contact matrix.

We hypothesized that genes would be co-expressed across the layers when they are interacting in all stages (both in CP and GZ), so we selected top 2% highest interacting regions of fetal brains both at GZ and CP (high interacting regions). We also selected (1) low interacting regions: top lowest interacting regions (0 interaction from normalized Hi-C contact matrix) of fetal brains both at GZ and CP, as well as (2) variant interacting regions: top 2% highest interacting regions from one stage (e.g. GZ) that are top 2%

lowest interacting regions from the other stage (e.g. CP) for comparison. Expression correlation values of the same regions were selected from the gene co-expression matrix, and expression correlations between different states (high interacting regions vs. low interacting regions and high interacting regions vs. variant interacting regions) were compared by two-sample Kolmogorov-Smirnov test.

### Epigenetic state enrichment for inter-chromosomal interacting regions

The fetal brain epigenetic 25 state model from Epigenomic roadmap[10] was used to generate the epigenetic state combination matrix, which was generated by marking loci where two interacting chromosomal bins (defined as bins with (1) interaction counts > 75% quantile interaction count for inter-chromosome and (2) interaction counts > 0 for intra-chromosome) share epigenetic signature. For example, the epigenetic combination matrix between the active transcription start site (TssA) and active enhancers (EnhA1) was generated by marking where interacting loci have TssA on one locus and EnhA1 on the other locus. Intra- and inter-chromosomal contact frequency maps were then compared to epigenetic state matrix by Fisher's exact test to calculate enrichment of shared epigenetic combinations in interacting regions.

### Compartment analysis

Expected interaction frequency was calculated from the normalized intra-chromosomal 40kb binned contact matrix based on the distance between two bins. We summed series of submatrices of 400kb window size with 40kb step size from the normalized Hi-C maps to generate observed and expected matrices. The Pearson's correlation matrix was computed from the observed/expected matrix, and PCA was conducted on correlation matrix. PC1 from each chromosome was used to identify compartments. Eigenvalues positively correlated with the gene density were set as compartment A, while those that are negatively correlated were set as compartment B.

### Gene expression and epigenetic state change across different compartments

Genomic regions were classified into three categories according to compartments: compartment A in cell type1 that changes to compartment B in cell type2 (A to B), compartment B in cell type1 that changes to compartment B in cell type2 (B to A), regions that do not change compartment between two cell types (stable).

Genes residing in each compartment category were selected and GO enrichment was performed. Gene expression fold-change (FC) between different cell types was calculated from Miller et al.[15] (comparison for CP vs. GZ) and CORTECON[17] (comparison for ES vs. CP and ES vs. GZ). Distribution of gene expression FC for genes in different compartment categories was compared by one-way ANOVA and Tukey's posthoc test.

15 state epigenetic marks from Epigenomic Roadmap[10] in genomic regions classified based on compartments were averaged across 40kb bins. The DHS FC[10] between different cell types (ES vs. CP and ES vs. GZ) was calculated and statistically evaluated as in the gene expression comparison. Each epigenetic state counts[10] for one compartment category was normalized by total epigenetic mark number of that compartment category and compared between ES and fetal brains.

### TAD analysis

We conducted TAD-level analysis as described previously[12]. Shortly, we quantified the directionality index by calculating the degree of upstream or downstream (2Mb) interaction bias of a given bin, which was processed by a hidden Markov model (HMM) to remove hidden directionality bias.

Regions in between TADs are titled as TAD boundaries when the regions are smaller than 400kb and unorganized chromatin when the regions are larger than 400kb.

### TAD-based epigenetic changes upon differentially expressed genes

Genes were subdivided into 20 groups based on expression FC between ES and most differentiated neuronal states in CORTECON[17]: genes that are upregulated and downregulated upon differentiation were grouped into 10 quantiles, respectively, based on the FC. TADs into which genes from one subdivision reside were selected, and epigenetic state changes (from Epigenomic roadmap's 15 state epigenetic marks in ES and fetal brains[10]) in those TADs were normalized with TAD length and compared between ES and fetal brains. As different types of epigenetic marks have different absolute numbers (e.g. there are more quiescent states than enhancer states in the genome), each epigenetic state change was scaled across different quantiles to allow comparison between different states.

### Identification of Hi-C interacting regions

We identified Hi-C interacting regions and target genes for (1) human-gained enhancers[8], (2) expression quantitative trait loci (eQTL) SNPs[24], and (3) schizophrenia SNPs[9]. As the highest resolution available for the current Hi-C data was 10kb, we assigned these enhancers/SNPs to 10kb bins, obtained Hi-C interaction profile for 1Mb flanking region (1Mb upstream to 1Mb downstream) of each bin. We also made a background Hi-C interaction profile by pooling (1) 255,698 H3K27ac sites from frontal and occipital cortex at 12 PCW for human-gained enhancers[8] and (2) 9,444,230 imputed SNPs for eQTL and schizophrenia SNPs[9]. To avoid significant Hi-C interactions affecting the distribution fitting as well as parameter estimation, we used the lowest 95 percentiles of Hi-C contacts and removed zero contact values. Using these background Hi-C interaction profiles, we fit the distribution of Hi-C contacts at each distance for each chromosome using *fitdistrplus* package (**Extended Data Fig. 4a**). Significance for a given Hi-C contact was calculated as the probability of observing a stronger contact under the fitted Weibull distribution matched by chromosome and distance. P-values were adjusted by computing FDR, and Hi-C contacts with FDR<0.01 were selected as significant interactions. Significant Hi-C interacting regions were overlapped with GENCODE19 gene coordinates (including 2kb upstream to transcription start sites (TSS) to allow detection of enhancer-promoter interactions) to identify interacting genes. Same analysis was performed on Hi-C contact maps from CP, GZ, and ES[11]. To address the functional significance of target genes, GO enrichment was performed for the interacting genes.

### Protein-coding genes interacting with human-specific evolutionary enhancers

Protein-coding genes based on biomaRt (GENCODE19) were selected and non-synonymous substitution (dN)/synonymous substitution (dS) ratio was calculated for homologs in mouse, rhesus macaque, and chimpanzee for representation of mammals, primates, and great apes, respectively. Log2(dN/dS) distributions for protein-coding genes interacting vs. non-interacting to human-specific evolutionary enhancers in each lineage were then compared by two-sample Kolmogorov-Smirnov test.

### LncRNAs interacting with human-specific evolutionary enhancers

Long non-coding RNAs (lncRNAs) classified according to evolutionary lineages[22] were used to assess whether lineage-specific lncRNAs are interacting to human-specific evolutionary enhancers. We randomly selected the same number of enhancers (2,104) to the human-specific ones from the total enhancer pool (255,698), identified interacting regions based on the null distribution generated from a background enhancer interaction profile. Significant interacting regions (FDR<0.01) identified by Hi-C were intersected

with lncRNA coordinates[22] and interacting lncRNAs for each lineage were counted. This step was repeated for 3,000 times to obtain the lncRNA lineage distribution. LncRNAs interacting with human-specific evolutionary enhancers were also identified and enrichment was tested by calculating P-values as the probability of observing more interacting lncRNAs for a given lineage under the null lncRNA lineage distribution.

### Epigenetic state enrichment for Hi-C interacting regions

The functional framework for (1) eQTL SNPs, (2) schizophrenia SNPs, and (3) human-gained enhancers-interacting regions was assessed for epigenetic state enrichment. We implemented the same approach as in GREAT[29] to analyze the epigenetic state enrichment for *cis*-regulatory regions. For example, to evaluate whether schizophrenia SNPs are enriched with DHS, fraction of genome annotated with DHS (p), the number of schizophrenia SNPs (n), and number of schizophrenia SNPs overlapping with DHS (s) were calculated. Significance of the overlaps was tested by binomial probability of $P = Pr_{binom}\ (k \geq s \mid n = n,\ p = p)$[29]. Histone marks and 15-chromatin states from fetal brains, adult frontal cortex, and IMR90[10] were used for epigenetic state enrichment.

### eQTL analysis

To address whether co-localization mediates gene regulation, we compared the association between chromosome conformation with eQTL. Although fetal brain eQTL data would be optimal, since this data is currently not available, we analyzed adult frontal cortex *cis*-acting eQTL data[24]. We selected SNPs associated with gene expression (FDR<0.01) and clustered them with association $P<1\times10^{-5}$ and $r^2>0.6$ to obtain index SNPs. Using summary association statistics and linkage disequilibrium (LD) structure for each index SNP, we applied *CAVIAR*[25] to quantify the probability of each variant to be causal. Among 121,273,364 SNP-transcript pairs from frontal cortex eQTL data, this process resulted in 42,190 SNP-transcript pairs (267 transcripts and 14,882 SNPs) that are potentially credible. We refer to 14,882 credible SNPs as credible SNPs. Credible SNP interacting genes were identified as described in "identification of Hi-C interacting regions" section.

Fisher's exact test was performed to evaluate the significance of the overlap between Hi-C interacting genes and eQTL transcripts. The background gene list for Fisher's exact test includes genes located in 1Mb flanking regions to credible SNPs that are also tested in eQTL analysis.

For 42,190 SNP-transcript pairs, we assigned credible SNPs and genes into 10kb bins, and obtained Hi-C contacts between credible SNPs and genes from the 10kb binned Hi-C contact maps. As a gene can span across multiple 10kb bins, the highest interaction in the gene to a credible SNP was selected as Hi-C contacts as previously defined[30]. We also calculated expected interaction frequency from the normalized 10kb binned contact matrix based on the distance between two bins. Opposite interaction frequency was calculated by obtaining Hi-C contacts for the opposite site to the credible SNP with the same distance. Because interaction counts differ in different chromosomes as well as in different cell types, we normalized interaction by chromosomes and cell types. We performed one-way ANOVA and Tukey's posthoc test for the comparison between different interaction paradigms.

### Identification of credible SNPs for schizophrenia GWAS loci

128 LD-independent SNPs with genome-wide significance ($P<5\times10^{-8}$)[9] were used as index SNPs to obtain schizophrenia credible SNPs. All SNPs that are associated with $P<1\times10^{-5}$ and in LD ($r^2>0.6$) with an index SNP were selected, and correlations among this set of SNPs (LD structure) were calculated. CAVIAR was applied to summary association statistics and LD structure for each index SNP, and potentially causal SNPs

for each index SNP were identified. Among 55,000 SNPs that are in LD with 128 index SNPs, 7,613 SNPs were selected as causal by CAVIAR. Here we refer to these CAVIAR-identified SNPs as credible SNPs. Genes interacting to credible SNPs were identified as described in "identification of Hi-C interacting regions" section for CP, GZ, and ES. A separate set of credible SNPs initially reported from the original study was also processed with the same method[9].

### Identification of schizophrenia GWAS SNP-associated genes

We classified credible SNPs based on potential functionality (flow chart in **Extended Data Fig. 7**). For credible SNPs classified as functional (stop gained variant, frameshift variant, splice donor variant, NMD transcript variant, and missense variant) from biomaRt, we selected genes in which those SNPs locate. For those that are not directly affecting the gene function, we selected SNPs that fall onto the promoter and TSS of genes (2kb upstream-1kb downstream to TSS). Remaining SNPs were tested for Hi-C interaction so that Hi-C interacting genes were identified. This pipeline gives total ~900 genes potentially associated with GWAS SNPs.

### Identification of closest genes and LD genes

Closest genes to human-gained enhancers and schizophrenia index SNPs were obtained by *closestBed* command from *bedtools*. Gene coordinates from GENCODE19 including 2kb upstream to TSS were used to identify the closest genes.

LD genes refer to all genes in the LD. Here, LD is defined as physically distinct schizophrenia-associated 108 genome-wide significant regions[9]. We overlapped gene coordinates from GENCODE19 with LD regions to find genes that reside in LD.

Closest genes and LD genes were compared with Hi-C interacting genes. Venn diagrams were generated by *Vennerable* package in R. Only protein-coding genes were included in plotting Venn diagrams.

### Calculation of distance between SNPs and genes

For LD genes and closest genes, the shortest distance between an index SNP and a target gene was selected. For credible SNPs, (1) the distance between functional credible SNPs and target genes was set as 0, because functional SNPs reside in the gene, (2) the distance between promoter credible SNPs and target genes was calculated as the distance between SNPs and TSS of a gene, (3) the distance between credible SNPs and Hi-C interacting genes was calculated based on the distance between SNPs and Hi-C interacting bins (note that this distance has a unit of 10kb). We then combined the distance distributions from the 3 categories.

## Figure Legends

**Figure 1. Chromosome conformation in fetal brains reflects genomic features. a.** Representative heatmap of the chromosome contact matrix of CP. Normalized contact frequency (contact enrichment) is color-coded according to the legend on the right. **b.** Pearson correlation of the leading principle component (PC1) of inter-chromosomal contacts at 100kb resolution between *in vivo* cortical layers and non-neuronal cell types (ES and IMR90). **c.** Spearman correlation of PC1 of chromatin interaction profile of fetal brain (GZ) with GC content (GC), gene number, DNase I hypersensitivity (DHS) of fetal brain, and gene expression level in fetal laminae. **d.** GO enrichment of genes located in the top 1000 highly interacting inter-chromosomal regions specific to CP vs. GZ (left), and CP vs. ES (right), indicating that genes located on dynamic chromosomal regions are enriched for neuronal development. **e.** The top 2% highest interacting regions of fetal brains both at GZ and CP (High) show positive correlation in gene expression, while the top 2% lowest interacting regions (Low) and top 2% highly variant regions (Variant) have no skew in distribution. P-values from Kolmogorov–Smirnov test. **f.** The epigenetic state combination in inter-chromosomal interacting regions in GZ. Inter-chromosomal contact frequency map is compared to epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1), transcriptional regulators (TxReg), and transcribed regions (Tx) interact highly to each other as marked in red. Colored bars on the left represent epigenetic marks associated with promoters and transcribed regions (orange), enhancers (red), and repressive marks (blue). Chr, chromosome. Annotation for epigenetic marks described in

http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Figure 2. Compartment and TADs provide insights into gene regulatory mechanism. a.** Leading principal component (PC1) of the intra-chromosomal contact matrix in CP, GZ, and ES, with the DNase I hypersensitivity (DHS) fold change (FC) between ES and fetal brain (FB). PC1 values indicate compartment status of a given region, where positive PC1 represents compartment A (red), and negative PC1 represents compartment B (green). **b.** Distribution of gene expression FC (left) and DHS FC (right) for genes/regions that change compartment status ("A to B" or "B to A") or that remain the same ("stable") in different cell types. P-values from one-way ANOVA. **c.** GO enrichment of genes that change compartment status from A to B (top) and B to A (bottom) in CP to GZ. **d.** Heatmap of PC1 values of the genome that change compartment status in different cell types. **e.** Percentage of epigenetic marks for genomic regions that change compartment status between ES and CP. Note that B to A shift in ES to CP is associated with increased proportion of active transcribed regions (TssA and Tx) and enhancers (Enh, top), while A to B shift in ES to CP is associated with increased proportions of repressive marks (Het and ReprPCWk, bottom). P-values from Fisher's exact test. **f.** Epigenetic changes in topological associating domains (TADs) mediate gene expression changes during neuronal differentiation. Genes were divided by expression FC between ES and differentiated neurons, and epigenetic marks in the TADs containing genes in each group were counted and compared between ES and CP. Upregulated genes in neurons locate in TADs with more active epigenetic marks in CP than in ES, while downregulated genes in neurons locate in TADs with more repressive marks in CP than in ES. Epigenetic states associated with activation and transcription of the genes were marked as a red bar, while those associated with repression were marked as blue bars on the right. Annotation for epigenetic marks

described in http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Figure 3. Genetic architecture of human-gained enhancers. a.** Fraction of epigenetic states for regions interacting to human-gained enhancers in CP and GZ. **b.** Proportions of whether human-gained enhancers and interacting regions are within the same topological associating domain (TAD) vs. outside of the TAD. **c.** Overlap between human-gained enhancer interacting genes (Hi-$C_{evol}$ genes) in CP and GZ with closest genes to human-gained enhancers (left) and Hi-$C_{evol}$ genes in ES (right). **d.** Representative interaction map of a 10kb bin, in which human-gained enhancers reside, with the corresponding 1Mb flanking regions. This interactome map provides genes of action that interact with human-gained enhancers. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; Evol, genomic coordinate for a 10kb bin in which human-gained enhancers reside; -log10(P-value), P-value for the significance of the interaction between human-gained enhancers and each 10kb bin, grey dotted line for FDR=0.01; TAD, TAD borders in CP, GZ, and ES. **e.** GO enrichment for Hi-$C_{evol}$ genes in CP (left) and GZ (right). **f.** Number of primate-specific long non-coding RNAs (lncRNAs) interacting with human-gained enhancers in CP (red vertical lines in the graph) against a background control generated from 3,000 permutations, where the number of lncRNAs interacting with the same number of enhancers pooled from all fetal brain enhancers was counted. **g.** Overrepresentation of Hi-$C_{evol}$ genes in different tissues and closest genes with a curated set of intellectual disability (ID) risk genes. *P<0.05, **P<0.01, *** P<0.001. TSS, transcription start site; OR, odds ratio; GPCR, G-protein coupled receptor; Hi-C genes: GZ, CP, ES, Hi-$C_{evol}$ genes in each tissue; Hi-C genes: FB, union of Hi-$C_{evol}$ genes in GZ and CP; Hi-C genes: ES-specific, Hi-$C_{evol}$ genes in ES but not in fetal brain (FB); Hi-C genes: FB-specific, Hi-$C_{evol}$ genes in FB (union) but not in ES; Closest genes, closest genes to human-gained enhancers.

**Figure 4. Annotation of significant chromatin interactions for schizophrenia-associated loci. a.** Overlap between closest genes to index SNPs (Closest), genes locating in linkage disequilibrium (LD), and genes identified through SNP categorization and chromatin contacts in CP and GZ (Hi-$C_{SCZ}$ genes, diagram in **Extended Data Fig. 7**). **b.** Number of closest genes and LD genes that interact to credible SNPs (Hi-C supported) and those that do not interact to credible SNPs (Hi-C non-supported, top). Number of genes that interact to credible SNPs that are closest to or in LD with index SNPs (Hi-C genes), and not closest to or in LD with index SNPs (Hi-C genes not, bottom). Note that Hi-C genes here contain physically interacting genes, but not genes identified by functional SNPs or promoter SNPs. **c.** Distance between CAVIAR/index SNPs and their target genes for closest genes to index SNPs (Closest), genes locating in linkage disequilibrium (LD), and Hi-$C_{SCZ}$ genes in CP (CP) and GZ (GZ) **d.** GO enrichment for Hi-$C_{SCZ}$ genes in CP (left) and GZ (right). **e.** Representative interaction map of a 10kb bin, in which credible SNPs reside, to the corresponding 1Mb flanking regions. This interactome provides target genes interacting to credible SNPs-containing region. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; SNP, genomic coordinate for a 10kb bin in which credible SNPs locate; -log10(P-value), P-value for the significance of the interaction between credible SNPs and each 10kb bin, grey dotted line for FDR=0.01; GWAS loci, LD region for the index SNP; TAD, topological associating domain borders in CP, GZ, and ES.

## Acknowledgements

This work was supported by NIH grants to D.H.G. (5R01MH060233:05; 5R01MH100027) and J.L.S (K99MH102357), Glenn/AFAR Postdoctoral Fellowship Program for Translational Research on Aging (20145357) and Basic Science Research Program through the National Research Foundation of Korea (NRF) (2013024227) to H.W., the California Institute for Regenerative Medicine (CIRM)-BSCRC Training Grant to L.T.U. (TG2-01169). The Hi-C library was sequenced by Broad Stem Cell Research Center (BSCRC), and fetal tissue was collected from Center for Aids Research (CFAR, 5P30 AI028697). eQTL data was kindly provided by Mina Ryten and Adaiklavan Ramasamy. We thank Suhua Feng, Michael Gandal, Yuan Tian, Vivek Swarup, Grant Belgard for helpful discussions and critical reading of the manuscript.

## Author Contributions

H.W. designed and performed experiments, interpreted results, and co-wrote the manuscript. L.T.U. performed sample collection and experiments. J.L.S., N.N.P., and F.H. analyzed data. C.L. helped establishing Hi-C protocol. J.E. and E.E. participated in the discussion of the results. D.H.G. supervised the experimental design and analysis, interpreted results, provided funding, and co-wrote the manuscript.

## Author Information

*Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles*

Hyejung Won, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Changhoon Lee, Daniel H. Geschwind

*Department of Biological Chemistry, University of California California Los Angeles*

Jason Ernst

*Department of Computer Science, University of California Los Angeles*

Farhad Hormozdiari, Eleazar Eskin

*Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles*

Daniel H. Geschwind, Eleazar Eskin, Jason Ernst

## References

1 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

2 Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

3 Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).

4 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294, doi:10.1038/nature12644 (2013).

5 Network & Pathway Analysis Subgroup of Psychiatric Genomics, C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209, doi:10.1038/nn.3922 (2015).

6 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021, doi:10.1016/j.cell.2013.10.031 (2013).

7 Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007, doi:10.1016/j.cell.2013.10.020 (2013).

8 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159, doi:10.1126/science.1260943 (2015).

9 Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

10 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

11 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

12 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

13 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).

14 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64, doi:10.1038/nature12593 (2013).

15 Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206, doi:10.1038/nature13185 (2014).

16 Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-1065, doi:10.1016/j.cell.2015.07.048 (2015).

17 van de Leemput, J. *et al.* CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. *Neuron* **83**, 51-68, doi:10.1016/j.neuron.2014.05.013 (2014).

18 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

19 Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).

20 Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465-1470, doi:10.1126/science.aaa1975 (2015).

21 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).

22 Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635-640, doi:10.1038/nature12943 (2014).

23 Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647, doi:10.1016/j.neuron.2013.10.045 (2013).

24 Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428, doi:10.1038/nn.3801 (2014).

25 Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508, doi:10.1534/genetics.114.167908 (2014).

26 McCarthy, S. E. *et al.* De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658, doi:10.1038/mp.2014.29 (2014).

27 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221, doi:10.1038/nature13908 (2014).

28 Miller, J. A., Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12698-12703, doi:10.1073/pnas.0914257107 (2010).

29 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

30 Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic acids research* **42**, 87-96, doi:10.1093/nar/gkt857 (2014).

**Extended Data Figure 1. Basic characterization of Hi-C libary. a.** Hi-C library sequencing information. Percentage for double-stranded (DS) reads indicates percentage of DS reads to all reads, and percentage for valid pairs and filtered reads indicates percentage of valid pairs and filtered reads to DS reads. **b.** Frequency distribution of Hi-C contacts in GZ (left) and CP (right) **c.** Size distribution of topological associating domains (TADs) in GZ (left) and CP (right). **d.** Size distribution of genomic regions in between TADs that are less than 400kb (TAD boundaries) in GZ (left) and CP (right). **e.** Size distribution of genomic regions in between TADs that are bigger than 400kb (unorganized chromosome) in GZ (left) and CP (right). Cis ratio, ratio of cis (intra-chromosomal) reads to the total number of reads; chr, chromosome.

**Extended Data Figure 2. Chromosome conformation is associated with various genomic features.** a. Spearman correlation of principal components (PCs) of chromatin interaction profile of CP with GC content (GC), gene number, DNase I hypersensitivity (DHS), and gene expression level of fetal brains. **b.** GO enrichment of genes located in the top 1000 regions that gain inter-chromosomal interactions in CP compared to ES (upper left), ES compared to CP (upper right), CP compared to GZ (lower left), and GZ compared to CP (lower right). **c.** Top 5% (left) and 10% (middle) highest interacting regions both in GZ and CP (High) show positive correlation in gene expression, while low interacting regions (Low) and variant interacting regions (Variant) have no skew in distribution. (Right) Mean (top) and median (bottom) values for gene expression correlation for high, low, and variant interacting regions with different cutoffs, indicating that higher the interaction, higher the correlation of gene expression. **d.** Percentage of epigenetic marks for genomic regions that change compartment status between ES and GZ. Note that B to A shift in ES to GZ is associated with increased proportion of active transcribed regions (TssA and Tx) and enhancers (Enh, top), while A to B shift in ES to GZ is associated with increased proportions of repressive marks (Het and ReprPCWk, bottom). P-values from Fisher’s exact test. Annotation for epigenetic marks described in a core 15-state model from
http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 3. Interacting regions share epigenetic states. a.** Epigenetic state combination in inter-chromosomal interacting regions in CP. Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1), transcriptional regulatory regions (TxReg), and transcribed regions (Tx) interact highly to each other as marked in red. **b-c.** Epigenetic state combination in intra-chromosomal interacting regions in GZ (**b**) and CP (**c**). Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1) and transcriptional regulatory regions (TxReg) interact highly to promoters (PromD1, PromD2) and transcribed regions (Tx5', Tx) as marked in red. Inter- and intra-chromosomal contact frequency map is compared to epigenetic state combination matrix by Fisher’s exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Colored bars on the left represent epigenetic marks associated with promoters and transcribed regions (orange), enhancers (red), and repressive marks (blue). Annotation for epigenetic marks described in a 25-state model from
http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 4. Characterization of chromatin interactome of human-gained enhancers. a.** Distribution fitting of normalized chromatin interaction frequency between human-gained enhancers with 1Mb upstream (top) and 100kb upstream (bottom) regions. Weibull distribution (red line) fits Hi-C interaction frequency the best for every distance range. **b.** Distribution of the number of significant interacting loci to human-gained enhancers in GZ (top), CP (middle), and ES (bottom). **c.** Fraction of histone states (left) and epigenetic mark enrichment (right) for regions interacting with

human-gained enhancers in GZ and CP. CDF, cumulative distribution function; Annotation for epigenetic marks described in http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 5. Human-gained enhancers interact to evolutionary lineage-specific long non-coding RNAs (lncRNAs). a.** Protein-coding genes interacting with human-gained enhancers in CP (CP) and GZ (GZ) have lower non-synonymous substitutions (dN)/synonymous substitutions (dS) ratio compared to protein-coding genes non-interacting to human-gained enhancers (All) in mammals (mouse), primates (rhesus macaque), and great apes (chimpanzee), indicative of purifying selection. **b.** Number of lineage-specific lncRNAs interacting to human-gained enhancers (red vertical lines in the graph) in GZ (top) and CP (bottom). Null distribution generated from 3,000 permutations, where the number of lncRNAs interacting to the same number of enhancers pooled from all fetal brain enhancers was counted.

**Extended Data Figure 6. Association between eQTL and Hi-C interaction. a.** Overlap between eQTL transcripts and genes physically interacting to eQTL SNPs in CP and GZ. Significance of the overlap between eQTL transcripts and Hi-C interacting genes described in the upper right (Fisher's exact test). Background gene list for Fisher's exact test is all transcripts assessed in eQTL study within 1Mb from eQTL SNPs. **b-d.** Histone state enrichment for eQTL SNPs in adult frontal cortex (FCTX, **b**), fetal brain (FB, **c**), and IMR90 (**d**). **e.** Hi-C interaction frequency between eQTL SNPs and transcripts is greater than expected by chance in the relevant cell type. Lowess smooth curve plotted with actual data points. CP, chromatin contact frequency in CP; GZ, chromatin contact frequency in GZ; ES, chromatin contact frequency in ES; Exp, expected interaction frequency given the distance between two regions; Opp, opposite interaction frequency: interaction frequency of SNPs and transcripts when the position of genes was mirrored relative to the eQTL SNP. ***P<0.001, P-values from repeated measure of ANOVA.

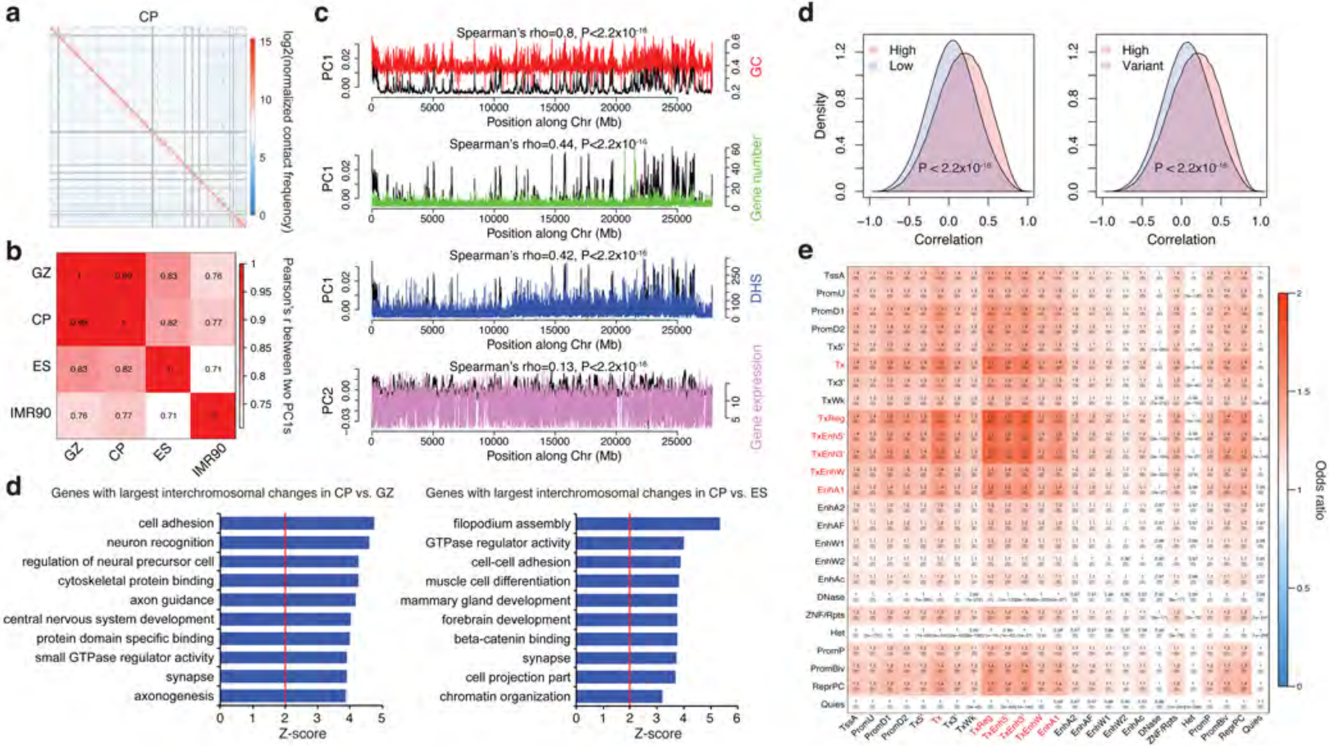
**Extended Data Figure 7. Defining schizophrenia risk genes based on functional annotation of credible SNPs.** Credible SNPs were selected using CAVIAR and categorized into functional SNPs, SNPs that fall onto gene promoters, and un-annotated SNPs. Histone state enrichment of credible SNPs was assessed in fetal brain (FB) and adult frontal cortex (FCTX). Functional SNPs and promoter SNPs were directly assigned to the target genes, while un-annotated SNPs were assigned to the target genes via Hi-C interactions in CP and GZ. GO enrichment for genes identified by each category is shown in the bottom. NMD, nonsense-mediated decay; TSS, transcription start site.

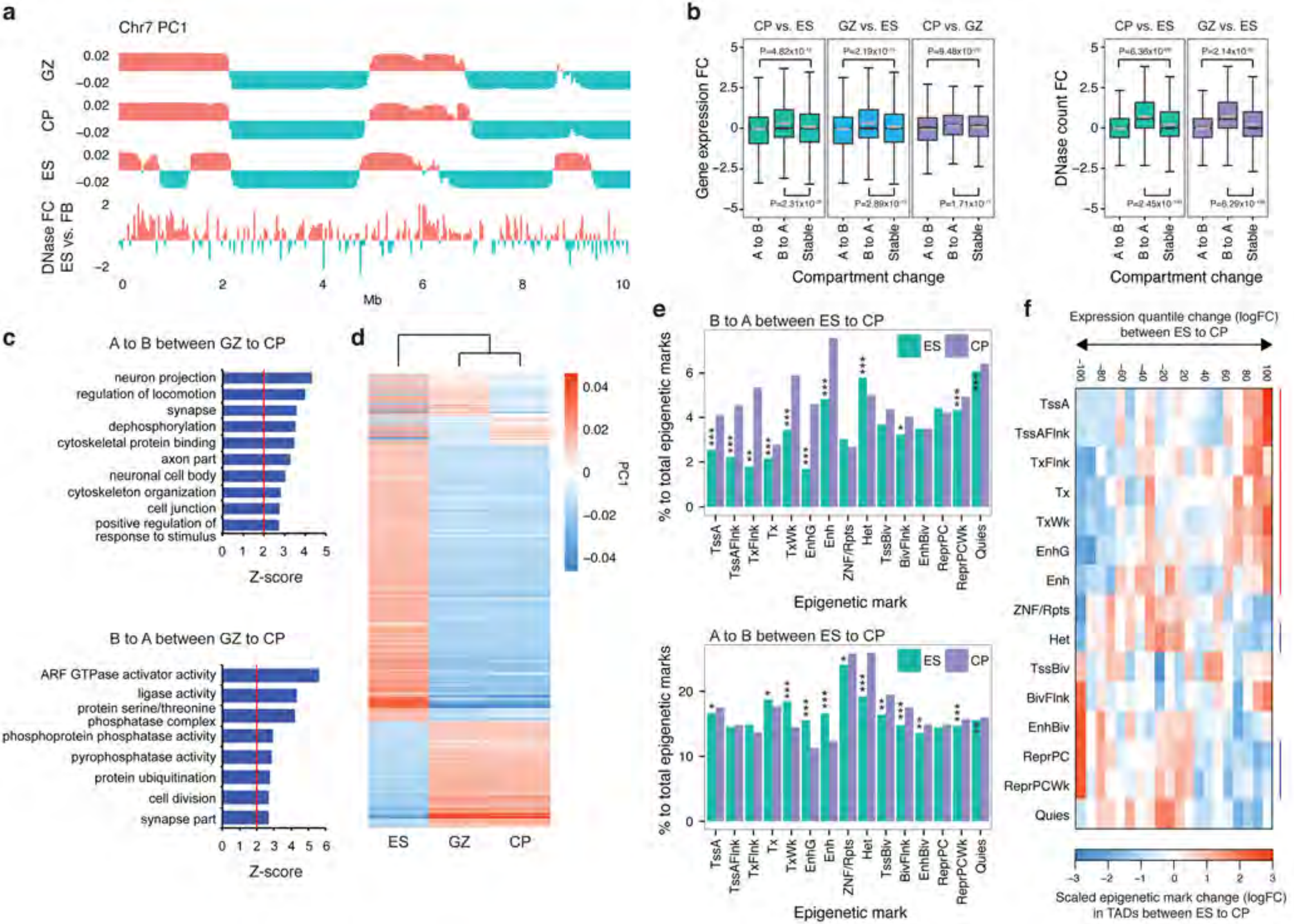
**Extended Data Figure 8. Representative interaction maps for credible SNPs to 1Mb flanking regions.** Interaction maps provide gene of actions for credible SNPs based on physical interaction. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; SNP, genomic coordinate for a 10kb bin in which credible SNPs locate; -log10(P-value), P-value for the significance of the interaction between credible SNPs and each 10kb bin, grey dashed line for FDR=0.01; GWAS loci, linkage disequilibrium (LD) region with the index SNP; TAD, TAD borders in CP, GZ, and ES.

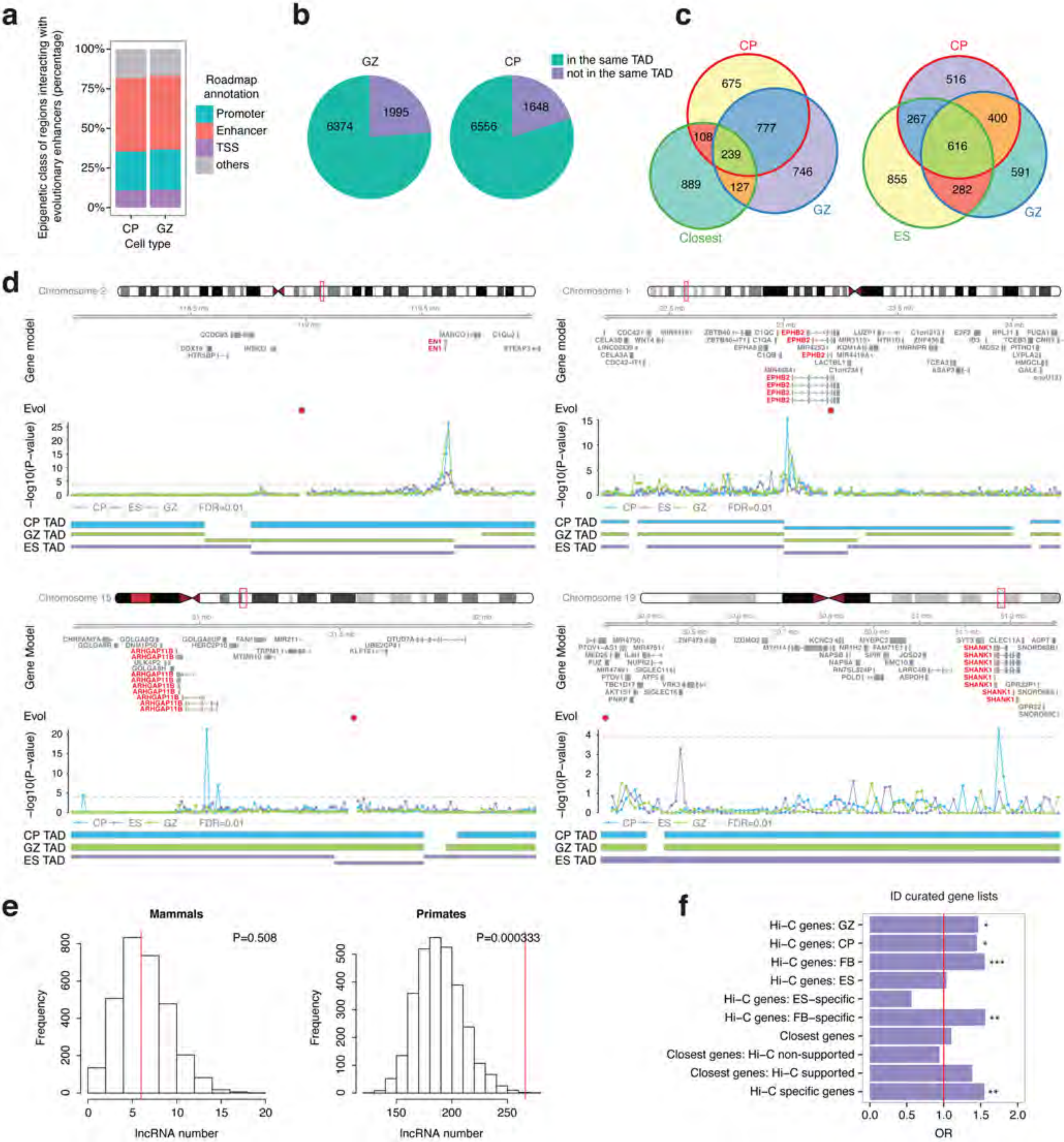
**Extended Data Figure 9. GO enrichment for schizophrenia risk genes curated by various methods. a-b.** GO enrichment for the closest genes to index SNPs (**a**) and genes in linkage disequilibrium (LD) with index SNPs (**b**) that are identified by a schizophrenia risk gene assessment pipeline in **Extended Data Fig. 7** (right) vs. not (left). **c.** GO enrichment for schizophrenia risk genes identified by a pipeline in **Extended Data Fig. 7** that are neither the closest genes nor in LD to index SNPs. Intersect and

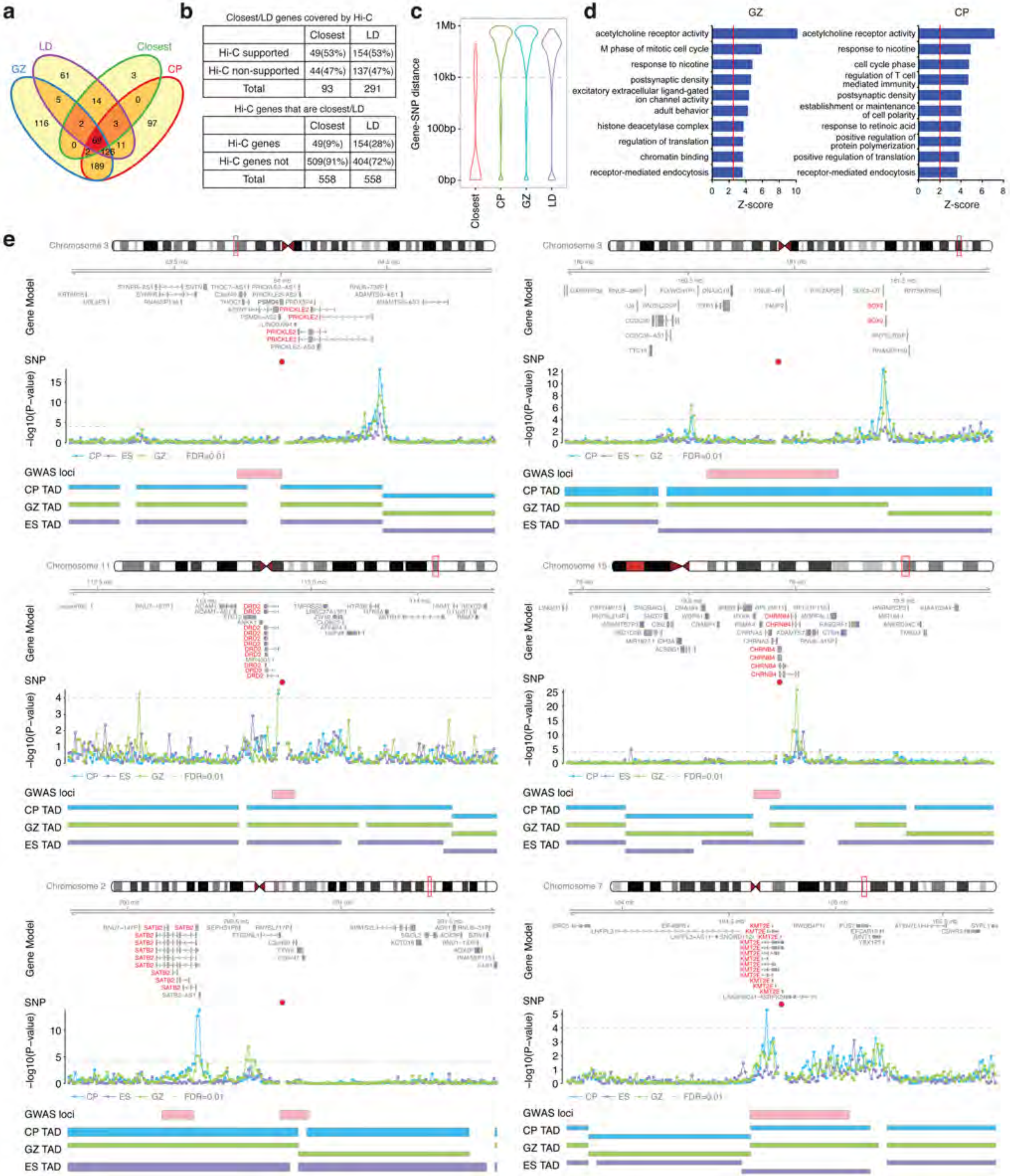
union between CP and GZ in left and right, respectively. Venn diagrams are marked in orange to depict the gene list assessed for GO enrichment.

**Extended Data Figure 10. Defining schizophrenia risk genes based on functional annotation of another set of credible SNPs.** Credible SNPs defined in the original study were categorized into functional SNPs, SNPs that fall onto gene promoters, and un-annotated SNPs. Overlap between credible SNPs identified by CAVIAR and credible SNPs originally identified indicates that two credible SNP lists overlap with each other. Histone state enrichment of credible SNPs in fetal brain (FB) and adult frontal cortex (FCTX). Functional SNPs and promoter SNPs were directly assigned to the target genes, while un-annotated SNPs were assigned to the target genes via Hi-C interactions in CP and GZ. GO enrichment for genes identified by each category and combined gene list is shown in the bottom. NMD, nonsense-mediated decay; TSS, transcription start site.


a
CP
log2(normalized contact frequency)
15
10
5
0
b
GZ
CP
ES
IMR90
1
0.99
0.83
0.76
0.99
1
0.82
0.77
0.83
0.82
1
0.71
0.76
0.77
0.71
1
Pearson's r between two PC1s
1
0.95
0.9
0.85
0.8
0.75
c
Spearman's rho=0.8, P<2.2x10^-16
PC1
GC
Spearman's rho=0.44, P<2.2x10^-16
Gene number
Spearman's rho=0.42, P<2.2x10^-16
DHS
Spearman's rho=0.13, P<2.2x10^-16
PC2
Gene expression
Position along Chr (Mb)
0
5000
10000
15000
20000
25000
d
High
Low
Variant
Density
Correlation
P < 2.2x10^-16
-1.0
-0.5
0.0
0.5
1.0
e
TssA
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
Odds ratio
2
1.5
1
0.5
0
d
Genes with largest interchromosomal changes in CP vs. GZ
cell adhesion
neuron recognition
regulation of neural precursor cell
cytoskeletal protein binding
axon guidance
central nervous system development
protein domain specific binding
small GTPase regulator activity
synapse
axonogenesis
Z-score
Genes with largest interchromosomal changes in CP vs. ES
filopodium assembly
GTPase regulator activity
cell-cell adhesion
muscle cell differentiation
mammary gland development
forebrain development
beta-catenin binding
synapse
cell projection part
chromatin organization
Z-score
0
1
2
3
4
5
6


a
Chr7 PC1
GZ
CP
ES
0.02
-0.02
DNase FC ES vs. FB
2
-2
0
2
4
6
8
10
Mb
b
CP vs. ES
GZ vs. ES
CP vs. GZ
Gene expression FC
5
2.5
0
-2.5
-5
P=4.82x10
P=2.19x10
P=9.48x10
P=2.31x10
P=2.89x10
P=1.71x10
A to B
B to A
Stable
Compartment change
DNase count FC
P=6.36x10
P=2.14x10
P=2.45x10
P=6.29x10
c
A to B between GZ to CP
neuron projection
regulation of locomotion
synapse
dephosphorylation
cytoskeletal protein binding
axon part
neuronal cell body
cytoskeleton organization
cell junction
positive regulation of response to stimulus
Z-score
B to A between GZ to CP
ARF GTPase activator activity
ligase activity
protein serine/threonine phosphatase complex
phosphoprotein phosphatase activity
pyrophosphatase activity
protein ubiquitination
cell division
synapse part
Z-score
d
ES
GZ
CP
PC1
0.04
0.02
0
−0.02
−0.04
e
B to A between ES to CP
A to B between ES to CP
ES
CP
% to total epigenetic marks
Epigenetic mark
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
***
**
*
f
Expression quantile change (logFC) between ES to CP
-100
-80
-60
-40
-20
20
40
60
80
100
Scaled epigenetic mark change (logFC) in TADs between ES to CP
−3
−2
−1
0
1
2
3


a
Epigenetic class of regions interacting with evolutionary enhancers (percentage)
100%
75%
50%
25%
0%
CP
GZ
Cell type
Roadmap annotation
Promoter
Enhancer
TSS
others
b
GZ
6374
1995
CP
6556
1648
in the same TAD
not in the same TAD
c
CP
675
108
239
777
889
127
746
GZ
Closest
CP
516
267
400
616
855
282
591
GZ
ES
d
Chromosome 2
Gene model
EN1
EN1
Evol
-log10(P-value)
CP
ES
GZ
FDR=0.01
CP TAD
GZ TAD
ES TAD
Chromosome 1
Gene model
EPHB2
Evol
-log10(P-value)
CP TAD
GZ TAD
ES TAD
Chromosome 15
Gene Model
ARHGAP11B
Evol
-log10(P-value)
CP TAD
GZ TAD
ES TAD
Chromosome 19
Gene Model
SHANK1
Evol
-log10(P-value)
CP TAD
GZ TAD
ES TAD
e
Mammals
P=0.508
Frequency
lncRNA number
Primates
P=0.000333
Frequency
lncRNA number
f
ID curated gene lists
Hi-C genes: GZ *
Hi-C genes: CP *
Hi-C genes: FB ***
Hi-C genes: ES
Hi-C genes: ES-specific
Hi-C genes: FB-specific **
Closest genes
Closest genes: Hi-C non-supported
Closest genes: Hi-C supported
Hi-C specific genes **
OR
0.0
0.5
1.0
1.5
2.0



b

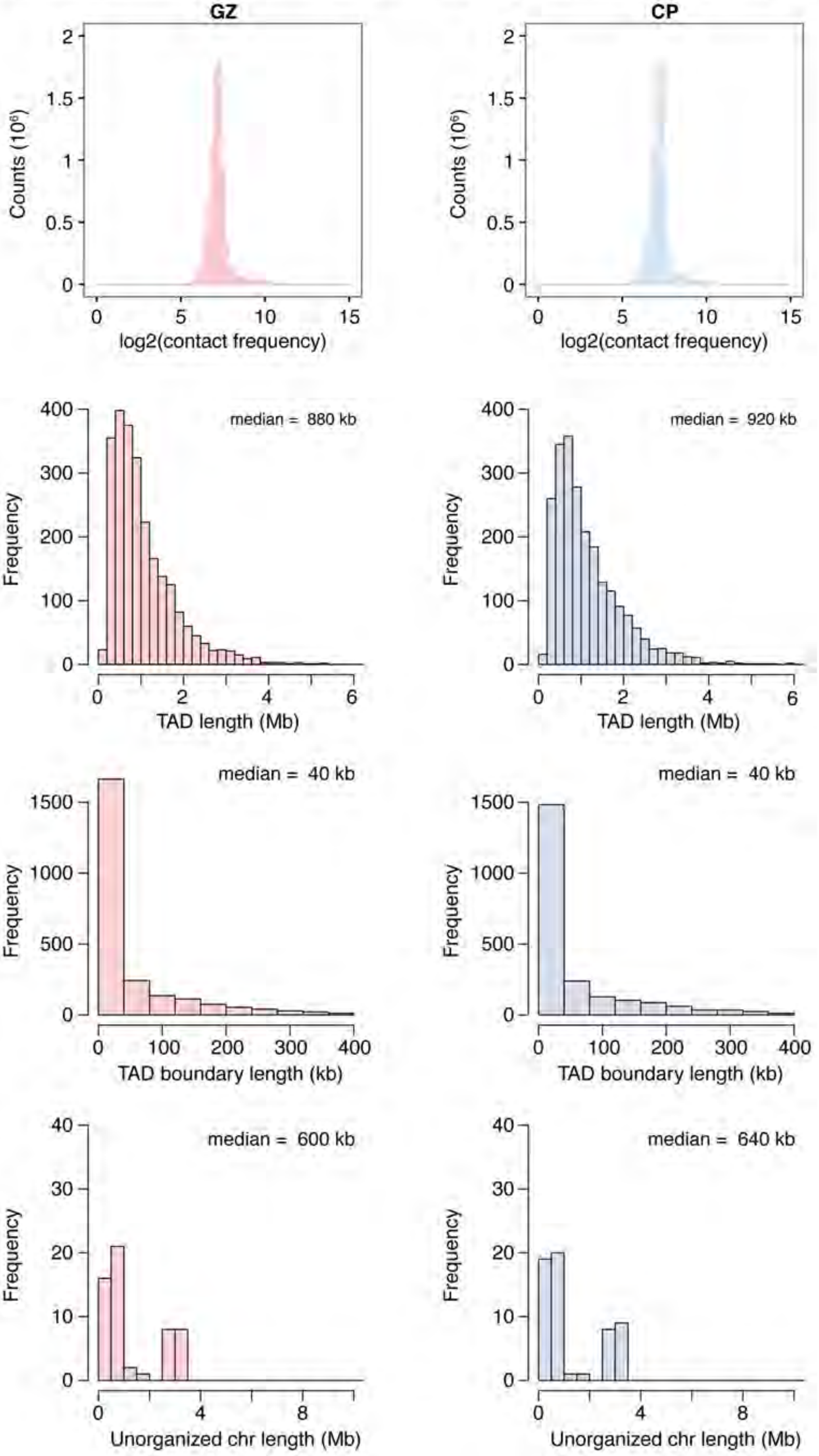
Closest/LD genes covered by Hi-C

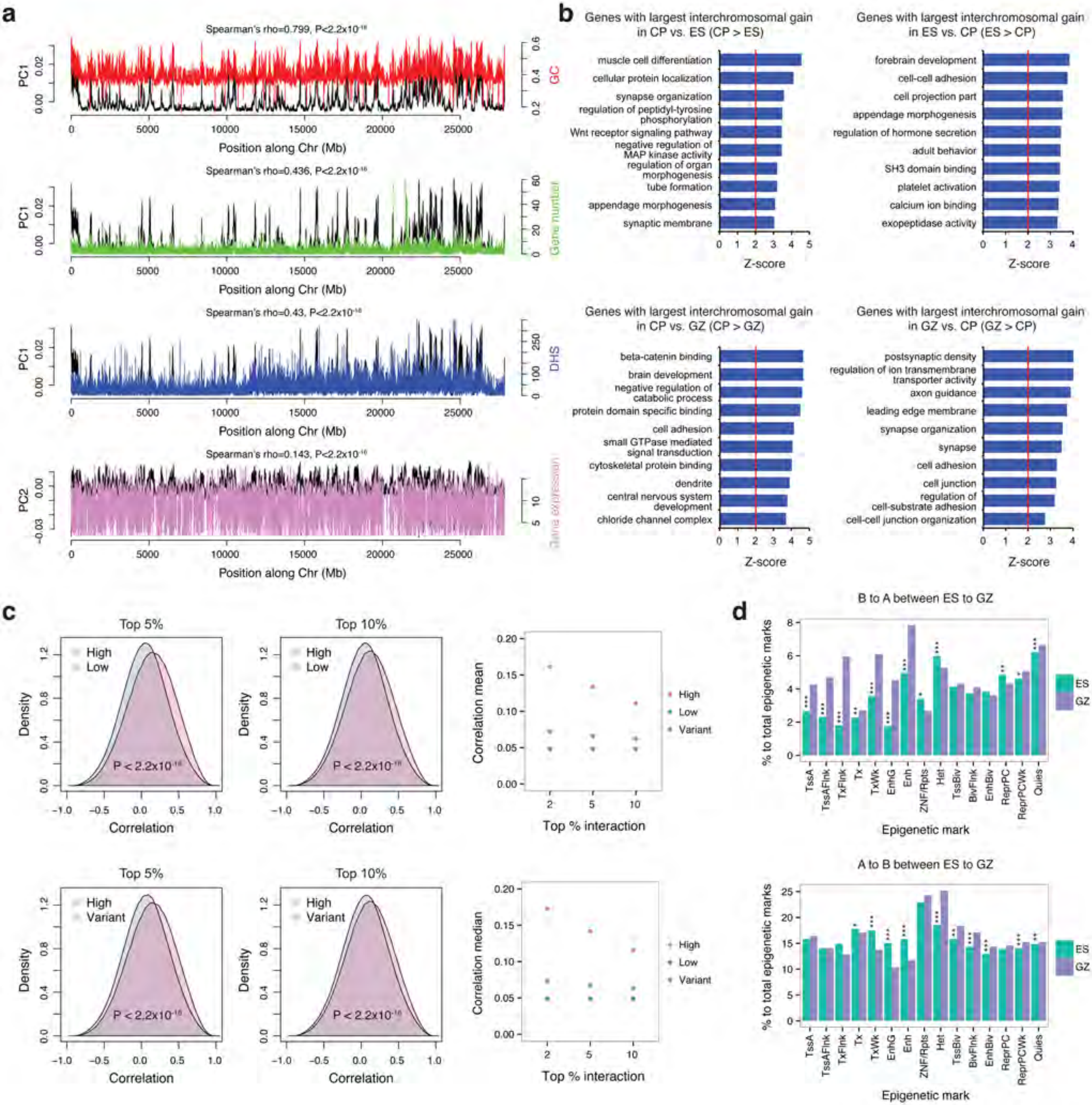
| | Closest | LD |
|---|---|---|
| Hi-C supported | 49(53%) | 154(53%) |
| Hi-C non-supported | 44(47%) | 137(47%) |
| Total | 93 | 291 |

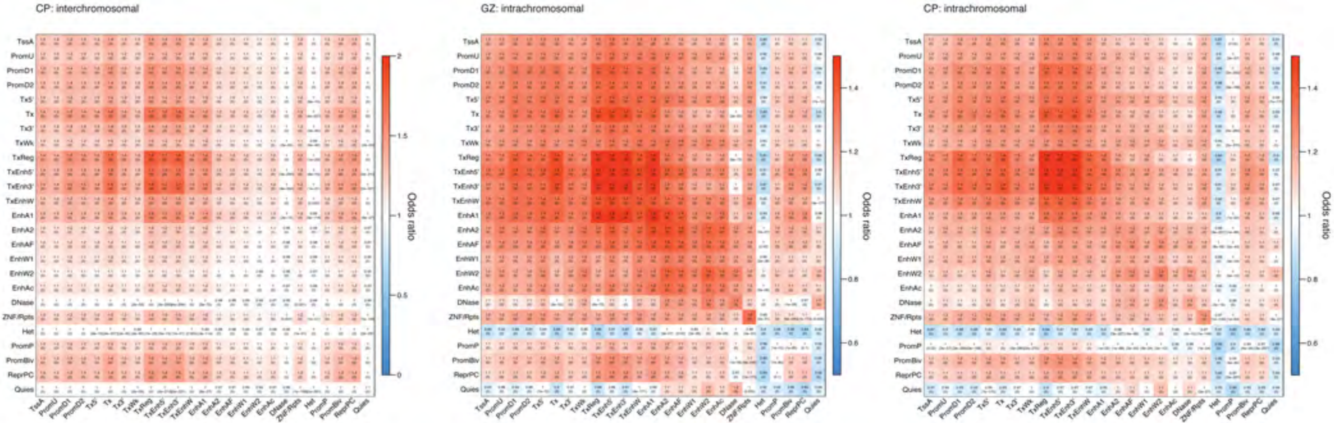
Hi-C genes that are closest/LD

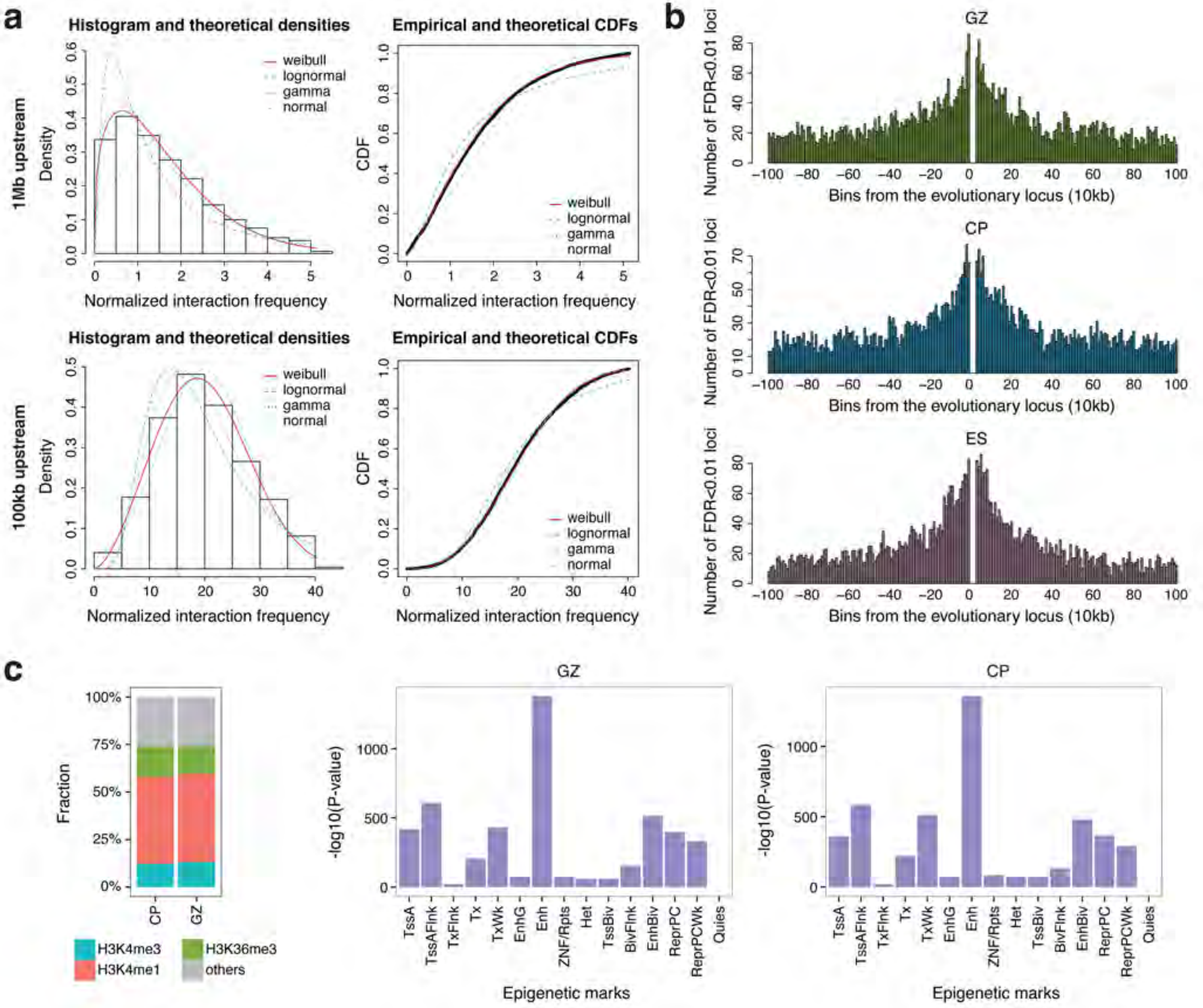
| | Closest | LD |
|---|---|---|
| Hi-C genes | 49(9%) | 154(28%) |
| Hi-C genes not | 509(91%) | 404(72%) |
| Total | 558 | 558 |

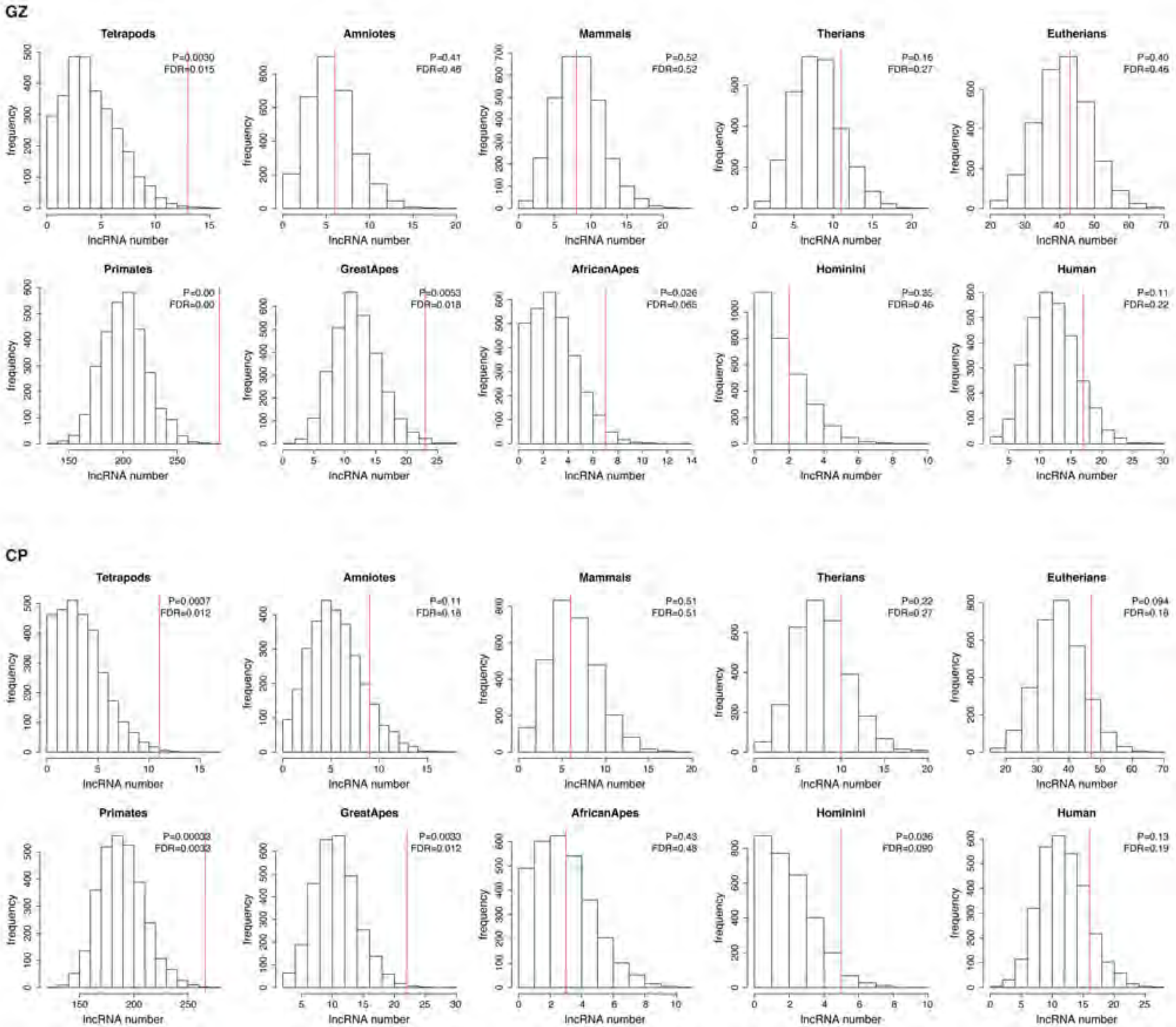
| Cell type | Cis ratio | All reads | DS mapped reads | Valid pairs | Filtered reads |
|---|---|---|---|---|---|
| GZ | 47.45% | 1,991,686,360 | 1,407,918,128<br>(70.69%) | 1,243,116,106<br>(88.29%) | 1,048,911,579<br>(74.50%) |
| CP | 46.40% | 1,958,637,304 | 1,352,951,087<br>(69.08%) | 1,225,315,488<br>(90.57%) | 1,022,593,960<br>(75.58%) |




a
Spearman's rho=0.799, P<2.2x10^-16
PC1
GC
Position along Chr (Mb)
Spearman's rho=0.436, P<2.2x10^-16
PC1
Gene number
Position along Chr (Mb)
Spearman's rho=0.43, P<2.2x10^-16
PC1
DHS
Position along Chr (Mb)
Spearman's rho=0.143, P<2.2x10^-16
PC2
Gene expression
Position along Chr (Mb)
b
Genes with largest interchromosomal gain in CP vs. ES (CP > ES)
muscle cell differentiation
cellular protein localization
synapse organization
regulation of peptidyl-tyrosine phosphorylation
Wnt receptor signaling pathway
negative regulation of MAP kinase activity
regulation of organ morphogenesis
tube formation
appendage morphogenesis
synaptic membrane
Z-score
Genes with largest interchromosomal gain in ES vs. CP (ES > CP)
forebrain development
cell-cell adhesion
cell projection part
appendage morphogenesis
regulation of hormone secretion
adult behavior
SH3 domain binding
platelet activation
calcium ion binding
exopeptidase activity
Z-score
Genes with largest interchromosomal gain in CP vs. GZ (CP > GZ)
beta-catenin binding
brain development
negative regulation of catabolic process
protein domain specific binding
cell adhesion
small GTPase mediated signal transduction
cytoskeletal protein binding
dendrite
central nervous system development
chloride channel complex
Z-score
Genes with largest interchromosomal gain in GZ vs. CP (GZ > CP)
postsynaptic density
regulation of ion transmembrane transporter activity
axon guidance
leading edge membrane
synapse organization
synapse
cell adhesion
cell junction
regulation of cell-substrate adhesion
cell-cell junction organization
Z-score
c
Top 5%
High
Low
P < 2.2x10^-16
Density
Correlation
Top 10%
High
Low
P < 2.2x10^-16
Density
Correlation
Correlation mean
Top % interaction
High
Low
Variant
Top 5%
High
Variant
P < 2.2x10^-16
Density
Correlation
Top 10%
High
Variant
P < 2.2x10^-16
Density
Correlation
Correlation median
Top % interaction
High
Low
Variant
d
B to A between ES to GZ
% to total epigenetic marks
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
Epigenetic mark
ES
GZ
A to B between ES to GZ
% to total epigenetic marks
Epigenetic mark
ES
GZ


CP: interchromosomal
GZ: intrachromosomal
CP: intrachromosomal
TssA
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
Odds ratio
0
0.5
1
1.5
2
0.6
0.8
1
1.2
1.4


a
Histogram and theoretical densities
1Mb upstream
Density
0.0 0.1 0.2 0.3 0.4 0.5 0.6
0 1 2 3 4 5
Normalized interaction frequency
weibull
lognormal
gamma
normal
Empirical and theoretical CDFs
CDF
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3 4 5
Normalized interaction frequency
weibull
lognormal
gamma
normal
Histogram and theoretical densities
100kb upstream
Density
0.0 0.1 0.2 0.3 0.4 0.5
0 10 20 30 40
Normalized interaction frequency
weibull
lognormal
gamma
normal
Empirical and theoretical CDFs
CDF
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40
Normalized interaction frequency
weibull
lognormal
gamma
normal
b
GZ
Number of FDR<0.01 loci
0 20 40 60 80
−100 −80 −60 −40 −20 0 20 40 60 80 100
Bins from the evolutionary locus (10kb)
CP
Number of FDR<0.01 loci
0 10 30 50 70
−100 −80 −60 −40 −20 0 20 40 60 80 100
Bins from the evolutionary locus (10kb)
ES
Number of FDR<0.01 loci
0 20 40 60 80
−100 −80 −60 −40 −20 0 20 40 60 80 100
Bins from the evolutionary locus (10kb)
c
Fraction
0% 25% 50% 75% 100%
CP GZ
H3K4me3
H3K36me3
H3K4me1
others
GZ
-log10(P-value)
0 500 1000
TssA TssAFlnk TxFlnk Tx TxWk EnhG Enh ZNF/Rpts Het TssBiv BivFlnk EnhBiv ReprPC ReprPCWk Quies
Epigenetic marks
CP
-log10(P-value)
0 500 1000
TssA TssAFlnk TxFlnk Tx TxWk EnhG Enh ZNF/Rpts Het TssBiv BivFlnk EnhBiv ReprPC ReprPCWk Quies
Epigenetic marks


GZ
Tetrapods
P=0.0030
FDR=0.015
Amniotes
P=0.41
FDR=0.46
Mammals
P=0.52
FDR=0.52
Therians
P=0.16
FDR=0.27
Eutherians
P=0.40
FDR=0.46
Primates
P=0.00
FDR=0.00
GreatApes
P=0.0053
FDR=0.018
AfricanApes
P=0.026
FDR=0.065
Hominini
P=0.35
FDR=0.46
Human
P=0.11
FDR=0.22
CP
Tetrapods
P=0.0007
FDR=0.012
Amniotes
P=0.11
FDR=0.18
Mammals
P=0.51
FDR=0.51
Therians
P=0.22
FDR=0.27
Eutherians
P=0.094
FDR=0.18
Primates
P=0.00033
FDR=0.0033
GreatApes
P=0.0033
FDR=0.012
AfricanApes
P=0.43
FDR=0.48
Hominini
P=0.036
FDR=0.090
Human
P=0.13
FDR=0.19
frequency
lncRNA number

## APPENDIX

### Content:

- Preprint of manuscript by Parikshak, N. N. et al. Global changes in patterning, splicing and primate specific lncRNAs in autism brain.
- Preprint of manuscript by Won, H. et al. Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development.

# Global changes in patterning, splicing and primate specific lncRNAs in ASD brain

Neelroop N. Parikshak[1,2,*], Vivek Swarup[1,*], T. Grant Belgard[1,2, †,*], Michael Gandal[1,2], Manuel Irimia[5,6], Virpi Leppa[1], Jennifer K. Lowe[1], Robert Johnson[7], Benjamin J. Blencowe[6], Steve Horvath[3-4], Daniel H. Geschwind[1-3]

1. Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
2. Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.
3. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
4. Department of Biostatistics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
5. EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, Barcelona 08003, Spain.
6. Donnelly Centre, University of Toronto, 160 College Street, Toronto, ON M5S 3E1, Canada; Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada.
7. NICHD Brain and Tissue Bank for Developmental Disorders, University of Maryland Medical School, Baltimore, Maryland 21201, USA.

*These authors contributed equally to this study.

†Current address: MRC Functional Genomics Unit, Department of Physiology, Anatomy & Genetics, University of Oxford, South Parks Road, Oxford, OX1 3PT, United Kingdom.

## Summary

We apply transcriptome-wide RNA sequencing in post mortem ASD brain and controls, identifying convergent alterations in the noncoding transcriptome, including primate specific lncRNA and transcript splicing in cerebral cortex, but not in cerebellum, of subjects who had ASD. We characterize an attenuation of patterning between frontal and temporal cortex in ASD and identify *SOX5*, a transcription factor involved in cortical neuron fate specification, as a likely driver of this pattern. We further show that a genetically defined subtype of ASD, Duplication 15q Syndrome, shares the core transcriptomic signature of idiopathic ASD, indicating that observed molecular convergence in autism brain is the likely consequence of manifold genetic alterations. Using co-expression network analysis, we show that diverse forms of genetic risk for ASD affect convergent, independently replicated, biological pathways and provide an unprecedented resource for understanding the molecular alterations associated with ASD in humans.

Autism spectrum disorder (ASD) is a neurodevelopmental syndrome characterized by deficits in social communication and mental flexibility[1]. Genetic risk factors contribute substantially to ASD risk, and recent studies support the potential contribution of more than a thousand genes to ASD risk[2-4]. However, given the shared cognitive and behavioral features across the autism spectrum, one hypothesis is that diverse risk factors may converge on common molecular, cellular, and circuit level pathways to result in the shared phenotype[5,6]. Analysis of the transcriptome has been used to identify common molecular pathways in the neocortex (CTX) from postmortem human brain tissue in individuals with ASD[7-11]. However, all transcriptomic studies in ASD to date have been limited to evaluating highly expressed mRNAs corresponding to protein coding genes. Moreover, most lack rigorous replication and do not assess gene expression patterns across brain regions.

We used rRNA-depleted RNA-seq (Methods) to evaluate transcriptomes from a large set of ASD and control (CTL) brain samples including neocortex (frontal and temporal) and cerebellum across 79 individuals (46 ASD, 33 CTL, 205 samples, Extended Data Fig. 1a-e, Supplementary Table 1). We first compared differential gene expression (DGE) between ASD and CTL individuals in CTX from a previously published[7] microarray study against new, independent gene expression profiles from RNA-seq to evaluate global reproducibility of DGE in ASD. We found a high degree of replication of DGE fold changes between the sample sets, despite evaluation on different gene expression platforms (fold changes at $P < 0.05$ in previously evaluated data correlate with new data with $R^2 = 0.60$, Extended Data Fig. 1f). We observed a much weaker overall signal and replication in cerebellum ($R^2 = 0.033$, Extended Data Fig. 1g). These analyses confirm the existence of a reproducible DGE signature in ASD CTX across different platforms and in independent samples.

We next combined samples from all individuals with idiopathic ASD into a covariate-matched "ASD Discovery Set" (Extended Data Fig. 1h) for CTX (106 samples, 26 ASD, 33 CTL individuals) and held out remaining samples for replication ("ASD Replication Set", Methods). For DGE analysis, we used a linear mixed effects model that accounts for biological and technical covariates (Methods) to identify 1156 genes differentially expressed in ASD CTX, 582 increased and 574 decreased (Benjamini-Hochberg FDR ≤ 0.05). Importantly, DGE analysis with additional covariates or different assumptions about the distribution of the data and test statistics yielded similar results (Extended Data Fig. 2a). Additionally this DGE signature clusters over two-thirds of ASD samples together and this clustering is not related to confounding factors such as cortical region, age, sex, and RNA quality (Figure 1a, Extended Data Fig. 2b). The most significantly down-regulated gene was *PVALB* (fold change = 0.53, FDR ≤ 0.05), a marker for GABAergic interneurons. *SST*, a marker for a different subpopulation of GABAergic interneurons, is also among the most downregulated (fold change = 0.61, FDR ≤ 0.05). Other down-regulated genes at FDR ≤ 0.05 include *NEUROD6*, involved in neuronal differentiation (fold change = 0.60), multiple ion channels, and *KDM5D*, a lysine demethylase (fold change = 0.66). In contrast, members of the complement cascade implicated in microglial-neuronal interactions (*C4A*, fold change = 1.94; *C1QB*, fold change = 1.65; both FDR ≤ 0.05) are upregulated in ASD CTX. Gene Ontology (GO) term enrichment analysis further supports the involvement of pathways implicated by these genes (Figure 1b), confirming previous findings[7]. Moreover, the upregulated set is enriched for astrocyte and microglia enriched genes, and the down-regulated set is enriched for synaptic genes (Extended Data Fig. 2c), consistent with previous observations[7,11].

We next sought to evaluate whether the transcriptional signature identified in the ASD Discovery Set generalizes to the ASD Replication set by assessing the 1st principal component of the DGE set, which summarizes the DGE expression pattern across all cortical samples. The ASD Discovery Set and ASD Replication Set share this pattern, which is significantly different for both sets compared to CTL (Figure 1c). Moreover, this pattern is highly associated with ASD diagnosis, but not other biological factors, technical factors, or scores on sub-domains of an ASD diagnostic tool (Figure 1d). These analyses demonstrate that convergent differences in ASD CTX are reproducible in independent samples and are not related to confounding factors.

We also detected 2715 lncRNAs expressed in cerebral cortex (after careful filtering for high-confidence transcripts, Supplementary Information), of which 62 were significantly dysregulated between ASD and CTL (33 long intergenic RNAs, lincRNAs; 19 antisense transcripts; and 10 processed transcripts at FDR $\leq 0.05$). Similar to the protein coding genes, these transcripts' expression patterns cluster ASD and CTL samples (Figure 1e). Most of these lncRNAs are developmentally regulated[12], have chromatin states indicative of transcription start sites (TSSs) near their 5′ end in brain[13], and are identified in other datasets[12,14] consistent with being valid, functional lncRNAs. Moreover, most (81%) exhibit primate-specific expression patterns in brain[15] (Supplementary Information). For example, Figure 1f depicts two lincRNAs, *LINC00693* and *LINC00689*, which are typically downregulated during development, yet are upregulated in ASD CTX relative to controls (Figure 1g), which we validated by RT-PCR (Extended Data Fig. 2d). *LINC00693* sequence is present, but poorly conserved in mouse, while *LINC00689* is primate-specific (present in macaque and other primates but not in any other species, Supplementary Information, Extended Data Fig. 3 for additional examples). These data indicate that dysregulation of lncRNAs, many of which are primate-specific and involved in brain development, is an important component of transcriptome dysregulation observed in ASD.

Previous work suggested that alterations in transcript splicing may contribute to transcriptomic changes in ASD[7,16,17] by evaluating splicing in a targeted manner and pooling samples across individuals[7,16,17]. Given the increased sequencing depth and reduced sequencing bias across transcript length in our dataset, we were able to perform an unbiased genome-wide analysis of differential alternative splicing (AS). We evaluated the percent spliced in (PSI, Extended Data Fig. 4a) for 34,025 AS events in CTX across the ASD Discovery Set, encompassing skipped exons (SE), alternative 5′ splice sites (A5SS), alternative 3′ splice sites (A3SS), and mutually exclusive exons (MXE) using the MATS pipeline[18] (Supplementary Information). We first asked whether there was a global signal, finding significant enrichment over background (Extended Data Fig. 4b). We identified 1127 events in 833 genes at FDR $\leq 0.5$ in CTX (similar to the number of events at uncorrected $P < 0.005$). Importantly, we obtained similar results with a different splice junction mapping and quantification approach (Extended Data Fig. 4c).

We performed PCR validations with nine AS events from the differential splicing set (*ASTN2, MEF2D, ERC2, MED31, SMARCC2, SYNE1, NRCAM, GRIN1, NCAM*) and found that validated changes in splicing patterns were concordant with RNA-seq (Extended Data Fig. 4d-e), demonstrating that our approach identifies alterations in AS with high specificity. Similar to our observations with lncRNA and DGE, AS changes clustered the samples by diagnosis (Figure 2a). The most significantly different event was the inclusion of an exon in *ASTN2* ($\Delta$PSI = 5.8 indicating a mean of 5.8% difference in inclusion in ASD vs CTL; $P = 7.8\text{x}10^{-6}$), a gene implicated by copy number variation (CNV) in ASD and other developmental disorders[19]. GO term analysis of the genes implicated by these pathways indicates involvement of biological processes related to neuronal projection, biological adhesion, and morphogenesis (Figure 2b), pathways where alternative isoforms are critical to specifying interactions between protein products. Moreover, the 1st principal component of the cortex differential splicing signature replicates in the ASD Replication Set and is not associated with other biological or technical factors (Figures 2c-d, Extended Data Fig. 5a). Importantly, many splicing alterations occur in genes that are not differentially expressed between ASD and CTL; removing AS events on genes exhibiting even nominal DGE ($P < 0.05$), still identified a strong difference between ASD and CTL CTX (Extended Data Fig. 5b).

A parallel analysis in cerebellum evaluating 32,954 AS events found no differentially regulated events significant at any multiple comparison correction thresholds (Extended Data Fig. 5c, Supplementary Table 3). There was no detectable global overlap between cerebellum and CTX above chance for events significant at $P < 0.05$ in both comparisons (fold enrichment = 1.1, P = 0.21). This suggests that AS alterations in ASD are largely confined to CTX cell types, consistent with the stronger overall DGE patterns observed in CTX versus cerebellum.

To further explore the underlying biology of AS dysregulation, we tested whether the shared splicing signature in ASD might be a product of perturbations in AS factors known to be important to neural development or preferentially expressed in neural tissue. We found that the expression levels of *RBFOX1*, *RBFOX2*, *SRRM4*, *NOVA1*, and *PTBP1* all had high correlations ($R^2 > 0.35$, $FDR \leq 0.05$) to AS alterations in CTX (Figure 2e), but not in cerebellum (Figure 2f). Furthermore, enrichment analysis revealed that most changes in cortical AS occur in neuron-specific exons that are excluded in ASD (exons with $\Delta PSI > 50\%$ in neurons overlap with exons excluded in ASD CTX, fold enrichment = 4.1, $P = 1.8x10^{-7}$, Extended Data Fig. 5d).

To validate a regulatory relationship between splicing factors and these events, we evaluated experimental data from knockout, overexpression, and knockdown experiments for Rbfox1[20], SRRM4[21], and PTBP1[22], respectively . We found that exons regulated by each of these splicing factors were significantly enriched in the set of exons excluded in ASD (Figure 2g), while in contrast, there was no enrichment for targets of ESRP[23], a splicing factor involved in epithelial cell differentiation but not expressed in CTX. This shows that alterations in three splicing factors dysregulated in ASD regulate AS of the neuron-specific exons whose inclusion is dysregulated in ASD in CTX and not cerebellum, indicating selective alteration of neuronal splicing in ASD CTX. Remarkably, the expression patterns of these three splicing factors (and others for which appropriate validation experiments were unavailable) results in distinct clusters (Extended Data Fig. 5e), suggesting that subsets of splicing factors act in different individuals to mediate a common downstream AS alteration.

Taken together these results indicate global transcriptional alterations in ASD cerebral cortex, but not cerebellum at the level of protein coding transcripts, lncRNA and AS. Therefore, to determine how these different transcriptomic subcategories relate to each other in ASD, we compared the 1st PC for each type of transcriptomic alteration across individuals (Figure 2h). Remarkably, the PCs are highly correlated ($R^2 > 0.8$) indicating that the transcriptomic alteration is a unitary phenomenon across protein coding, noncoding, and splicing levels, rather than distinct forms of molecular alteration.

Previous analysis with gene expression microarrays in a small cohort suggested that the typical pattern of transcriptional differences between the frontal and temporal cortex may be attenuated in ASD[7]. To further test this possibility, we evaluated DGE between CTX regions (Supplementary Information) in 16 matched frontal and temporal CTX sample pairs from ASD and CTL subjects and found 551 genes differentially expressed between regions in controls, but only 51 in ASD ($FDR \leq 0.05$; Figure 3a). We refer to the set of 523 genes with this pattern in CTL, but not ASD as the “Attenuated Cortical Patterning” set. The attenuation of patterning is evident from the global distribution of test statistics between frontal and temporal CTX in ASD and CTL and genes in this set do not show a greater difference in variability in ASD versus controls compared to other genes (Kolmogorov-Smirnov test, two-tailed P = 0.11, Extended Data Fig. 6a).

We complemented this analysis with a machine learning approach using all 123 cortical samples, training a regularized regression model[24] to classify frontal versus temporal CTX with independent gene expression data from BrainSpan[25] (Extended Data Fig. 6b, Supplementary Information). Multiple approaches to training the classifier with BrainSpan can differentiate between frontal and temporal CTX in both CTL and

ASD (Extended Data Fig. 6c-e), demonstrating that dissection and sample quality in our samples are of high quality. Loss of classification accuracy in ASD compared to CTL was observed when restricting the model to the genes with the most attenuated patterning in ASD (Extended Data Fig. 6f), demonstrating that attenuation of patterning generalizes across all samples. The Attenuated Cortical Patterning set includes multiple genes known to be involved in cell-cell communication and cortical patterning, such as *PCDH10*, *PCDH17*, *CDH12, MET*, and *PDGFD*, which was recently shown to mediate human specific aspects of cerebral cortical development[26]. GO term enrichment analysis of the Attenuated Cortical Patterning set identified enrichment for G protein coupled signaling, Wnt receptor signaling, and calcium binding, among several developmental processes (Figure 3b), and cell type enrichment analysis did not identify a strong preference for a particular cell type (Extended Data Fig. 6g).

To identify potential drivers of the alteration in cortical patterning, we evaluated transcription factor binding site enrichment upstream of genes in the Attenuated Cortical Patterning set (Supplementary Information), and found an enrichment of *SOX5* binding motifs (upstream of 364/523 genes, Figure 3c). Remarkably, *SOX5* itself belongs to the Attenuated Cortical Patterning set: while *SOX5* is differentially expressed between frontal and temporal CTX in CTL, it is not in ASD (Figure 3d). We thus predicted that if *SOX5* regulates cortically patterned genes, its expression should correlate with target gene transcript levels. Consistent with this prediction, we found that genes in the Attenuated Cortical Patterning set are anti-correlated with *SOX5* in CTL CTX, but not in ASD CTX (Figure 3e, top left; Wilcoxon rank sum test of R values, P = 0.01), suggesting that the normal role of SOX5 as a transcriptional repressor may be disrupted in ASD. We reasoned that a true loss of SOX5-mediated cortical patterning would be specific to the predicted SOX5 targets. Consistent with this, we find a loss of correlations between *SOX5* and predicted targets, but no difference in correlations between *SOX5* and non-targets in the Attenuated Cortical Patterning set (Figure 3e). Taken together, these findings show that a loss of regional patterning downstream of the transcriptional repressor *SOX5*, which plays a crucial role in glutamatergic neuron development[27,28], contributes to the loss of regional identity in ASD.

Gene expression changes in postmortem brain may be a consequence of genetic factors, environmental factors, or both. Brain tissue from individuals with ASD that harbor known, penetrant genetic causes are very rare. However, we were able to identify postmortem brain tissue from 8 subjects with one of the more common recurrent forms of ASD, Duplication 15q Syndrome (dup15q, which is present in about 0.5-1% of ASD cases, see Extended Data Fig. 7a for characterization of duplications). We performed RNA-seq across frontal and temporal cortex and compared DGE changes in dup15q with those observed in individuals with idiopathic ASD to better understand the extent to which the observed molecular pathology overlaps. As expected, most genes in the 15q11.1-13.2 duplicated region have higher expression in dup15q CTX compared to CTL (Figure 4a), although *SNRPN* and *SNURF* were notably downregulated. Conversely, no significant upregulation of genes in this region were identified in idiopathic ASD or controls. Strikingly, when we assessed genome-wide expression changes, we observed a strong signal of DGE in dup15q that widely overlaps with that of idiopathic ASD (fold changes at FDR $\leq 0.05$ in dup15q correlate with idiopathic ASD with $R^2 = 0.79$, Figure 4b). Moreover, the slope of the best-fit line through these changes is 2.0, indicating that on average, the transcriptional changes in dup15q CTX are highly similar, but twice the magnitude of those observed in ASD CTX.

Next, we sought to evaluate AS changes in dup15q. There is only one significant splicing change in the dup15q region (Supplementary Table 3), consistent with the idea that duplication in this region duplicates all isoforms of the genes, resulting in minimal alteration of transcript structure. Similar to DGE, global AS analysis in dup15q CTX vs to CTL CTX revealed a stronger, but highly overlapping signature with idiopathic

ASD CTX (fold changes at FDR ≤ 0.2 in dup15q agree correlate with idiopathic ASD with $R^2 = 0.66$) indicating that splicing changes in dup15q syndrome recapitulate those of idiopathic ASD (Figure 4c). The slope of the best-fit line through the PSI for spliced exons in dup15q CTX compared to those in ASD CTX is 2.5 similar to DGE. Notably, both gene expression and AS changes in dup15q implicated similar pathways as those found in idiopathic ASD (Extended Data Fig. 7c-d). Clustering dup15q samples and CTL samples using both the DGE set and the differential AS set showed that all dup15q samples cluster together (Figure 4d), as opposed to the more variable clustering of idiopathic ASD, supporting the hypothesis that this shared genetic abnormality leads to a more homogeneous molecular phenotype.

Next, to test whether this molecular ASD signature may be due to independent of postmortem or reactive effects (Supplementary Information), we compared our data with gene expression profiles from a iPSC-derived neurons (nIPSCs)[29] from dup15q were available, we could use these data to definitively reveal which changes in dup15q CTX are independent of postmortem or reactive effects (Supplementary Information), since such effects are not present *in vitro*. We observe that DGE in the 15q region is concordant with that seen in the nIPSCs (Figure 4e), even though the sample size is small and the analysis is likely underpowered. Upregulated changes in dup15q are also seen in nIPSCs (Figure 4f), consistent with our other statistical analyses showing limited effects of potential confounders. The very immature, fetal state of the nIPSCs[30] likely explains the absence of an enrichment signal for genes downregulated in postnatal ASD brain, which are enriched for genes involved in neurons with more mature synapses.

We next applied gene network analysis to construct an organizing framework to understand shared biological functions across idiopathic ASD and dup15q (combining the ASD Discovery Set, ASD Replication Set, and dup15q set). We utilized Weighted Gene Co-expression Network Analysis (WGCNA), which identifies groups of genes with shared expression patterns across samples (modules) from which shared biological function is inferred. Modules identified via WGCNA can than be related to a range of relevant phenotypes and potential confounders[31,32]. We applied signed co-expression analysis and used bootstrapping to ensure the network was robust, and not dependent on any subset of samples (Supplementary Information), while controlling for technical factors and RNA quality ("Adjusted FPKM" levels, Methods). WGCNA identified 16 co-expression modules (Extended Data Fig. 8a), which are further characterized by their association to ASD (Extended Data Fig. 8b), enrichment for cell-type specific genes (Extended Data Fig. 8c), and enrichment for GO terms (Extended Data Fig. 9). Of the downregulated modules, three are associated with ASD and dup15q (M1/10/17) and one with dup15q only (M11). Five of the upregulated modules are associated with ASD and dup15q (M4/5/6/9/12) and one is specific to dup15q (M13) (Figure 5a, top). Additionally, we identified a module strongly enriched for genes from the Attenuated Cortical Patterning set and *Wnt* signaling that contains *SOX5* (M12; fold enrichment = 3.0, $P = 3x10^{-8}$), verifying the strong relationship observed between the *Wnt* pathway regulating TF *SOX5* and attenuation of cortical patterning[33].

Notably, the modules identified here significantly overlap with previous patterns identified in ASD ($asdM12_{array}$ and $asdM16_{array}$[7]; Figure 5a, middle). We found that the ASD-associated modules identified by our larger sample size and RNA-seq provide significant refinement of previous observations by identifying more discrete biological processes related to cortical development[34], the post-synaptic density[35], and lncRNAs (Figure 5a, bottom). For example, M1 overlaps a subset of $asdM12_{array}$ (fold enrichment = 5.7) and developmental modules (devM16 fold enrichment = 3.7), and is enriched for proteins found in the PSD and genes involved in calcium signaling and gated ion channel signaling. Another subset of $asdM12_{array}$, M10 (fold enrichment = 11) overlaps more with a mid-fetal upregulated cortical development module (devM13 fold enrichment = 4.0), and genes involved in secretory pathways and intracellular signaling. A third module, M17 shows the least overlap with $asdM12_{array}$ (fold enrichment = 2.2) and is related to energy metabolism. Notably, these three modules are enriched for neuron-specific genes (Extended Data Fig. 8c), but not all neuronal modules are down regulated in ASD (M3 is not altered in ASD CTX). Taken together, specific neurobiological processes are affected in individuals with ASD related to developmentally regulated neurodevelopmental processes.

The most upregulated modules, M5 and M9, both strongly overlap (fold enrichments > 20) with previously identified upregulated co-expression module asdM16$_{array}$. M5 is enriched for microglial cell markers and immune response pathways, whereas M9 is enriched for astrocyte markers and immune-mediated signaling and immune cell activation (Extended Data Fig. 8c, Extended Data Fig. 9). This analysis clearly separates the contributions of the coordinated biological processes of microglial activation and reactive astrocytosis, which were previously not distinguishable as separate modules[7]. Thus, our analysis pinpoints more specific biological pathways in idiopathic ASD than those previously identified and reveals that similar changes occur downstream of the genetic perturbation in dup15q.

We evaluated the relationship between the five modules most strongly associated with ASD (M1/5/9/10/17, which are supported by module-trait association analysis and gene set enrichment analysis, Supplementary Information), and found that there was a remarkably high anti-correlation between the eigengene of M5 and downregulated modules, particularly M1 ($R^2 = 0.76$) (Figure 5b). M1 (Figure 5c) is downregulated in ASD and enriched for genes at the PSD and genes involved in synaptic transmission, while M5 (Figure 5d) is enriched for microglial genes and cytokine activation. This strong anti-correlation between microglial signaling and synaptic signaling in ASD and dup15q provides evidence in humans for dysregulation of microglia-mediated synaptic pruning, as previously suggested[36].

Next, to determine the role of causal genetic variation, we evaluated enrichment of both rare genetic variants, focusing on genes affected by ASD associated gene disrupting (LGD) *de novo* mutations[37], and common variants[38,39]. Genes within three modules, M1, M3, and M12, show enrichment for common variation signal for ASD (Figure 5e, Methods). Remarkably, M12 (Figure 5f), which is related to cortical patterning and Wnt signaling, also exhibit GWAS signal enrichment, providing the first evidence that risk conferred by common variation in ASD may affect regionalization of the cortex. Interestingly, M3 is significantly enriched for both schizophrenia (SCZ) and ASD common variants, is related to synaptic transmission, nervous system development, and regulation of ion channel activity (Extended Data Fig. 9), consistent with the notion that ASD and SCZ share common and rare genetic risk[1,40-43].

We only identified one module, M2 (Figure 5g), as significantly enriched in protein disrupting (nonsense, splice site, or frameshift) rare *de novo* variants previously associated with SCZ and ASD. M2 overlaps with a cortical developmental module implicated in ASD[34] (devM2 fold enrichment = 5.1). Notably, M2 is not differential between ASD and CTL in our dataset, consistent with the observation that these genes are primarily expressed during early neuronal development in fetal brain[34]. Remarkably, M2 contains an unusually large fraction of lncRNAs (15% of the genes in M2 are classified as lncRNAs, while other modules are 1-5% lncRNA). We hypothesize that, in addition to protein coding genes involved in transcriptional and chromatin regulation, rare *de novo* variants may also affect lncRNAs in ASD, a prediction that will be testable once large sets of whole genome sequences are available.

These combined transcriptomic and genetic analyses reveal that different forms of genetic variation affect biological processes involved in multiple stages of cortical development. Common genetic risk is enriched in M3, M1, and M12, which reflect early glutamatergic neurogenesis, later neuronal function, and cortical patterning, respectively. We also observe that rare *de novo* variation, which is enriched in M2, affects distinct biology related to transcriptional regulation and chromatin modification. These findings are consistent with transcriptomic analyses of early prenatal brain development and ASD risk mutations that implicate chromatin regulation and glutamatergic neuron development[34,44].

We provide the first comprehensive picture of largely unexplored aspects of transcription in ASD, lncRNA and alternative splicing, and identify a strong convergent signal in these, as well as protein coding genes[7]. These results will aid in interpreting genetic variation outside of the known exome, as whole genome sequencing supplants current methods. A role of lncRNAs has been previously explored in ASD[45], but only two individuals were evaluated with targeted microarrays. We evaluate lncRNAs in an unbiased manner across many individuals, notably identifying an enrichment of lncRNAs in M2, most of which are uncharacterized in brain and arose on the primate lineage. The involvement of lncRNAs in this early developmental program that is enriched for *de novo* mutations implicated in ASD suggests their study will be particularly relevant to understanding the emergence of primate higher cognition on the mammalian lineage, and by extension human brain evolution[15,46,47].

We also provide the first confirmation of an attenuation of genes that typically show differential expression between frontal and temporal lobe in ASD CTX and further identified *SOX5*, known to regulate cortical laminar development[50,51], as a putative regulator of this disruption. That M12, which is enriched for genes exhibiting cortical regionalization and is also enriched in ASD GWAS signal, supports the prediction that attenuation of patterning may be mediated by common genetic variation in or near the *SOX5* target genes. Disruption of cortical lamination by direct effects on glutamatergic neurogenesis and function has been predicted by independent data, including network analyses of rare ASD associated variants identified in exome sequencing studies[34,44].

These data, in conjunction with previous studies, reveal a consistent picture of the ASD's emerging postnatal and adult pathology. Specific neuronal signaling and synaptic molecules are downregulated and astrocyte and microglial genes are upregulated in over 2/3 of cases. Microglial infiltration has been observed in ASD cortex with independent methods[52], and normal microglial pruning has been shown to be necessary for brain development[36]. Our findings further suggest that aberrant microglial-neuronal interactions may be pervasive in ASD and related to the gene expression signature seen in a majority of individuals. In our comprehensive AS analysis, we identify three splicing factors upstream of the altered splicing signature observed in ASD CTX. These factors are known to be involved in coordinating sequential processes in neuronal development[17,21] and maintaining neuronal function[48,49]. It may therefore be sufficient to disrupt any one of these factors to induce a similar outcome during brain development, which would be consistent with the shared downstream perturbation observed here.

Finally, evaluation of the transcriptome in dup15q supports the enormous value of the "genotype first" approach of studying syndromic forms of ASD, with known penetrant genetic lesions[53]. It is highly unlikely that the shared transcriptional dysregulation in dup15q is due to a shared environmental insult. Thus, the most parsimonious explanation for the convergent transcriptomic pathology seen in all dup15q and over 2/3 of the cases of idiopathic ASD is that it represents an adaptive or maladaptive response to a primary genetic insult, which in most cases of ASD will be genetic[2,54]. As future investigations pursue the full range of causal genetic variation contributing to ASD risk, these analyses and data will be valuable for interpreting genetic and epigenetic studies of ASD as well as those of other neuropsychiatric disorders.

## Figures & Figure Legends

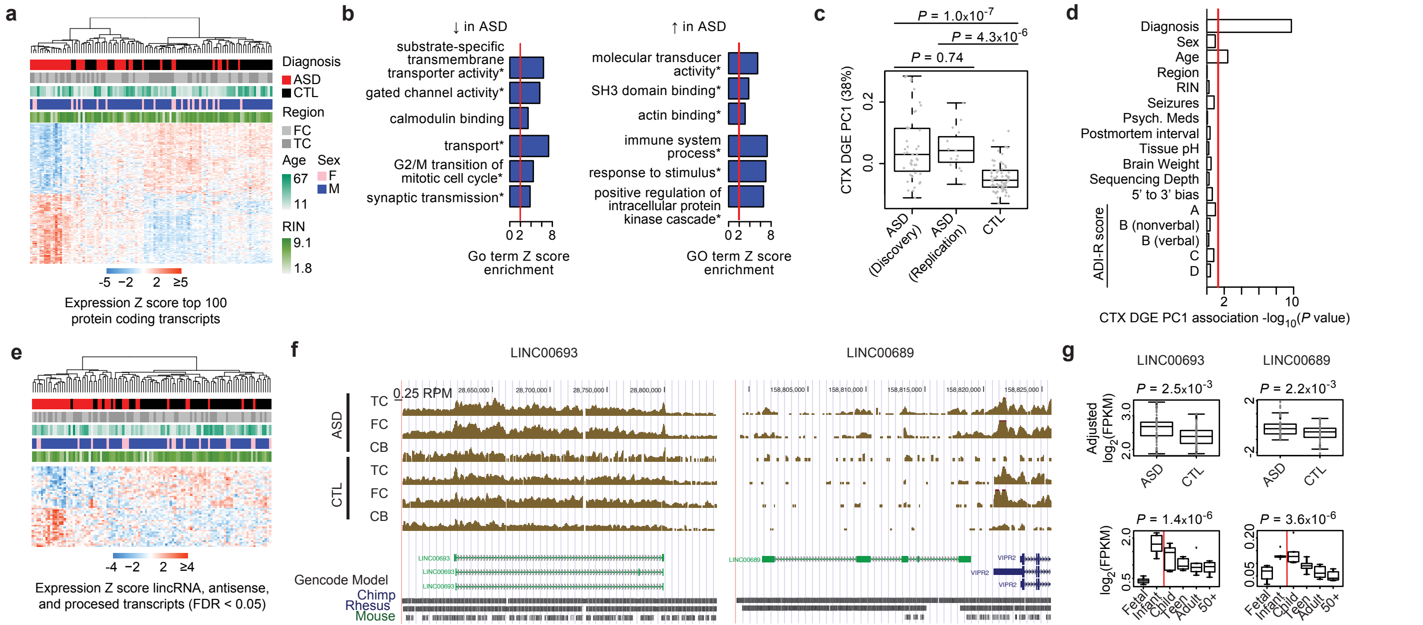


Figure 1 | Transcriptome-wide differential gene expression in ASD. a, Average linkage hierarchical clustering of samples in the ASD Discovery Set using the top 100 upregulated and top 100 downregulated protein coding genes. b, Gene Ontology (GO) term enrichment analysis of upregulated and downregulated genes in ASD. *FDR ≤ 0.05 across all GO terms and gene sets. c, 1st principal component of the CTX DGE set (CTX DGE PC1) is able to distinguish ASD and CTL samples, including independent samples from the ASD Replication Set. d, CTX DGE PC1 is primarily associated with diagnosis, and not other factors. e, Average linkage hierarchical clustering of ASD Discovery Set using all lncRNAs in the DGE set. f, UCSC genome browser track displaying reads per million (RPM) in a representative ASD and CTL sample, superimposed over the gene models and sequence conservation for genomic regions including *LINC00693* and *LINC00689*. g, *LINC00693* and *LINC00689* are upregulated across ASD samples and downregulated during frontal cortex development. Abbreviations: FC, frontal cortex; TC, temporal cortex; RIN, RNA integrity number; ADI-R score, Autism Diagnostic Interview Revised score; FPKM, fragments per kilobase million mapped reads.

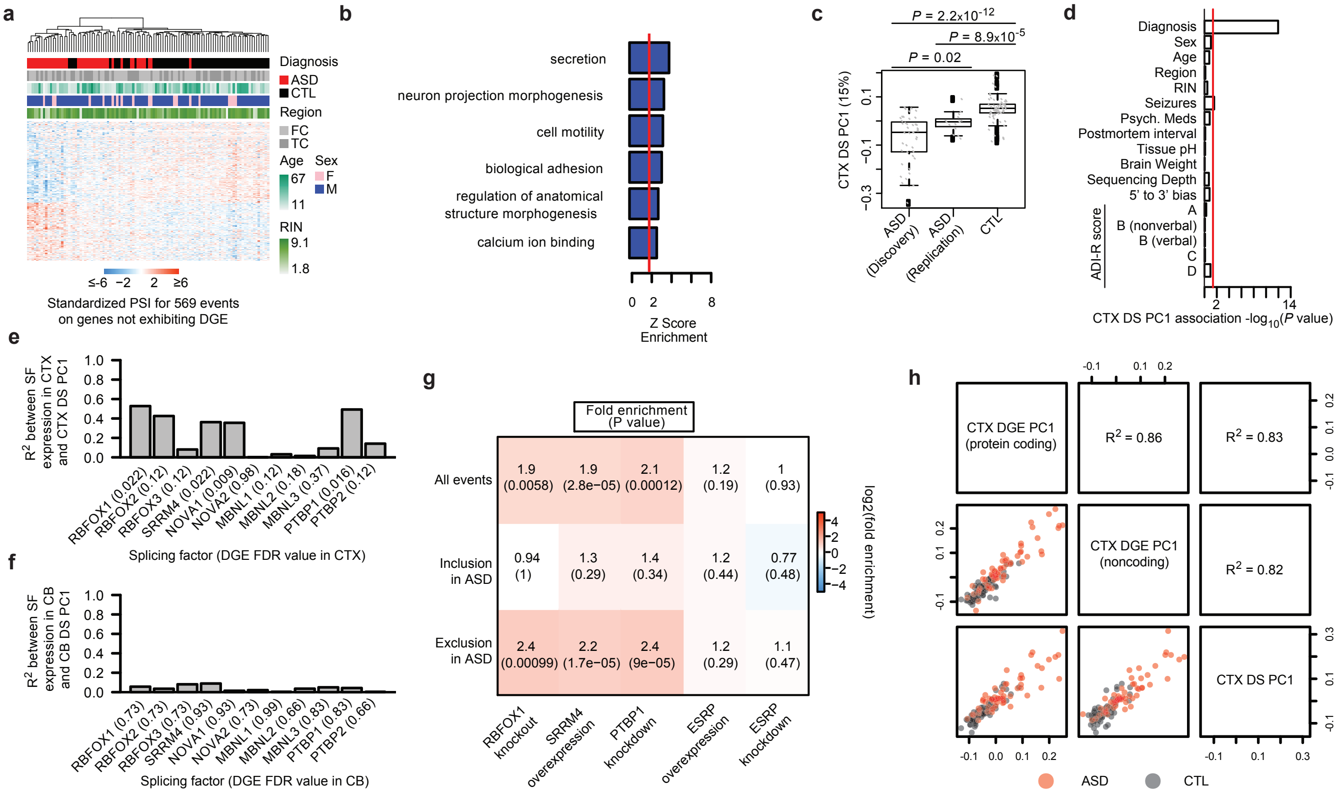


Figure 2 | Alteration of alternative splicing in ASD. a, Average linkage hierarchical clustering of ASD discovery set using top 100 differentially included and top 100 differentially excluded exons from the differential splicing (DS) set across the ASD Discovery Set. b, Gene Ontology term enrichment analysis of genes with DS in ASD. c, 1st principal component 1 of the CTX differential alternative splicing set (CTX DS PC1) is able to distinguish ASD and CTL samples using independent samples from the ASD Replication Set. d, CTX DS PC1 is primarily associated with diagnosis, and not other factors. e, Correlation between CTX DS PC1 and gene expression of neuronal splicing factors in CTX. f, Correlation between 1st principal component of cerebellum differential splicing (CB DS PC1) and gene expression of neuronal splicing factors in cerebellum. g, Overlap between DS set and splicing events regulated by splicing factors where experimental data was available. h, Scatterplots and correlations between the 1st principal component across the ASD versus CTL DGE sets for different transcriptome subcategories. Abbreviations: FC, frontal cortex; TC, temporal cortex; RIN, RNA integrity number; ADI-R score, Autism Diagnostic Interview Revised score; FPKM, fragments per kilobase million mapped reads.

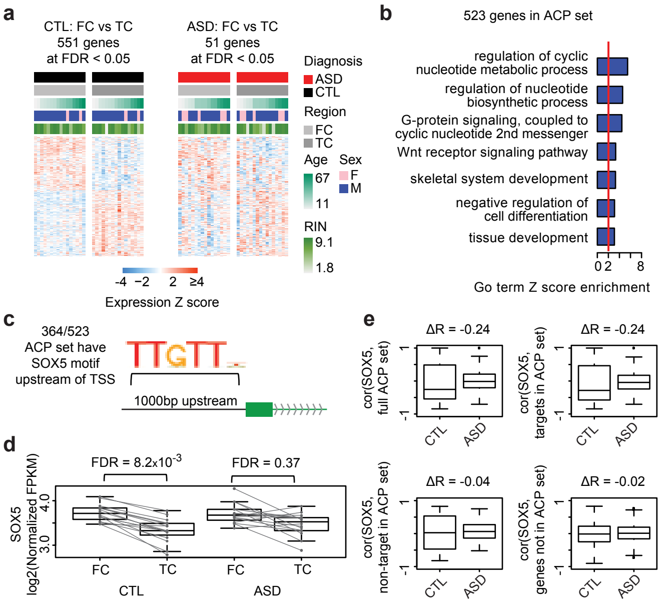


Figure 3 | Attenuation of cortical patterning in ASD cortex. a, Heatmap of 551 genes exhibiting cortical patterning between frontal cortex (FC) and temporal cortex (TC) in ASD, with samples sorted by diagnostic status and brain region. b, Gene ontology term enrichment analysis of genes exhibiting attenuated cortical patterning (ACP). c, Schematic of transcription factor motif enrichment upstream of genes in the ACP set, with the *SOX5* motif sequence logo. d, The *SOX5* gene exhibits attenuated cortical patterning in ASD CTX compared to CTLs. Lines connect FC-TC pairs that are from the same individual. e, Correlation between *SOX5* gene expression and predicted targets in CTL and ASD, with all ACP genes (top left), SOX5 targets from the ACP set (top right), SOX5 non-targets from the ACP set (bottom left), and all genes not in the ACP set (bottom right). Plots show the difference in correlation between *SOX5* and other genes in ASD and CTL (ΔR).

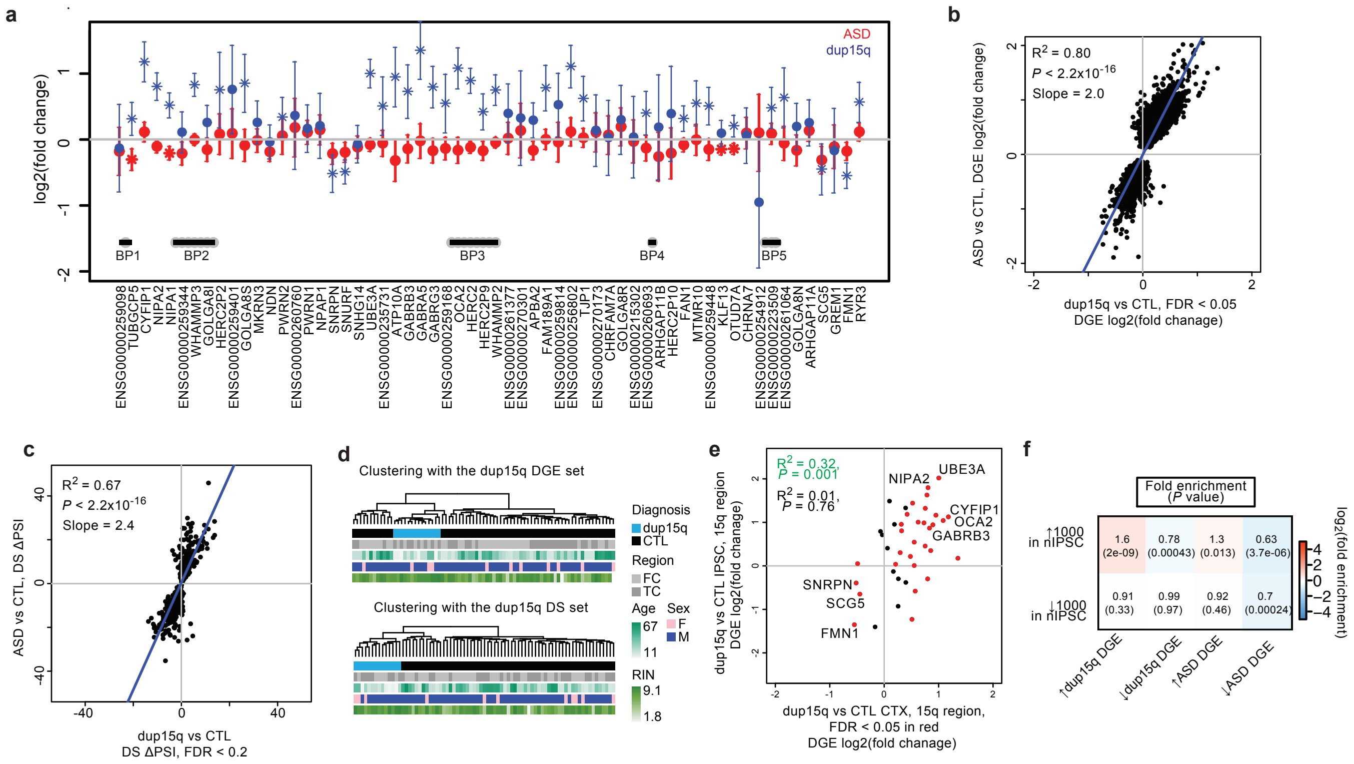


Figure 4 | Duplication 15q Syndrome recapitulates transcriptomic changes in idiopathic ASD. a, DGE changes across the 15q11-13.2 region for ASD and dup15q compared to CTL, error bars are +/- 95% confidence intervals for the fold changes. b, Comparison of effect sizes in dup15q vs CTL and ASD vs CTL, with changes in dup15q at FDR ≤ 0.05 highlighted. c, Comparison of differential splicing (DS) changes in dup15q vs CTL and ASD vs CTL, highlighting 402 events at FDR ≤ 0.2 in dup15q. d, Average linkage hierarchical clustering of dup15q samples and controls using the DGE and DS gene sets. e, Plot of fold changes between induced pluripotent stem cells differentiated into neurons (nIPSCs) from dup15q vs CTL and postmortem CTX DGE from dup15q vs CTL in the 15q region. f, Heatmap overlapping the top 1000 genes up- and down-regulated in the nIPSC comparison to the up- and down- regulated genes in dup15q and idiopathic ASD CTX.

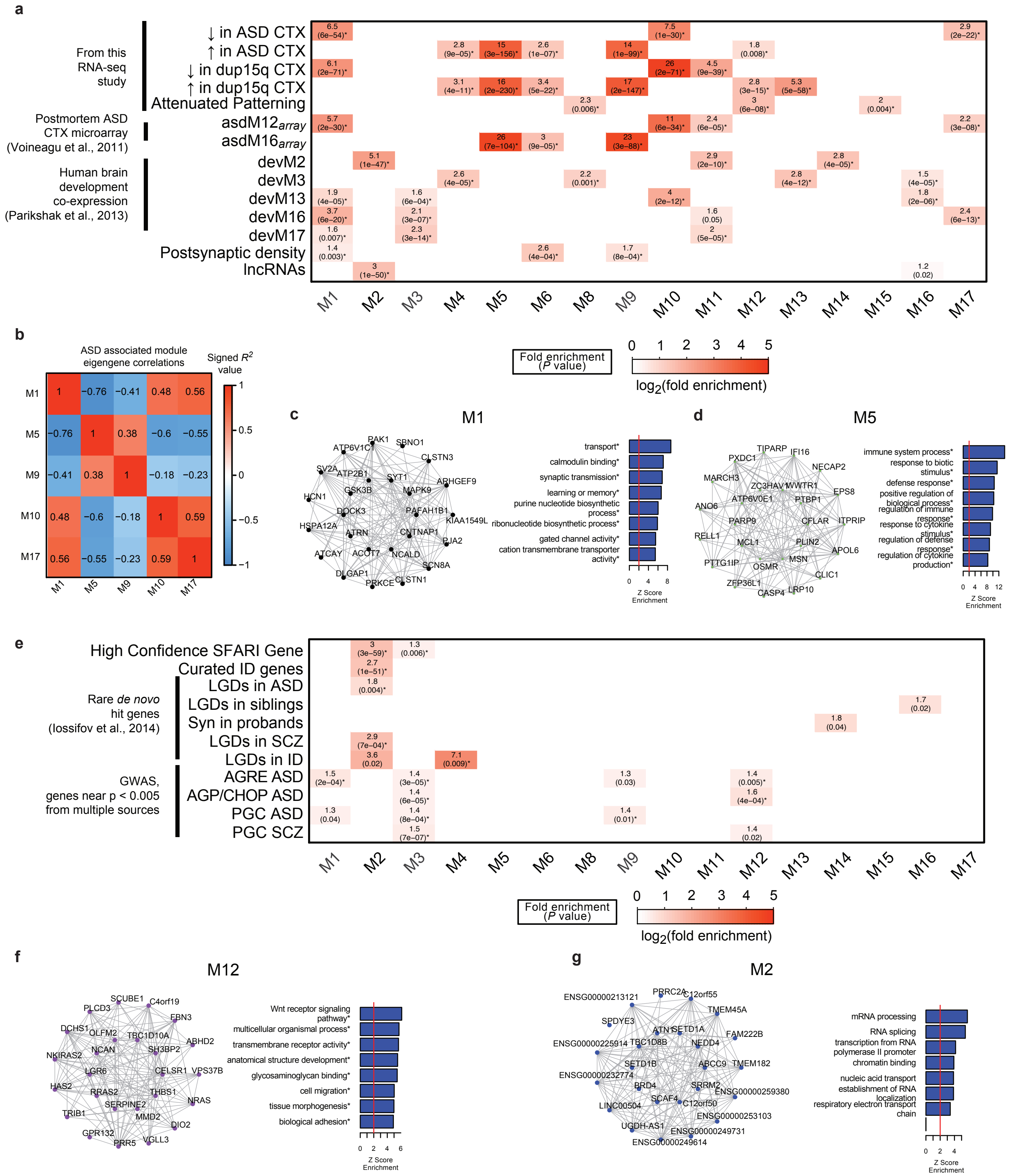


Figure 5 | Co-expression network analysis across all ASD and CTL samples in CTX. a, Gene set enrichment analyses comparing the 16 co-expression modules with multiple gene sets from this RNA-seq study, from postmortem ASD CTX microarray, from human brain development, from the postsynaptic density and set of all brain-expressed lncRNAs. b, Comparison of five ASD-associated modules against each other by correlating module eigengenes. c, Module plot of M1 displaying the top 25 hub genes along with the module's Gene Ontology term enrichment. d, similar to c, but for M5. e, Gene set enrichment analysis with genome-wide whole-exome sequencing data (Rare *de novo* hit genes) and genome-wide association study (GWAS) results in ASD, schizophrenia (SCZ), and intellectual disability (ID). Boxes are filled if the odds ratio is greater than 0, and the enrichment $P < 0.05$. Asterisks* indicate FDR ≤ 0.05 across all comparisons in a and e. f,g, similar to c, but for M12 and M2, respectively. Abbreviations: LGD, likely gene disrupting, genes affected by nonsense, nonsynonymous, or splice-site mutations or frame-shift indels; AGRE, AGP/CHOP, and PGC refer to consortia that collect genetic data (Supplementary Information for details).

## Methods

Sample description: Brain tissue for ASD and control individuals was acquired from the Autism Tissue Program (ATP) brain bank at the Harvard Brain and Tissue Bank and the University of Maryland Brain and Tissue Bank (a Brain and Tissue Repository of the NIH NeuroBioBank). Sample acquisition protocols were followed for each brain bank, and samples were de-identified prior to acquisition. Brain sample and individual level metadata is available in Supplementary Table 1.

RNA-seq methodology: Starting with 1ug of total RNA, samples were rRNA depleted (RiboZero Gold, Illumina) and libraries were prepared using the TruSeq v2 kit (Illumina) to construct unstranded libraries with a mean fragment size of 150bp (range 100-300bp) that underwent 50bp paired end sequencing on an Illumina HiSeq 2000 or 2500 machine. Paired-end reads were mapped to hg19 using Gencode v18 annotations[55] via Tophat2[56]. Gene expression levels were quantified using union exon models with HTSeq[57]. For additional and information on sequencing and read alignment parameters, please see Supplementary Information.

Sample sets for analysis: For differential gene expression and splicing analysis, we defined an age matched set, referred to as the ASD Discovery Set (106 samples in CTX, 51 in cerebellum) of idiopathic ASD and control samples for the discovery set, and held out younger or unmatched samples as the ASD Discovery Set (17 in CTX, 8 in cerebellum). Dup15q individuals were analysed separately, utilizing the full set of controls from the ASD Discovery Set. For co-expression network analysis, we combined the discovery set, replication set, and dup15q individuals for a total of 137 CTX samples and 59 cerebellum samples.

Differential Gene Expression (DGE): DGE analysis was performed with expression levels adjusted for gene length, library size, and G+C content (referred to as “Normalized FPKM”) Supplementary Information. CTX samples (frontal and temporal) were analyzed separately from cerebellum samples. A linear mixed effects model framework was used to assess differential expression in log2(Normalized FPKM) values for each gene for cortical regions (as multiple brain regions were available from the same individuals) and a linear model was used for cerebellum (where one brain region was available in each individual, with a handful of technical replicates removed). Individual brain ID was treated as a random effect, while age, sex, brain region (except in the case of cerebellum, where there is only one region), and diagnoses were treated as fixed effects. We also used technical covariates accounting for RNA quality, library preparation, and batch effects as fixed effects into this model (Supplementary Information).

Reproducibility analyses: We assessed replication between datasets by evaluating the concordance between independent sample sets by comparing the squared correlation ($R^2$) of fold changes of genes in each sample set at a non-stringent P value threshold. This general approach has been shown to be effective for identifying reproducible gene expression patterns[58], and we modify it such that the P value threshold is set in one sample set (the *x* axis in the scatterplots), and the $R^2$ with fold changes in these genes are evaluated in an independent sample set (the *y* axis in the scatterplots).

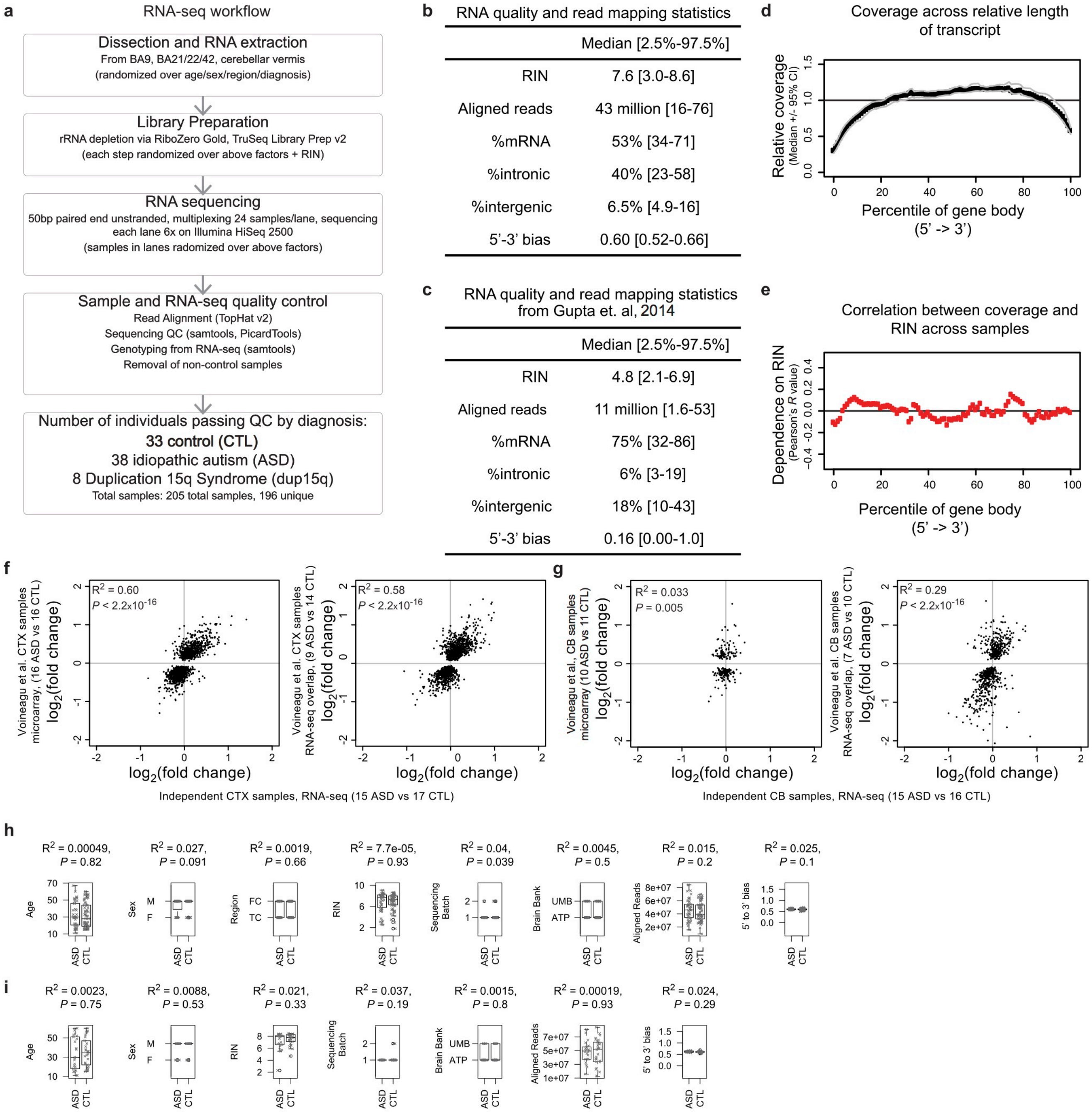
Differential Splicing Analysis: Alternative splicing was quantified using the percent spliced in (PSI) metric using Multivariate Analysis of Transcript Splicing (MATS, v3.08)[18]. For each event, MATS reports counts supporting the inclusion (I) or exclusion (E) of a splicing event. To reduce spurious events due to low counts, we required at least 80% of samples to have I + S >= 10. For these events, the percent spliced in is calculated as PSI = I / (I + S) (Extended Data Fig. 4a). Statistical analysis for differential splicing was performed utilizing the linear mixed effects model regression framework as described above for DGE. This approach is advantageous over existing methods as it allows modeling of covariates and takes into consideration the variability in PSI across samples when assessing event significance with ASD (Supplementary Information).

Genotyping dup15q: For Dup15q samples, the type of duplication and copy number in the breakpoint 2-3 region were available for these brains[59]. To expand this to the regions between each of the recurrent breakpoint in these samples, 7/8 dup15q brains were genotyped (one was not genotyped due to limitations in tissue availability). The number of copies between each of the breakpoints is reported in Extended Data Fig. 7a.

Co-expression network analysis: The R package weighted gene co-expression network analysis (WGCNA) was used to construct co-expression networks using the technical variation normalized data[31,60] (referred to as "Adjusted FPKM"). We used the biweight midcorrelation to assess correlations between log2(Normalized FPKM) and parameters for network analysis are described in Supplementary Information. Notably, we utilized a modified version of WGCNA that involves bootstrapping the underlying dataset 100 times and constructing 100 networks. The consensus of these networks (50th percentile across all edges) was then used as the final network [32], ensuring that a handful of samples do not determine the network structure. For module-trait analyses, 1st principal component of each module (eigengene) was related to ASD diagnosis, age, sex, and brain region in a linear mixed effects framework as above, only replacing the expression values of each gene with the eigengene.

Enrichment analysis of gene sets and GWAS: Enrichment analyses were performed either with Fisher's exact test (cell type and splicing factor enrichments) or logistic regression (all enrichment analyses in Figure 5). We used logistic regression in the latter case to control for gene length or other biases that may influence enrichment analysis (Supplementary Information). All GO term enrichment analysis was performed using GO Elite[61] with 10,000 permutations. We focused on molecular function and biological process terms for display purposes.

## Extended Data Figures & Legends



a RNA-seq workflow

Dissection and RNA extraction
From BA9, BA21/22/42, cerebellar vermis
(randomized over age/sex/region/diagnosis)

Library Preparation
rRNA depletion via RiboZero Gold, TruSeq Library Prep v2
(each step randomized over above factors + RIN)

RNA sequencing
50bp paired end unstranded, multiplexing 24 samples/lane, sequencing each lane 6x on Illumina HiSeq 2500
(samples in lanes radomized over above factors)

Sample and RNA-seq quality control
Read Alignment (TopHat v2)
Sequencing QC (samtools, PicardTools)
Genotyping from RNA-seq (samtools)
Removal of non-control samples

Number of individuals passing QC by diagnosis:
**33 control (CTL)**
38 idiopathic autism (ASD)
8 Duplication 15q Syndrome (dup15q)
Total samples: 205 total samples, 196 unique

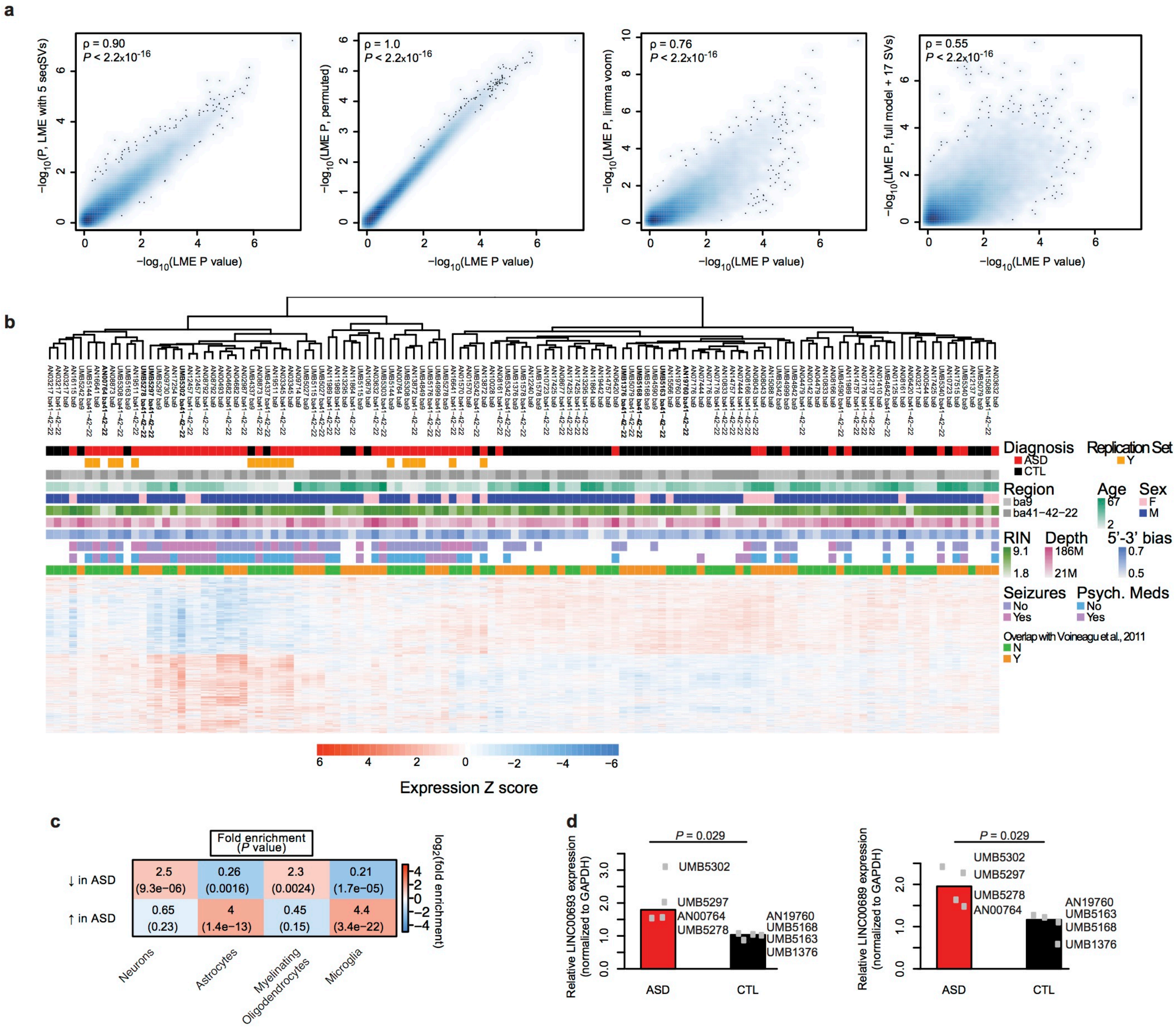
b RNA quality and read mapping statistics

| | Median [2.5%-97.5%] |
|---|---|
| RIN | 7.6 [3.0-8.6] |
| Aligned reads | 43 million [16-76] |
| %mRNA | 53% [34-71] |
| %intronic | 40% [23-58] |
| %intergenic | 6.5% [4.9-16] |
| 5'-3' bias | 0.60 [0.52-0.66] |

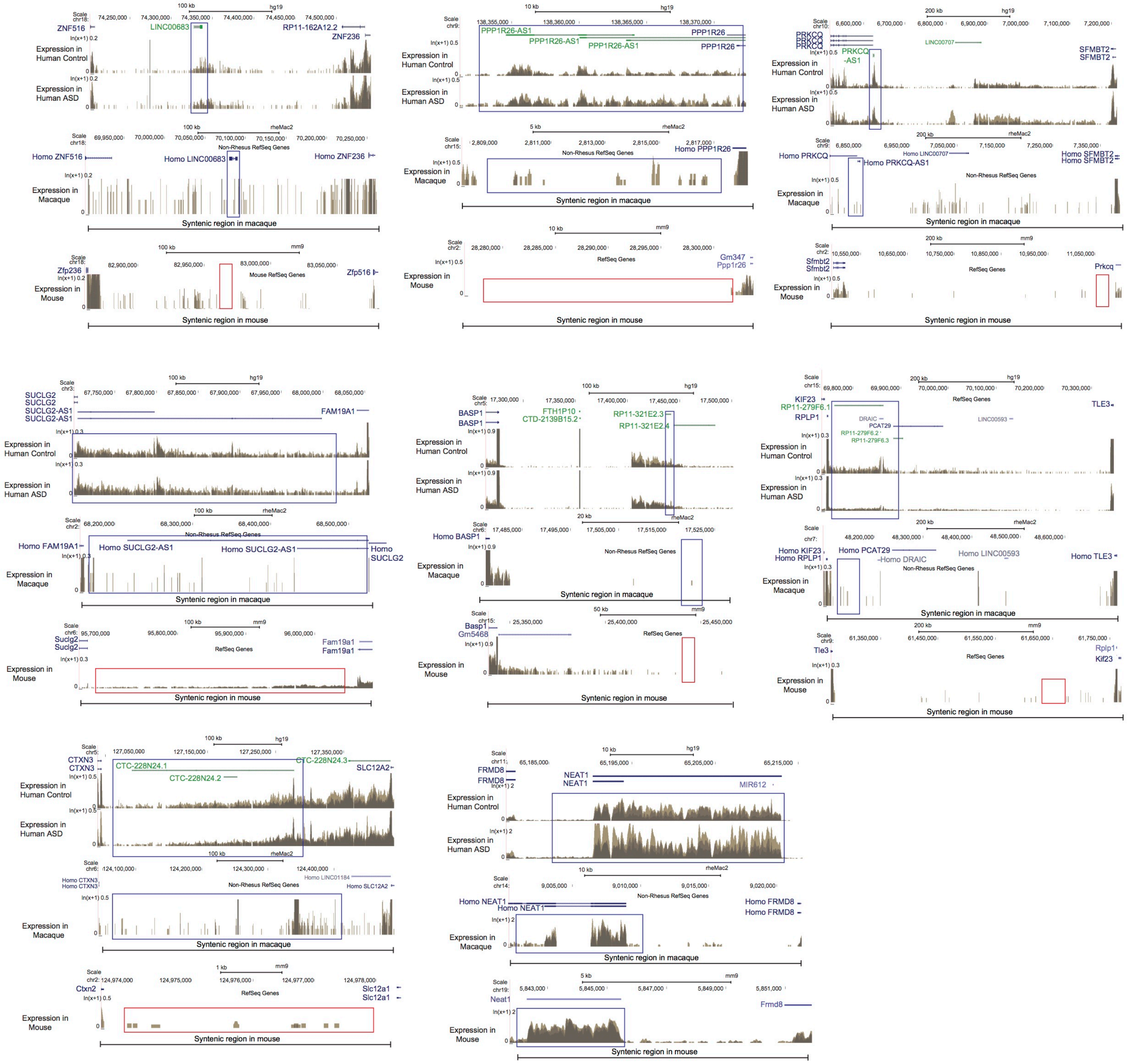
c RNA quality and read mapping statistics from Gupta et. al, 2014

| | Median [2.5%-97.5%] |
|---|---|
| RIN | 4.8 [2.1-6.9] |
| Aligned reads | 11 million [1.6-53] |
| %mRNA | 75% [32-86] |
| %intronic | 6% [3-19] |
| %intergenic | 18% [10-43] |
| 5'-3' bias | 0.16 [0.00-1.0] |

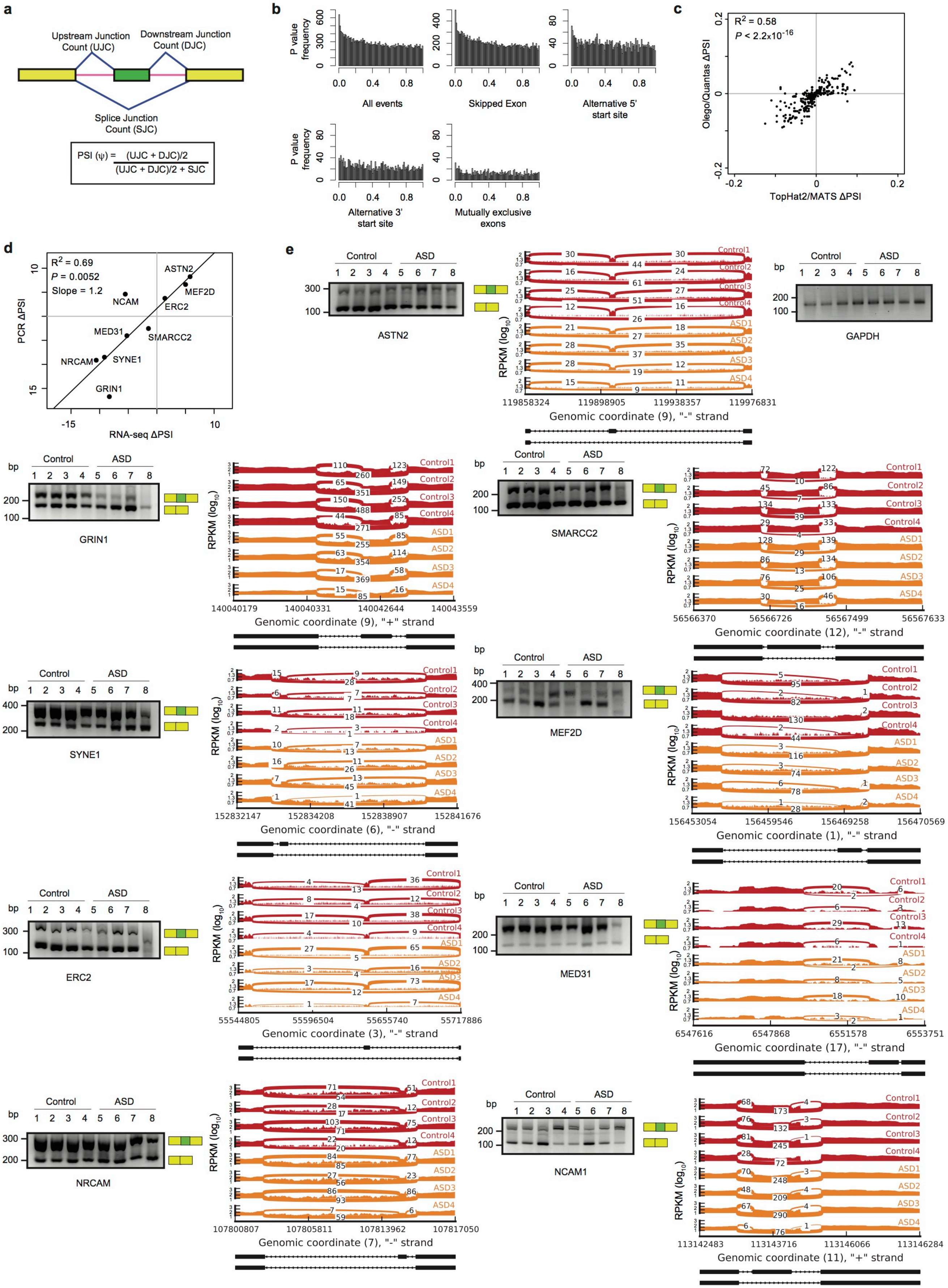
Extended Data Figure 1 | Methodology, quality control, and differential expression replication analysis. a, RNA-seq workflow, including RNA extraction, library preparation, sequencing, read alignment, and quality control. b, RNA-seq quality and alignment statistics from this study, including RNA integrity number (RIN), number of aligned reads, proportion of reads mapping to different genomic features (mRNA, intronic, intergenic), and bias in coverage from the 5' to the 3' end of the top 1000 expressed transcripts (statistics compiled using PicardTools). c, Similar statistics as in b for another RNA-seq study that utilized polyA tail selection mRNA-seq to evaluate the transcriptome in ASD cortex[11] (primarily BA19, visual cortex, but also including some BA10/44 samples, frontal cortex). d, RNA-seq read coverage relative to normalized gene length across transcripts from the 5' to the 3' end in this study. e, Dependence between coverage and RIN across gene body (correlation between RIN and coverage in d across samples). f, Correlation of ASD vs CTL fold changes between previously evaluated and new ASD samples in CTX by microarray (left) and RNA-seq (right) using genes that were at $P < 0.05$ the samples from Voineagu et al., 2011. g, Correlation between effect sizes as in f, but for cerebellum (CB) samples. h,i, Correlation between covariates and ASD vs CTL status in CTX (h) and CB (i) in the ASD Discovery Set.



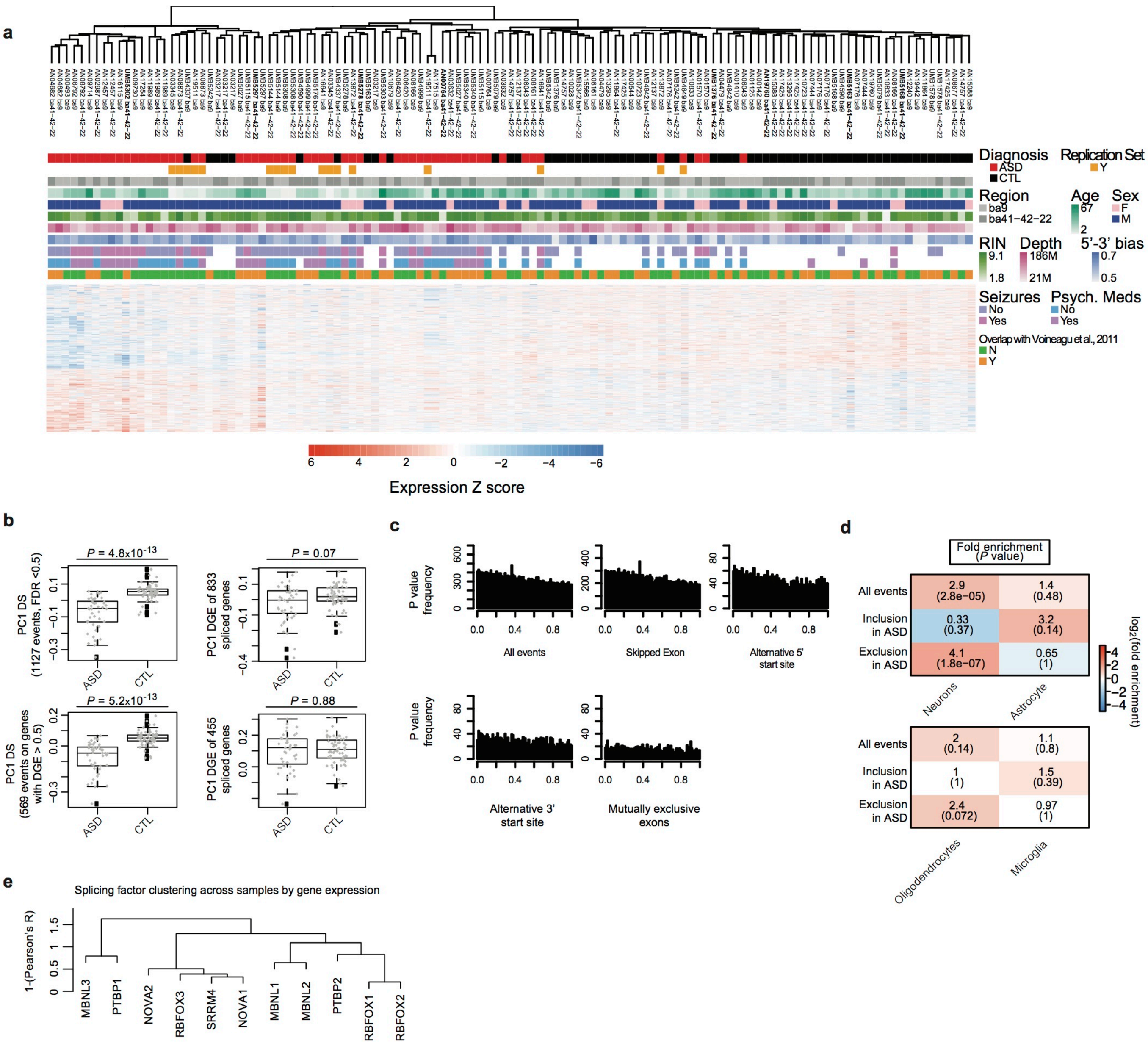
Extended Data Figure 2 | Transcriptome-wide differential gene expression (DGE) analysis in CTX. a, Comparison of P value rankings across different methods for DGE with Spearman's correlation. From left to right: removal of three additional principal components of sequencing statistics (Supplementary Information) related to RNA-sequencing quality, application of a permutation analysis for DGE P value computation, application of variance-weighted linear regression for DGE[62], and using surrogate variable analysis for DGE[63]. b, Average linkage hierarchical clustering heatmap using all genes DGE in the ASD Discovery Set, but including all idiopathic ASD frontal cortex (FC) and temporal cortex (TC) samples across 123 samples, combining the ASD Discovery set and the ASD Replication set. Bolded samples in the dendrogram are used for validation in d. c, Enrichment analysis of cell-type specific gene sets (5-fold enriched in the cell type compared to all other cells) with genes decreased and increased in ASD. d, RT-PCR validation of the two lincRNAs shown in Figure 1f-g, P values are computed with the Wilcoxon rank-sum test.



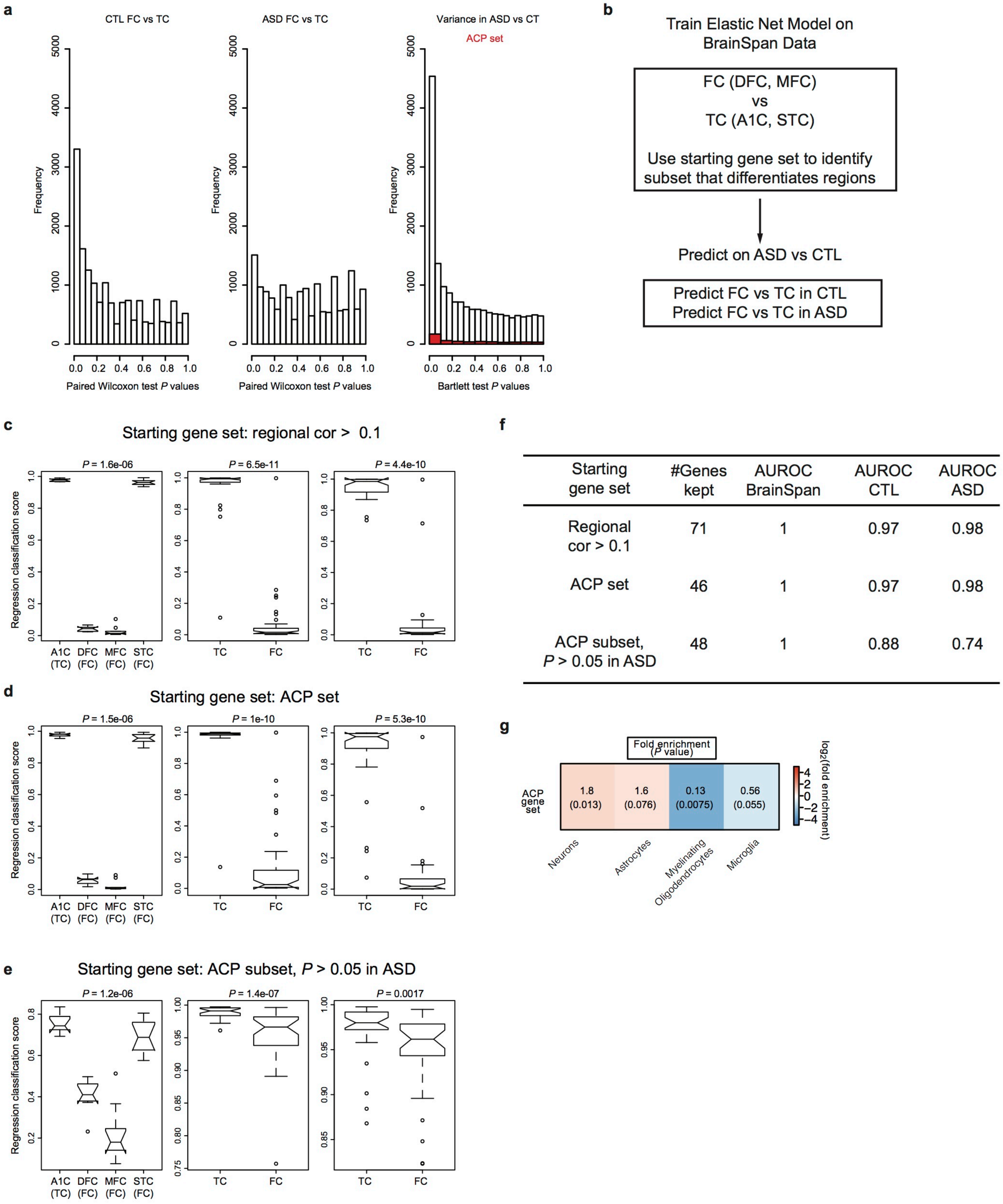
Extended Data Figure 3 | Gene browser tracks for selected primate-specific lncRNAs. For each lncRNA, expression for representative samples for ASD vs CTL (top) in human, macaque (middle), and mouse (bottom) are shown. The genome location for macaque and mouse displayed is syntenic to the human region, with the expected location of the lncRNA highlighted.


a
Upstream Junction Count (UJC)
Downstream Junction Count (DJC)
Splice Junction Count (SJC)
PSI (ψ) = (UJC + DJC)/2 / ((UJC + DJC)/2 + SJC)
b
P value frequency
All events
Skipped Exon
Alternative 5' start site
Alternative 3' start site
Mutually exclusive exons
c
R² = 0.58
P < 2.2x10⁻¹⁶
Olego/Quantas ΔPSI
TopHat2/MATS ΔPSI
d
R² = 0.69
P = 0.0052
Slope = 1.2
PCR ΔPSI
RNA-seq ΔPSI
ASTN2
MEF2D
ERC2
NCAM
MED31
SMARCC2
SYNE1
NRCAM
GRIN1
e
Control
ASD
bp
ASTN2
GAPDH
GRIN1
SMARCC2
SYNE1
MEF2D
ERC2
MED31
NRCAM
NCAM1
RPKM (log₁₀)
Control1
Control2
Control3
Control4
ASD1
ASD2
ASD3
ASD4
Genomic coordinate (9), "-" strand
Genomic coordinate (9), "+" strand
Genomic coordinate (12), "-" strand
Genomic coordinate (6), "-" strand
Genomic coordinate (1), "-" strand
Genomic coordinate (3), "-" strand
Genomic coordinate (17), "-" strand
Genomic coordinate (7), "-" strand
Genomic coordinate (11), "+" strand

Extended Data Figure 4 | Splicing analyses and validation in ASD. a, Schematic describing how the percent spliced in (PSI) metric is computed. b, Distribution of *P* values for changes in the PSI between ASD and CTL in CTX for all events (left) and event subtypes (SE, spiced exon; A5SS, alternative 5' splice site; A3SS, alternative 3' splice site; MXE, mutually exclusive exons). c, Comparison of the CTX splicing analyses in when using PSI values obtained via read alignment by TopHat2[64] followed by the MATS[18] pipeline (used throughout this study) against read alignment by OLego followed by Quantas[65]. d, Comparison of ΔPSI values in nine splicing events between PCR and RNA-seq. e, PCR validation and sashimi plots for the nine splicing events delineated in d, from the samples highlighted in Extended Data Fig. 5a.



Extended Data Figure 5 | Additional splicing analyses in ASD. a, Average linkage hierarchical clustering heatmap using all differentially spiced (DS) events from the ASD Discovery Set, but including all idiopathic ASD neocortical samples (FC and TC) across 123 samples, combining the ASD Discovery set and the ASD Replication set. Bolded samples in the dendrogram were used for PCR validation in Extended Data Fig. 4. b, Top: difference between ASD and CTL in the DS set based on PC1 of the DS set at the PSI level, and PC1 of the gene expression levels of genes in the DS set. Bottom: Same comparison after differentially expressed genes (p < 0.05) are removed. c, Distribution of P values for changes in the PSI between ASD and CTL in cerebellum. d, Cell-type enrichment analysis of splicing events from CTX. e, Average-linkage hierarchical clustering using 1-(Pearson's correlation) to compare the gene expression patterns of the splicing factors investigated in Figure 2.



| Starting gene set | #Genes kept | AUROC BrainSpan | AUROC CTL | AUROC ASD |
|---|---|---|---|---|
| Regional cor > 0.1 | 71 | 1 | 0.97 | 0.98 |
| ACP set | 46 | 1 | 0.97 | 0.98 |
| ACP subset, *P* > 0.05 in ASD | 48 | 1 | 0.88 | 0.74 |

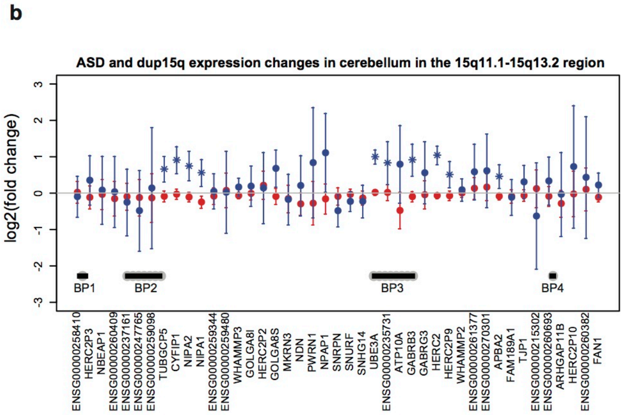
Extended Data Figure 6 | Attenuation of cortical patterning in ASD. a, Histograms of P values from paired Wilcoxon rank-sum test differential gene expression between 16 frontal cortex (FC) and 16 temporal cortex (TC) in CTL and ASD and a histogram of Bartlett's test P values for differences in gene expression variance between ASD and CTL for all genes (white) and genes in the Attenuated Cortical Patterning (ACP) set (red). c, Approach to training the elastic net model on BrainSpan and application of the model on 123 cortical samples in this study. c-e, Results of learned cortical region classifications with different starting gene sets, with the BrainSpan training set (left), CTL samples (middle), and ASD samples (right) in each panel and the Wilcoxon rank-sum test P value of FC vs TC difference for each comparison. f, Summary of results form c-e. g, Cell type enrichment analysis for genes in the ACP set. Abbreviations: A1C, primary auditory cortex; DFC, dorsolateral prefrontal cortex; MFC, medial prefrontal cortex; STC, superior temporal cortex; FC, frontal cortex; TC, temporal cortex; AUROC, area under the receiver-operator characteristic curve.

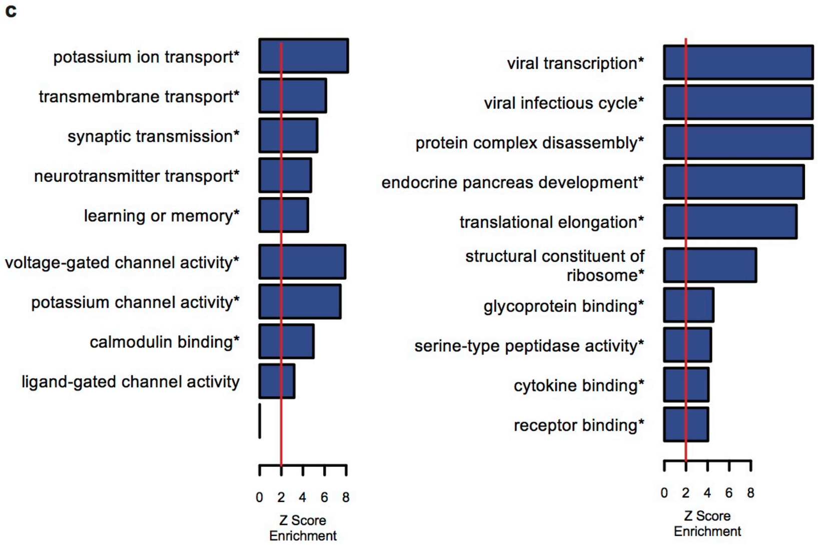
a

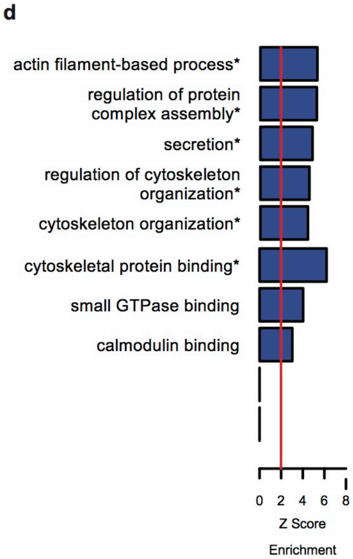
Duplication 15q breakpoints across individuals

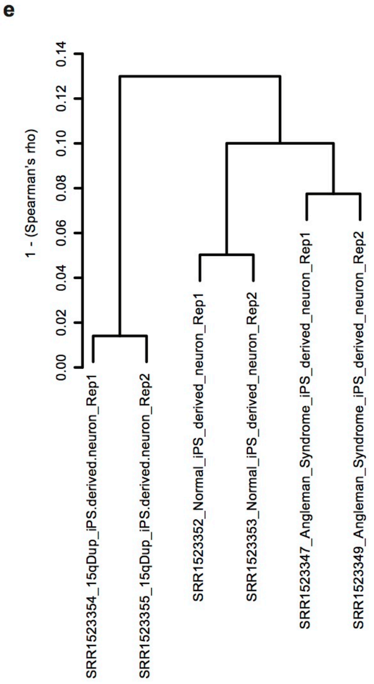
| Sample | BP1-2 | BP2-3 | BP3-4 | BP4-5 |
|---|---|---|---|---|
| AN09402 | 4 | 4,b | 2 | 2 |
| AN14829 | 4 | 4 | 4 | 3 |
| AN17138 | 4 | 4 | 2 | 2 |
| AN03935 | 4 | 4 | 4 | 3 |
| AN05983 | 4 | 4 | 4 | 3 |
| AN06365 | 4 | 4 | 4 | 3 |
| AN11931 | 4 | 4 | 4 | 3 |
| AN14762 | - | 4,a | - | - |

a,Obtained from Scoles et al., 2011 who evaluated duplication in this region by RT-PCR of SNRPN/GABRB3/UBE3A vs B2M
b,Discrepancy with Scoles et al., who report 5 here

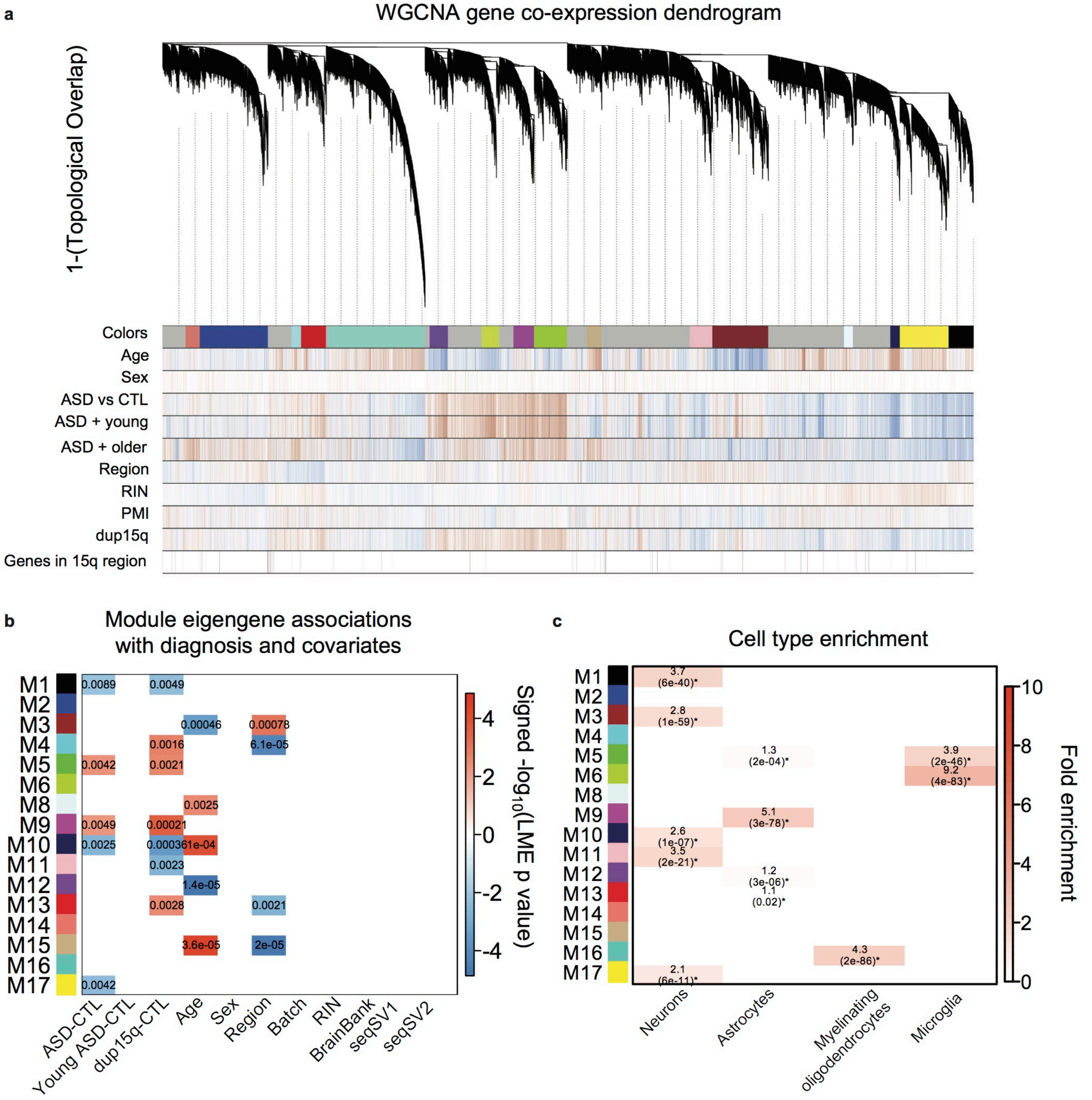






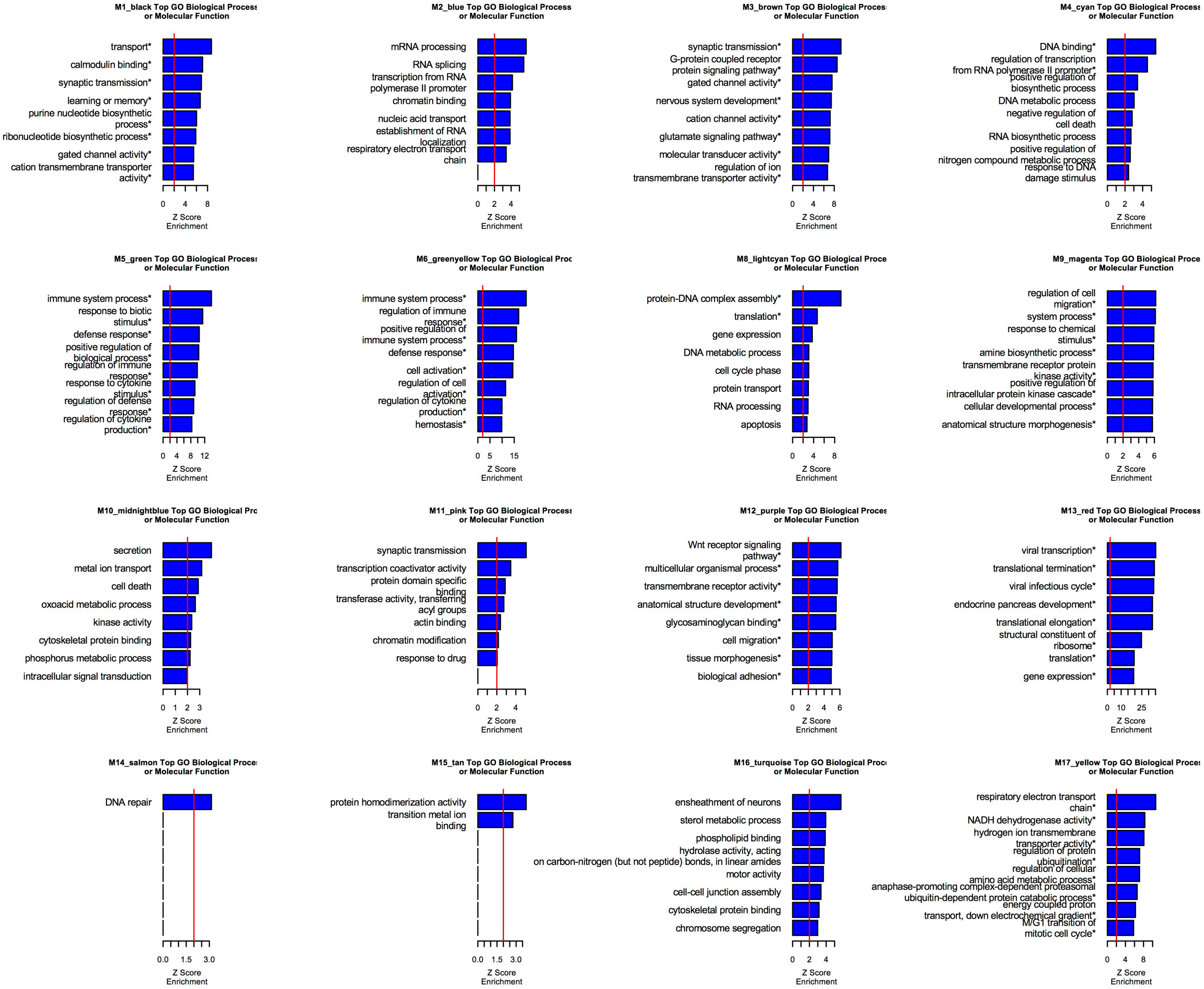


Extended Data Figure 7 | Dup15q syndrome analyses. a, Copy number between breakpoints (BP) in the 15q region. Genome-wide CNV analysis allowed evaluation of copy number in additional regions from previous studies[59,66]. b, Differential expression across the 15q region of interest in dup15q vs CTL and ASD vs CTL cerebellum, note only 3 samples were available for dup15q cerebellum so additional analyses were not pursued. c, Gene Ontology term enrichment analysis for the dup15q CTX differential expression set. d, Gene Ontology term enrichment analysis for the dup15q CTX differential splicing (DS) set. e, Hierarchical clustering of iPSC-derived neurons from dup15q, Angelman syndrome, and a control[29].



Extended Data Figure 8 | Co-expression network analysis in ASD CTX. a, Modules identified from a dendrogram constructed from a consensus of 100 bootstrapped datasets using the 137 CTX samples. Correlations for each gene to each measured factor are delineated below the dendrogram (blue = negative, red

= positive correlation). b, Module-trait associations as computed by a linear mixed effects model with all factors on the x-axis used as covariates. All P values are displayed where the coefficient passed p < 0.01. Note that this alternative approach to module-trait association agrees with the Fisher's exact test used in Figure 5a when the fold enrichment for module overlap with DGE sets is > 2.8, and we use an intersection of both methods for the modules focused on in Figure 5b. c, Module enrichments for cell type specific gene expression patterns.



Extended Data Figure 9 | GO term enrichments for all modules. *FDR < 0.05 across all GO enrichments across all modules.

References

1. Geschwind, D. H. Genetics of autism spectrum disorders. *Trends Cogn. Sci. (Regul. Ed.)* 15, 409–416 (2011).
2. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* 46, 881–885 (2014).
3. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* (2015). doi:10.1016/S1474-4422(15)00044-7
4. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* 17, 782–790 (2014).
5. Chen, J. A., Peñagarikano, O., Belgard, T. G., Swarup, V. & Geschwind, D. H. The emerging picture of autism spectrum disorder: genetics and pathology. *Annu Rev Pathol* 10, 111–144 (2015).
6. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 9, 341–355 (2008).
7. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384 (2011).
8. Purcell, A. E., Jeon, O. H., Zimmerman, A. W., Blue, M. E. & Pevsner, J. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* 57, 1618–1628 (2001).
9. Garbett, K. *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiology of Disease* 30, 303–311 (2008).
10. Chow, M. L. *et al.* Age-Dependent Brain Gene Expression and Copy Number Anomalies in Autism Suggest Distinct Pathological Processes at Young Versus Mature Ages. *PLoS Genet.* 8, e1002592 (2012).
11. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Comms* 5, 5748 (2014).
12. Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Publishing Group* 18, 154–161 (2015).
13. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).
14. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* 47, 199–208 (2015).
15. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640 (2014).
16. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Reports* 6, 1139–1152 (2014).
17. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159, 1511–1523 (2014).
18. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* 40, e61–e61 (2012).
19. Lionel, A. C. *et al.* Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. *Human Molecular Genetics* 23, 2752–2768 (2014).
20. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* 20, 1434–1442 (2013).
21. Raj, B. *et al.* A Global Regulatory Mechanism for Activating an Exon Network Required for Neurogenesis. *Molecular Cell* 56, 90–103 (2014).
22. Gueroussov, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* 349, 868–873 (2015).
23. Dittmar, K. A. *et al.* Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Molecular and Cellular Biology* 32, 1468–1482 (2012).
24. Tibshirani, R., Johnstone, I., Hastie, T. & Efron, B. Least angle regression. *The Annals of Statistics* 32,

407–499 (2004).

25. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* 41, D996–D1008 (2013).
26. Lui, J. H. *et al.* Radial glia require PDGFD–PDGFRβ signalling in human but not mouse neocortex. *Nature* 515, 264–268 (2014).
27. Lai, T. *et al.* SOX5 Controls the Sequential Generation of Distinct Corticofugal Neuron Subtypes. *Neuron* 57, 232–247 (2008).
28. Kwan, K. Y. *et al.* SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16021–16026 (2008).
29. Germain, N. D. *et al.* Gene expression analysis of human induced pluripotent stem cell-derived neurons carrying copy number variants of chromosome 15q11-q13.1. *Mol Autism* 5, 44 (2014).
30. Stein, J. L. *et al.* A Quantitative Framework to Evaluate Modeling of Cortical Development by Neural Stem Cells. *Neuron* 83, 69–86 (2014).
31. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* 4, Article17 (2005).
32. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 1, 54 (2007).
33. Morales, P. L. M., Quiroga, A. C., Barbas, J. A. & Morales, A. V. SOX5 controls cell cycle progression in neural progenitors by interfering with the WNT–β-catenin pathway. *EMBO reports* 11, 466–472 (2010).
34. Parikshak, N. N. *et al.* Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* 155, 1008–1021 (2013).
35. Bayés, À. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* 14, 19–21 (2010).
36. Schafer, D. P. *et al.* Microglia Sculpt Postnatal Neural Circuits in an Activity and Complement-Dependent Manner. *Neuron* 74, 691–705 (2012).
37. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221 (2014).
38. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Human Molecular Genetics* 21, 4781–4792 (2012).
39. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. 459, 528–533 (2009).
40. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* 45, 984–994 (2013).
41. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506, 185–190 (2014).
42. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res* 25, 142–154 (2015).
43. Gilman, S. R. *et al.* Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nat. Neurosci.* 15, 1723–1728 (2012).
44. Willsey, A. J. *et al.* Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell* 155, 997–1007 (2013).
45. Ziats, M. N. & Rennert, O. M. Aberrant Expression of Long Noncoding RNAs in Autistic Brain. *J Mol Neurosci* 49, 589–593 (2012).
46. Geschwind, D. H. & Rakic, P. Cortical Evolution: Judge the Brain by Its Cover. *Neuron* 80, 633–647 (2013).
47. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated Recruitment of New Brain Development Genes into the Human Genome. *PLoS Biol* 9, e1001179 (2011).
48. Fogel, B. L. *et al.* RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Human Molecular Genetics* 21, 4171–4186 (2012).

49. Gehman, L. T. *et al.* The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nat Genet* 43, 706–711 (2011).
50. Greig, L. C., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat Rev Neurosci* 14, 755–769 (2013).
51. Srinivasan, K. *et al.* A network of genetic repression and derepression specifies projection fates in the developing neocortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19071–19078 (2012).
52. Morgan, J. T. *et al.* Abnormal microglial–neuronal spatial organization in the dorsolateral prefrontal cortex in autism. *Brain Research* 1456, 72–81 (2012).
53. Stessman, H. A., Bernier, R. & Eichler, E. E. A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell* 156, 872–877 (2014).
54. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet* 15, 133–141 (2014).
55. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7, S4–9 (2006).
56. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31, 46–53 (2012).
57. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).
58. Shi, L. *et al.* The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics* 9, S10 (2008).
59. Scoles, H. A., Urraca, N., Chadwick, S. W., Reiter, L. T. & LaSalle, J. M. Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. *Mol Autism* 2, 19 (2011).
60. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559 (2008).
61. Zambon, A. C. *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* 28, 2209–2210 (2012).
62. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15, R29 (2014).
63. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* 3, e161 (2007).
64. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578 (2012).
65. Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res* 41, 5149–5163 (2013).
66. Wintle, R. F. *et al.* A genotype resource for postmortem brain samples from the Autism Tissue Program. *Autism Res* 4, 89–97 (2011).

End Notes

None.

Author Contributions

NNP, VS, and TGB performed dissections and RNA-seq analyses and differential gene expression analysis. NNP and VS performed splicing and co-expression network analysis. NNP and TGB performed analyses with Duplication 15q Syndrome individuals. NNP and MG reviewed clinical information and performed meta-analysis of ASD gene expression studies. VL and JKL performed genotyping and CNV analysis on dup15q samples. VS performed validation experiments for gene and splicing level alterations in ASD. MI and BJB assisted with splicing analyses.  RJ performed dissections. SH provided guidance on differential gene expression and co-expression analyses. DHG provided guidance on all experiments and analyses. NNP and DHG wrote the manuscript. All authors contributed to revising and finalizing the manuscript.

## Title

Genome-wide chromosomal conformation elucidates regulatory relationships in human brain development

## Authors and affiliations

Hyejung Won[1], Luis de la Torre-Ubieta[1], Jason L. Stein[1], Neelroop N. Parikshak[1], Farhad Hormozdiari[3], Changhoon Lee[1], Eleazar Eskin[3,4], Jason Ernst[2,4], Daniel H. Geschwind[1,4*]

[1] Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles

[2] Department of Biological Chemistry, David Geffen School of Medicine, University of California Los Angeles

[3] Department of Computer Science, University of California Los Angeles

[4] Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles

[5]Department of Molecular, Cell and Developmental Biology, University of California Los Angeles, Los Angeles

[*] Correspondence: dhg@mednet.ucla.edu

## Introduction

The demonstration that chromatin exhibits a complex 3 dimensional organization, whereby short and long distance physical interactions correspond to complex gene regulatory processes has opened a new window on understanding the functional organization of the human genome[1-4]. Recently, chromatin remodeling has also been causally implicated in several neurodevelopmental disorders, including autism and schizophrenia[5-7]. However, it remains unclear whether knowledge of chromosome organization in a tissue specific manner might inform our understanding of gene regulation in brain development or disease. Here we determined the genome-wide landscape of chromosome conformation during early human cortical development by performing Hi-C analysis in the mitotically active and post mitotic laminae of human fetal brain. We integrate Hi-C data with transcriptomic and epigenomic data and utilize chromosome contact information to delineate physical gene-gene regulatory interactions for non-coding regulatory elements. We show how these data permit large-scale functional annotation of non-coding variants identified in schizophrenia GWAS and of human specific enhancers[8,9]. These data provide a rubric that illustrates the power of tissue-specific annotation of non-coding regulatory elements, as well as novel insights into the pathogenic mechanisms of neurodevelopmental disorders and the evolution of higher cognition.

Recent advances in high-throughput sequencing have unveiled the epigenomic landscape of multiple human cell types, as well as 3 dimensional folding principles of chromatin[10,11]. In particular, chromosome conformation capture experiments demonstrate that chromatin is organized into hierarchical structures, which include compartments (a few megabase (Mb))[1], topological associating domains (TADs, sub-Mb)[12], and loops (ranging from few kilobase (kb) to few hundred kb)[2,4]. These structures are thought to play a role in gene regulation and biological function by defining functional genomic units and mediating the effects of *cis*-regulatory elements via both short- and long-range physical interactions (e.g. promotor-enhancer interactions), relationships that cannot simply be predicted by linear adjacency in chromosomes. Coupled with epigenomic data, such higher order chromatin interactions should facilitate systemic annotation of *cis*-regulatory elements, as well as intergenic and intronic variants, which will further expand our understanding of tissue specific developmental programs, as well as disease pathogenesis.

We constructed multiple Hi-C libraries in mid gestation fetal cerebral cortex from three individuals during the peak of neurogenesis and migration (gestation week, GW17-18). We reasoned that it would be useful to analyze mitotically active neuronal precursors involved in neurogenesis separately from post-mitotic migrating and maturing neurons, so we dissected the cortical anlage into two major structures: the cortical and subcortical plate (CP), consisting primarily of post mitotic neurons and the germinal zone (GZ), containing primarily mitotically active neural progenitors (representative heatmap in **Fig. 1a**, **Extended Data Fig. 1a-b**). For comparison with non-neuronal cell types, we also used publicly available Hi-C data on human embryonic stem (ES) cells and IMR90 cells[11,12]. To provide grounding for our data and compare global chromosome architecture between different cell types, we performed principal component analysis (PCA)[13] on the genome-wide inter-chromosomal contact matrices of CP, GZ, ES, and IMR90. As previously demonstrated, global chromosome architecture does not change dramatically between different cell types[13]. However, the first principal components (PC1s) from neuronal tissues (CP and GZ) have significantly higher correlation than the PC1s between different cell types (**Fig. 1b**), consistent with the higher similarity between tissues from brain, versus the two other cell lines.

**3D chromatin structure reflects gene regulation during neural differentiation.**

Previous studies have shown that genome-wide chromosome conformation captures multiple levels of genomic features related to biological function, ranging from GC content and gene number to marks of open chromatin, such as DNase I hypersensitivity sites (DHS)[13]. Most human-relevant Hi-C has been conducted in cell lines[1,2,4,11,12,14] and not in complex tissue, such as developing brain. As an initial first step to insure the quality and validity of our data, we analyzed the relationships between the major component of the inter-chromosomal interaction matrix with these major genomic features, finding high correlation with GC content, gene number, DHS[10], and to a lesser extent, gene expression[15] (**Fig. 1c**, **Extended Data Fig. 2a**), as has been previously observed in non-neural cell lines[13].

To further explore the biological significance of chromosome contact changes during neural differentiation, we explored whether the genes in regions of dynamic chromatin structure were related to neural differentiation by comparing the inter-chromosome contact matrices (binned to 100kb) in different cell types and selecting bins with the highest chromatin contact count changes between two cell types (**Methods**). Genes located in the regions of highest inter-chromosomal interaction changes between CP and GZ were enriched for neuronal genes, represented by the gene ontology (GO) categories of neuron recognition, axon guidance, central nervous system (CNS)

development, and synapse (**Fig. 1d**, **Extended Data Fig. 2b**; **Methods**). Genes located in regions with highest inter-chromosomal interaction changes between CP and ES cells were enriched for developmental genes involved in forebrain development and chromatin organization (**Fig. 1d**, **Extended Data Fig. 2b**), indicating that these interactions reflect tissue relevant developmental gene regulation.

To further explore how these physical chromatin interactions relate to biological function, we hypothesized that highly interacting chromatin regions would be more likely to be co-regulated. To test this, we compared the distribution of correlation patterns for genes locating in (1) the regions of highest interaction values in both CP and GZ, (2) the lowest interacting regions in both CP and GZ, and (3) the regions of differential interaction values (the regions of highest interaction values in CP and lowest interaction values in GZ and vice versa). Highly interacting regions tend to be biased toward positive correlations, while there was no bias in correlation for low and differential interacting regions (**Fig. 1e**). Interestingly, the positive correlation for high interacting regions becomes even higher when more stringent cutoffs are used, supportive of the quantitative nature of interaction-driven co-expression, whereby the relationship between physical 3D chromatin interactions and expression is mostly driven by the top percentiles of interacting regions (**Extended Data Fig. 2c**). To further elucidate the epigenetic regulatory mechanisms behind the apparent interaction-mediated co-expression, we marked bins in which epigenetic marks from two loci appear together. By comparing the epigenetic mark combination matrix with the Hi-C contact matrix, we observed that interacting regions exhibit shared epigenetic patterns at the level of both inter- and intra-chromosomal interactions (**Fig. 1f**, **Extended Data Fig. 3**; **Methods**). In particular, regions associated with positive transcriptional regulation and enhancers are more likely to physically interact with each other, consistent with their co-regulation.

One of the core functional units of general genome organization recently uncovered by chromatin capture methods across a wide variety of cell types is the compartment, a relatively large, dynamic domain[1], which is comprised of smaller, sub-Mb regions of topologically associating domains (TADs)[12]. Compartments are divided into two types, type A compartments that consist primarily of euchromatin and actively transcribed genes and type B compartments, which are heterochromatic and repressed. TADs have been previously shown to be relatively stable, whereas compartments have been shown to change during lineage specification in stem cells[11]. Consistent with this, we observed dynamic compartment switching between CP and GZ, enriched for GO categories related to neuronal genes and phosphatase activity (**Fig. 2c**), as well as compartment switching between CP and ES (**Fig. 2a,d**). Genes that change compartments from ES to CP are decreased for A to B transitions across differentiation and increased for changes from the B to A compartments (**Fig. 2b**), as expected. Compartment changes are also accompanied by epigenetic changes, so that the B to A compartment shift is associated with increased DHS and active epigenetic marks indicative of open chromatin, whereas the A to B shift is associated with decreased DHS and increased repressive marks (**Fig. 2b,e**). The same pattern was observed for GZ vs. ES and CP vs. GZ (**Fig. 2b,e**, **Extended Data Fig. 2d**), demonstrating that gene expression changes across development are tightly linked to epigenetic changes coupled with compartment switching.

TADs are thought to mediate co-transcriptional regulation primarily within their boundaries (100kb-1Mb) through physical “looping” interactions of promotors and enhancers in co-regulated genes[4,16]. Since TAD boundaries are conserved across different cell types[12], we hypothesized that changes in epigenetic marks in TADs, rather than the boundaries of TADs, would be most associated with gene expression changes

across development. To test this, we divided genes based on their fold change in expression between ES and differentiated neurons[17] (both increased and decreased), and assessed changes in epigenetic marks within the TADs where these genes reside (**Extended Data Fig. 1c-e**, **Methods**). Notably, active marks including enhancers and elements related to transcribed regions are increased in TADs that contain upregulated genes, whereas repressive marks are increased in TADs that contain downregulated genes (**Fig. 2f**). Collectively, these results indicate that our Hi-C data reflects the major elements of global chromosome architecture in fetal brains, providing a framework for exploring gene regulatory mechanism related to human neural development and function.

Next, to demonstrate how knowledge of intra-chromosomal contacts could significantly advance understanding of important gene regulatory relationships in the nervous system, we performed two integrative experiments. In the first, we used these chromatin contact data to functionally annotate specific non-coding regulatory elements in the developing brain. We leveraged recent efforts that have identified >2000 developmental enhancers gained specifically in the human cerebral cortex, providing a remarkable resource for understanding the evolution of human cognition[8]. Usually, in the absence of such tissue specific data, regulatory elements are assigned to the closest gene[18,19], a convention that we compared with our Hi-C derived interactions. We reasoned that our Hi-C data from fetal brain could be used to identify the target genes for many of these enhancers, which based on previously chromatin looping analyses in cell lines are often not the closest gene[4,16,18,19].

We derived an interaction map of human-gained enhancers, defined as significant interacting regions (at a 1% false discovery rate, FDR) compared to the null distribution generated by fitting the contact frequencies of all fetal brain enhancers identified in the same study[8] (**Extended Data Fig. 4a**, **Methods**). We defined the search space as including the 1Mb flanking regions, since most enhancer-promoter interactions are within this range[4]. Although statistically significant interactions are increased upon proximity to the enhancer, the majority of interactions are at relatively long-ranges (>100kb, **Extended Data Fig. 4b**) and are not restricted to the adjacent genes. Indeed, ~65% of the closest genes to human-gained enhancers are not identified through fetal brain Hi-C interactions, revealing that the majority of enhancers are not interacting with the most adjacent gene (**Fig. 3c**). Compared to the original study[8], which used human-gained enhancer hotspot TADs in ES cells and IMR90 cells due to the lack of Hi-C data from relevant tissue, our approach provides genes of action with higher resolution in the matching tissue (fetal cortices) from which evolutionary enhancers were identified. Human-gained enhancer-interacting regions were enriched with enhancers, promoters, and transcription start sites (TSSs) (**Fig. 3a**, **Extended Data Fig. 4c**), consistent with the previous findings that enhancers interact with promoters, as well as other enhancers[16]. The majority of interactions (>75%) were in the same TADs (**Fig. 3b**), also consistent with observations in cell lines that most enhancer-promoter interactions are in the same TAD[16,19]. Human-gained enhancer interacting genes (Hi-$C_{evol}$ genes) are involved in GTPase regulation as well as G-protein coupled receptor (GPCR) and CREB signaling, and are enriched with GO terms representing synaptic and axon guidance genes (**Fig. 3e**, representative interactions in **Fig. 3d**). One striking example is a human-gained enhancer that interacts with *ARHGAP11B*, a human-specific gene implicated in the expansion of human neocortex[20] (**Fig. 3d**).

Given the high conservation of protein-coding genes across the vertebrate lineage, comparative genomics have suggested that human-specific traits most likely result from changes in regulatory elements[8,21]. Indeed, protein-coding Hi-$C_{evol}$ genes have a lower

non-synonymous substitution (dN)/synonymous substitution (dS) ratio compared to Hi-C non-interacting protein-coding genes in multiple lineages (**Extended Data Fig. 5**). These results indicate that human-gained enhancers are interacting with protein-coding genes that undergo purifying selection, further supporting the hypothesis that non-coding elements undergo evolutionary selection to induce species-specific changes in gene expression[8,21]. We also investigated whether human-gained enhancers are interacting with lineage-specific long non-coding RNAs (lncRNAs)[22]. We observed that lineage-specific interactions with human-gained enhancers were enriched for primate-specific lncRNAs, as well as evolutionary conserved lncRNAs (**Fig. 3f**, **Extended Data Fig. 5**). Thus, while human-gained enhancers interact and possibly regulate evolutionary conserved protein-coding genes, they are more likely to interact with primate-specific lncRNAs.

Since the development of human higher cognition is dependent on the development of the human cerebral cortex via elaboration of novel gene regulatory relationships[8,23], we reasoned, as have others[8] that the genes regulated by these human specific enhancers would be associated with intellectual functioning in humans. Remarkably, we found that the Hi-$C_{evol}$ genes in fetal brain, but not the genes defined by proximity to the enhancers are significantly enriched with intellectual disability (ID) risk genes[6] (**Fig. 3g**). This result provides experimental support for the contention that human-gained enhancers are associated with the evolution of human cognitive function[8]. This enrichment was tissue-specific, as Hi-$C_{evol}$ genes defined by Hi-C interactions in ES cells did not show enrichment for ID risk genes (**Fig. 3g**). Indeed, ~56% of the Hi-$C_{evol}$ genes in neuronal tissue were not identified through chromatin contacts in ES cells, emphasizing the importance of defining tissue-relevant chromatin contacts, as well as importance of using the relevant tissue for Hi-C analysis (**Fig. 4c**).

Since most disease related common genetic variation is located in non-protein coding regions, we next assessed the ability of Hi-C data for functional annotation of common single nucleotide polymorphisms (SNPs). As a first line verification that Hi-C data could identify known functional relationships between SNPs and gene expression we used *cis*-expression quantitative trait loci (eQTL) data from adult frontal cortex[24], since such data is not yet available from fetal brain. For each significant eQTL locus, we obtained a set of significant eQTL SNPs with >95% likelihood of containing the causal SNP from association statistics and linkage disequilibrium (LD; 1000 Genomes) structure using CAVIAR[25]. We then identified genes interacting to likely causal eQTL SNPs via the chromatin contact matrix (Hi-$C_{eQTL}$ genes, **Methods**), and compared Hi-$C_{eQTL}$ genes with the known associated gene from the eQTL study, finding that Hi-$C_{eQTL}$ genes significantly overlapped with eQTL transcripts (**Extended Data Fig. 6a**). There were many Hi-$C_{eQTL}$ genes that were not identified as eQTL transcripts, likely due to a combination of factors, including low power of the eQTL sample, limited resolution of Hi-C (SNP-transcript interactions within 20kb cannot be detected), and the difference in age of tissues used for each analysis. Indeed, eQTL SNPs identified by CAVIAR were highly enriched with adult frontal cortex, but not fetal brain, enhancers (**Extended Data Fig. 6b-d**). Despite this, eQTL SNP-transcript pairs exhibit higher chromatin contact frequency than expected by chance across all distance ranges (**Extended Data Fig. 6e**), further supporting the utility of Hi-C to infer the biological function of regulatory variation.

Next, we applied a similar logic to advance our understanding of 108 genome-wide significant schizophrenia-associated loci, most of which are in relatively uncharacterized non-coding regions of the genome[9]. We obtained credible SNPs using CAVIAR, and split SNPs into those without known function and likely functional SNPs (SNPs that cause missense, frameshift, and splice variants and SNPs that fall onto gene promoters;

Methods). Credible SNPs were enriched with enhancers in fetal brain and adult frontal cortex, confirming the likely regulatory role of these SNPs in the brain (**Extended Data Fig. 7**). SNPs defined as likely functional SNPs and promoter SNPs were directly assigned to their target genes. For the remaining intergenic and intronic SNPs that were un-annotated, and therefore without clear function, we used the chromatin contact matrix to find genes with which the regions where the SNPs are located are physically interacting (diagram in **Extended Data Fig. 7**).

Combining genes annotated as functional SNPs, promoter SNPs, and by Hi-C interactions, we obtained a total of ~900 genes (Hi-$C_{SCZ}$ genes) associated with schizophrenia risk variants. Hi-C contacts identified numerous genes that were neither adjacent to index SNPs nor in LD with them (**Fig. 4a-c**, **Extended Data Fig. 9**). While almost 70-80% of the LD genes and closest genes were identified as Hi-$C_{SCZ}$ genes, only half of them were identified by chromatin contacts, indicating that many of them were identified by functional SNPs residing in the genes. Moreover, 70-90% of the Hi-$C_{SCZ}$ genes were not identified by using LD genes or the closest genes to the association signal, consistent with observations that the linear organization of genes and regulatory elements on the chromosome does not reflect regulatory interactions[4,18,19].

Hi-C analysis showed that schizophrenia-associated common variants converge into specific molecular pathways related to neuronal function, including the postsynaptic density, acetylcholine receptors, cell cycle, and chromatin remodelers (**Fig. 4d-e**, **Extended Data Fig. 7-8**). To insure that this was not an artifact of the method used for credible SNP selection, we used a different method to define the set of credible SNPs[9] (**Extended Data Fig. 9**) and found the same enrichments, demonstrating the robustness of the genes identified through the Hi-C analysis. One notable example is illustrated by credible SNPs (rs4245150, rs17602038, rs4938021, rs4936275, rs4936276) that reside upstream of the *Dopamine D2 Receptor* (*DRD2*), the target of antipsychotic drugs. Although these SNPs are close to the *DRD2* TSS, they are not within the gene, which complicates interpretation of their biological function. Hi-C analysis demonstrates for the first time that indeed these SNPs are interacting with the TSS of *DRD2* (**Fig 4e**), providing biological insights into the function of these SNPs.

Another relevant example is an index SNP (rs79212538) interacting with *GRIA1*, an ionotropic glutamate receptor subunit, although *GRIA1* is neither the closest gene nor in LD with the index SNP (**Extended Data Fig. 8**). Additionally, Hi-C shows that schizophrenia associated non-coding SNPs interact with multiple genes involved in excitatory synaptic transmission, including *CACNA1C*, *GRIN2A*, and *NLGN4X*, further supporting glutamatergic transmission defects in schizophrenia pathophysiology (**Extended Data Fig. 8**). Interestingly, Hi-$C_{SCZ}$ genes significantly overlap with ASD *de novo* likely gene-disrupting (LGD) targets (CP: OR=2.4, P=$1.6 \times 10^{-5}$, GZ: OR=1.8, P=0.006), consistent with a shared genetic etiology between ASD and schizophrenia[26]. The fact that genes with LGD mutations in ASD are associated with regulatory variants in schizophrenia suggests that complete abrogation of these genes may cause developmental defects as in ASD, while regulatory changes in these genes may cause later-onset of neuropsychiatric symptoms as in schizophrenia. Collectively, genes annotated by chromatin contact information provide novel insights into schizophrenia pathogenesis.

In conclusion, we demonstrate how a comprehensive analysis of genome-wide chromatin configuration during human neural development informs our view of gene regulation. This chromatin contact landscape provides important biological insights on gene regulatory mechanisms, such that co-expressed genes share epigenetic co-regulation of interacting regions, and that changes in functional epigenetic marks are tightly linked to TADs and compartment switching to induce changes in gene expression.

We also annotated non-coding regulatory elements in the genome based on long-range chromatin contacts to identify enhancer-promoter interactions during human brain development, as well as genes of actions for eQTL. In turn, we show how these interactions can be used to inform our biological interpretation of risk variants for schizophrenia, which serves as a template for understanding the role of non-coding variation more broadly in neuropsychiatric disorders.

## Methods

### Fetal brain layer dissection

Human fetal cortical tissues from three individuals were collected from frontoparietal cortex at gestation week (GW) 17-18 (one sample from GW17 and two samples from GW18). In cold DMEM/F-12 (ThermoFisher, 11320-033), frontoparietal cortex was first dissected to thin (~1mm) slices to visualize layers. Under the light field microscope, cortical slice was dissected to germinal zone (GZ) and cortical plates (CP). GZ contains ventricular zone and subventricular zone, and hence comprised of proliferating neurons. CP refers to intermediate zone, cortical plate, and marginal zone, which are mainly composed of differentiated and migrating neurons. By dissecting layers from same fetal cortices, we can compare progenitors to differentiated neurons with same genotype and minimize intersample heterogeneity.

### Hi-C

Collected tissue was dissociated with trypsin and cell number was counted. Ten million cells were fixed in 1% formaldehyde for 10 min. Cross-linked DNA was digested by restriction enzyme HindIII (NEB, R0104). Digested chromatin ends were filled and marked with biotin-14-dCTP (ThermoFisher, 19518-018). Resulting blunt-end fragments were ligated under dilute concentration to minimize random intermolecular ligations. DNA purified after crosslinking was reversed by proteinase K (NEB, P8107) treatment. Biotins from unligated ends were removed by exonuclease activity of T4 DNA polymerase (ThermoFisher, 18005). DNA was sheared by sonication (Covaris, M220) and 300-600bp fragments were selected. Biotin-tagged DNA, which is intermolecular ligation products, was pulled down with streptavidin beads (Invitrogen, 65001), and ligated with Illumina paired end adapters. Resulting Hi-C library was amplified by PCR (KAPA Biosystems HiFi HotStart PCR kit, KK2502) with the minimum number of cycle (typically 12-13 cycles), and sequenced by Illumina 50bp paired-end sequencing.

### Hi-C reads mapping and pre-processing

Note that mapping and filtering of the reads, as well as normalization of experimental and intrinsic biases of Hi-C contact matrices were conducted with the following method regardless of cell types to minimize potential variance in the data obtained from different platforms. We implemented *hiclib* (https://bitbucket.org/mirnylab/hiclib) to perform initial analysis on Hi-C data from mapping to filtering and bias correction. Briefly, quality analysis was performed using a phred score, and sequenced reads were mapped to hg19 human genome by *Bowtie2* (with increased stringency, *--score-min -L 0.6,0.2--very-sensitive*) through iterative mapping. Read pairs were then allocated to HindIII restriction enzyme fragments. Self-ligated and unligated fragments, fragments from repeated regions of the genome, PCR artifacts, and genome assembly errors were removed. Filtered reads were binned at 10kb, 40kb, and 100kb resolution to build a genome-wide contact matrix at a given bin size. This contact map depicts contact frequency between any two genomic loci. Biases can be introduced to contact matrices by experimental procedures and intrinsic properties of the genome. To decompose biases from the contact matrix and yield a true contact probability map, filtered bins were subjected to iterative correction[13], the basic assumption of which is that each locus has uniform coverage. Bias correction and normalization results in a corrected heatmap of bin-level resolution. 100kb resolution bins were assessed for inter-chromosomal interactions, 40kb for TAD analysis, and 10kb for gene loop detection.

### Inter-chromosomal principal component analysis

Principal component analysis (PCA) was conducted in a genome-wide inter-chromosome contact map (100kb binned) as described previously[13]. Since intra-

chromosome conformation may drive the PCA results, *cis* contacts were iteratively replaced to random *trans* counts. After removing diagonal and poorly covered regions, we performed PCA using *hiclib* command *doEig*.

Pearson's correlations between the first principal components (PC1) from different cell types (CP, GZ, ES, and IMR90[12]) were calculated to compare similarities in inter-chromosomal interactions between different cell types.

Spearman's correlations between PC1/PC2 and biological traits (GC content, gene density, DNase I hypersensitivity (DHS), gene expression) were calculated. GC content (%) for each 100kb bin was calculated by *gcContentCalc* command from R package *Repitools*. Gene density (number of genes in 100kb bin) was obtained based on longest isoforms from GENCODE19. DHS of fetal brains from Epigenomic roadmap[10] and gene expression level of prenatal cortical layers from Miller et al.[15] were used and average values per 100kb bin were calculated.

### Gene enrichment analysis

Gene ontology (GO) enrichment was performed by GO-Elite Pathway Analysis (http://www.genmapp.org/go_elite/). All genes in the genome except the ones located in the chromosome Y and mitochondrial DNA were used as a background gene list. Because Hi-C interaction is measured in bins, sometimes we cannot dissect the individual genes when they are clustered in the genome (i.e. PCDH locus). To prevent several gene clusters overriding entire GO terms, we removed GO mainly defined by gene clusters (for 100kb or 40kb binned data) or we randomly included one gene per cluster (e.g. PCDHA1 for PCDHA1-13 cluster) prior to GO analysis (for 10kb binned data).

Gene enrichment for the curated gene lists was performed using binomial generalized linear model to regress out exome length. Autism spectrum disorder (ASD) *de novo* gene list and intellectual disability (ID) curated gene list from Iossifov et al.[27] and Pariskshak et al.[6] were used for the enrichment test, respectively. Protein-coding genes based on biomaRt were used as a background gene list.

### Identification of the regions with largest inter-chromosomal conformation changes

Chromosome contact matrix was normalized with the total interaction counts between two cell types for comparison. Intra-chromosomal interactions were masked from the genome-wide contact matrix, and top 1000 bins with the largest interaction changes between different cell types (GZ vs. CP or ES vs. CP) were selected. As one bin is comprised of two loci that are interacting with each other, this would give ~2000 sites in the genome. Genes located in those ~2000 sites were combined to perform GO analysis.

### Co-expression of inter-chromosomal interacting regions

Using transcriptome from fetal cortical layers[28], average expression values per 100kb bin were calculated. Pearson correlation matrix was calculated from 100kb binned expression data from all layers to generate gene co-expression matrix. At this step, gene co-expression matrix has the same dimension as inter-chromosomal contact matrix.

We hypothesized that genes would be co-expressed across the layers when they are interacting in all stages (both in CP and GZ), so we selected top 2% highest interacting regions of fetal brains both at GZ and CP (high interacting regions). We also selected (1) low interacting regions: top lowest interacting regions (0 interaction from normalized Hi-C contact matrix) of fetal brains both at GZ and CP, as well as (2) variant interacting regions: top 2% highest interacting regions from one stage (e.g. GZ) that are top 2%

lowest interacting regions from the other stage (e.g. CP) for comparison. Expression correlation values of the same regions were selected from the gene co-expression matrix, and expression correlations between different states (high interacting regions vs. low interacting regions and high interacting regions vs. variant interacting regions) were compared by two-sample Kolmogorov-Smirnov test.

### Epigenetic state enrichment for inter-chromosomal interacting regions

The fetal brain epigenetic 25 state model from Epigenomic roadmap[10] was used to generate the epigenetic state combination matrix, which was generated by marking loci where two interacting chromosomal bins (defined as bins with (1) interaction counts > 75% quantile interaction count for inter-chromosome and (2) interaction counts > 0 for intra-chromosome) share epigenetic signature. For example, the epigenetic combination matrix between the active transcription start site (TssA) and active enhancers (EnhA1) was generated by marking where interacting loci have TssA on one locus and EnhA1 on the other locus. Intra- and inter-chromosomal contact frequency maps were then compared to epigenetic state matrix by Fisher's exact test to calculate enrichment of shared epigenetic combinations in interacting regions.

### Compartment analysis

Expected interaction frequency was calculated from the normalized intra-chromosomal 40kb binned contact matrix based on the distance between two bins. We summed series of submatrices of 400kb window size with 40kb step size from the normalized Hi-C maps to generate observed and expected matrices. The Pearson's correlation matrix was computed from the observed/expected matrix, and PCA was conducted on correlation matrix. PC1 from each chromosome was used to identify compartments. Eigenvalues positively correlated with the gene density were set as compartment A, while those that are negatively correlated were set as compartment B.

### Gene expression and epigenetic state change across different compartments

Genomic regions were classified into three categories according to compartments: compartment A in cell type1 that changes to compartment B in cell type2 (A to B), compartment B in cell type1 that changes to compartment B in cell type2 (B to A), regions that do not change compartment between two cell types (stable).

Genes residing in each compartment category were selected and GO enrichment was performed. Gene expression fold-change (FC) between different cell types was calculated from Miller et al.[15] (comparison for CP vs. GZ) and CORTECON[17] (comparison for ES vs. CP and ES vs. GZ). Distribution of gene expression FC for genes in different compartment categories was compared by one-way ANOVA and Tukey's posthoc test.

15 state epigenetic marks from Epigenomic Roadmap[10] in genomic regions classified based on compartments were averaged across 40kb bins. The DHS FC[10] between different cell types (ES vs. CP and ES vs. GZ) was calculated and statistically evaluated as in the gene expression comparison. Each epigenetic state counts[10] for one compartment category was normalized by total epigenetic mark number of that compartment category and compared between ES and fetal brains.

### TAD analysis

We conducted TAD-level analysis as described previously[12]. Shortly, we quantified the directionality index by calculating the degree of upstream or downstream (2Mb) interaction bias of a given bin, which was processed by a hidden Markov model (HMM) to remove hidden directionality bias.

Regions in between TADs are titled as TAD boundaries when the regions are smaller than 400kb and unorganized chromatin when the regions are larger than 400kb.

**TAD-based epigenetic changes upon differentially expressed genes**

Genes were subdivided into 20 groups based on expression FC between ES and most differentiated neuronal states in CORTECON[17]: genes that are upregulated and downregulated upon differentiation were grouped into 10 quantiles, respectively, based on the FC. TADs into which genes from one subdivision reside were selected, and epigenetic state changes (from Epigenomic roadmap's 15 state epigenetic marks in ES and fetal brains[10]) in those TADs were normalized with TAD length and compared between ES and fetal brains. As different types of epigenetic marks have different absolute numbers (e.g. there are more quiescent states than enhancer states in the genome), each epigenetic state change was scaled across different quantiles to allow comparison between different states.

**Identification of Hi-C interacting regions**

We identified Hi-C interacting regions and target genes for (1) human-gained enhancers[8], (2) expression quantitative trait loci (eQTL) SNPs[24], and (3) schizophrenia SNPs[9]. As the highest resolution available for the current Hi-C data was 10kb, we assigned these enhancers/SNPs to 10kb bins, obtained Hi-C interaction profile for 1Mb flanking region (1Mb upstream to 1Mb downstream) of each bin. We also made a background Hi-C interaction profile by pooling (1) 255,698 H3K27ac sites from frontal and occipital cortex at 12 PCW for human-gained enhancers[8] and (2) 9,444,230 imputed SNPs for eQTL and schizophrenia SNPs[9]. To avoid significant Hi-C interactions affecting the distribution fitting as well as parameter estimation, we used the lowest 95 percentiles of Hi-C contacts and removed zero contact values. Using these background Hi-C interaction profiles, we fit the distribution of Hi-C contacts at each distance for each chromosome using *fitdistrplus* package (**Extended Data Fig. 4a**). Significance for a given Hi-C contact was calculated as the probability of observing a stronger contact under the fitted Weibull distribution matched by chromosome and distance. P-values were adjusted by computing FDR, and Hi-C contacts with FDR<0.01 were selected as significant interactions. Significant Hi-C interacting regions were overlapped with GENCODE19 gene coordinates (including 2kb upstream to transcription start sites (TSS) to allow detection of enhancer-promoter interactions) to identify interacting genes. Same analysis was performed on Hi-C contact maps from CP, GZ, and ES[11]. To address the functional significance of target genes, GO enrichment was performed for the interacting genes.

**Protein-coding genes interacting with human-specific evolutionary enhancers**

Protein-coding genes based on biomaRt (GENCODE19) were selected and non-synonymous substitution (dN)/synonymous substitution (dS) ratio was calculated for homologs in mouse, rhesus macaque, and chimpanzee for representation of mammals, primates, and great apes, respectively. Log2(dN/dS) distributions for protein-coding genes interacting vs. non-interacting to human-specific evolutionary enhancers in each lineage were then compared by two-sample Kolmogorov-Smirnov test.

**LncRNAs interacting with human-specific evolutionary enhancers**

Long non-coding RNAs (lncRNAs) classified according to evolutionary lineages[22] were used to assess whether lineage-specific lncRNAs are interacting to human-specific evolutionary enhancers. We randomly selected the same number of enhancers (2,104) to the human-specific ones from the total enhancer pool (255,698), identified interacting regions based on the null distribution generated from a background enhancer interaction profile. Significant interacting regions (FDR<0.01) identified by Hi-C were intersected

with lncRNA coordinates[22] and interacting lncRNAs for each lineage were counted. This step was repeated for 3,000 times to obtain the lncRNA lineage distribution. LncRNAs interacting with human-specific evolutionary enhancers were also identified and enrichment was tested by calculating P-values as the probability of observing more interacting lncRNAs for a given lineage under the null lncRNA lineage distribution.

**Epigenetic state enrichment for Hi-C interacting regions**

The functional framework for (1) eQTL SNPs, (2) schizophrenia SNPs, and (3) human-gained enhancers-interacting regions was assessed for epigenetic state enrichment. We implemented the same approach as in GREAT[29] to analyze the epigenetic state enrichment for *cis*-regulatory regions. For example, to evaluate whether schizophrenia SNPs are enriched with DHS, fraction of genome annotated with DHS (p), the number of schizophrenia SNPs (n), and number of schizophrenia SNPs overlapping with DHS (s) were calculated. Significance of the overlaps was tested by binomial probability of $P = Pr_{binom}\ (k \geq s \mid n = n,\ p = p)$[29]. Histone marks and 15-chromatin states from fetal brains, adult frontal cortex, and IMR90[10] were used for epigenetic state enrichment.

**eQTL analysis**

To address whether co-localization mediates gene regulation, we compared the association between chromosome conformation with eQTL. Although fetal brain eQTL data would be optimal, since this data is currently not available, we analyzed adult frontal cortex *cis*-acting eQTL data[24]. We selected SNPs associated with gene expression (FDR<0.01) and clustered them with association $P<1\times10^{-5}$ and $r^2>0.6$ to obtain index SNPs. Using summary association statistics and linkage disequilibrium (LD) structure for each index SNP, we applied *CAVIAR*[25] to quantify the probability of each variant to be causal. Among 121,273,364 SNP-transcript pairs from frontal cortex eQTL data, this process resulted in 42,190 SNP-transcript pairs (267 transcripts and 14,882 SNPs) that are potentially credible. We refer to 14,882 credible SNPs as credible SNPs. Credible SNP interacting genes were identified as described in "identification of Hi-C interacting regions" section.

Fisher's exact test was performed to evaluate the significance of the overlap between Hi-C interacting genes and eQTL transcripts. The background gene list for Fisher's exact test includes genes located in 1Mb flanking regions to credible SNPs that are also tested in eQTL analysis.

For 42,190 SNP-transcript pairs, we assigned credible SNPs and genes into 10kb bins, and obtained Hi-C contacts between credible SNPs and genes from the 10kb binned Hi-C contact maps. As a gene can span across multiple 10kb bins, the highest interaction in the gene to a credible SNP was selected as Hi-C contacts as previously defined[30]. We also calculated expected interaction frequency from the normalized 10kb binned contact matrix based on the distance between two bins. Opposite interaction frequency was calculated by obtaining Hi-C contacts for the opposite site to the credible SNP with the same distance. Because interaction counts differ in different chromosomes as well as in different cell types, we normalized interaction by chromosomes and cell types. We performed one-way ANOVA and Tukey's posthoc test for the comparison between different interaction paradigms.

**Identification of credible SNPs for schizophrenia GWAS loci**

128 LD-independent SNPs with genome-wide significance ($P<5\times10^{-8}$)[9] were used as index SNPs to obtain schizophrenia credible SNPs. All SNPs that are associated with $P<1\times10^{-5}$ and in LD ($r^2>0.6$) with an index SNP were selected, and correlations among this set of SNPs (LD structure) were calculated. CAVIAR was applied to summary association statistics and LD structure for each index SNP, and potentially causal SNPs

for each index SNP were identified. Among 55,000 SNPs that are in LD with 128 index SNPs, 7,613 SNPs were selected as causal by CAVIAR. Here we refer to these CAVIAR-identified SNPs as credible SNPs. Genes interacting to credible SNPs were identified as described in "identification of Hi-C interacting regions" section for CP, GZ, and ES. A separate set of credible SNPs initially reported from the original study was also processed with the same method[9].

### Identification of schizophrenia GWAS SNP-associated genes

We classified credible SNPs based on potential functionality (flow chart in **Extended Data Fig. 7**). For credible SNPs classified as functional (stop gained variant, frameshift variant, splice donor variant, NMD transcript variant, and missense variant) from biomaRt, we selected genes in which those SNPs locate. For those that are not directly affecting the gene function, we selected SNPs that fall onto the promoter and TSS of genes (2kb upstream-1kb downstream to TSS). Remaining SNPs were tested for Hi-C interaction so that Hi-C interacting genes were identified. This pipeline gives total ~900 genes potentially associated with GWAS SNPs.

### Identification of closest genes and LD genes

Closest genes to human-gained enhancers and schizophrenia index SNPs were obtained by *closestBed* command from *bedtools*. Gene coordinates from GENCODE19 including 2kb upstream to TSS were used to identify the closest genes.

LD genes refer to all genes in the LD. Here, LD is defined as physically distinct schizophrenia-associated 108 genome-wide significant regions[9]. We overlapped gene coordinates from GENCODE19 with LD regions to find genes that reside in LD.

Closest genes and LD genes were compared with Hi-C interacting genes. Venn diagrams were generated by *Vennerable* package in R. Only protein-coding genes were included in plotting Venn diagrams.

### Calculation of distance between SNPs and genes

For LD genes and closest genes, the shortest distance between an index SNP and a target gene was selected. For credible SNPs, (1) the distance between functional credible SNPs and target genes was set as 0, because functional SNPs reside in the gene, (2) the distance between promoter credible SNPs and target genes was calculated as the distance between SNPs and TSS of a gene, (3) the distance between credible SNPs and Hi-C interacting genes was calculated based on the distance between SNPs and Hi-C interacting bins (note that this distance has a unit of 10kb). We then combined the distance distributions from the 3 categories.

## Figure Legends

**Figure 1. Chromosome conformation in fetal brains reflects genomic features. a.** Representative heatmap of the chromosome contact matrix of CP. Normalized contact frequency (contact enrichment) is color-coded according to the legend on the right. **b.** Pearson correlation of the leading principle component (PC1) of inter-chromosomal contacts at 100kb resolution between *in vivo* cortical layers and non-neuronal cell types (ES and IMR90). **c.** Spearman correlation of PC1 of chromatin interaction profile of fetal brain (GZ) with GC content (GC), gene number, DNase I hypersensitivity (DHS) of fetal brain, and gene expression level in fetal laminae. **d.** GO enrichment of genes located in the top 1000 highly interacting inter-chromosomal regions specific to CP vs. GZ (left), and CP vs. ES (right), indicating that genes located on dynamic chromosomal regions are enriched for neuronal development. **e.** The top 2% highest interacting regions of fetal brains both at GZ and CP (High) show positive correlation in gene expression, while the top 2% lowest interacting regions (Low) and top 2% highly variant regions (Variant) have no skew in distribution. P-values from Kolmogorov–Smirnov test. **f.** The epigenetic state combination in inter-chromosomal interacting regions in GZ. Inter-chromosomal contact frequency map is compared to epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1), transcriptional regulators (TxReg), and transcribed regions (Tx) interact highly to each other as marked in red. Colored bars on the left represent epigenetic marks associated with promoters and transcribed regions (orange), enhancers (red), and repressive marks (blue). Chr, chromosome. Annotation for epigenetic marks described in

http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Figure 2. Compartment and TADs provide insights into gene regulatory mechanism. a.** Leading principal component (PC1) of the intra-chromosomal contact matrix in CP, GZ, and ES, with the DNase I hypersensitivity (DHS) fold change (FC) between ES and fetal brain (FB). PC1 values indicate compartment status of a given region, where positive PC1 represents compartment A (red), and negative PC1 represents compartment B (green). **b.** Distribution of gene expression FC (left) and DHS FC (right) for genes/regions that change compartment status ("A to B" or "B to A") or that remain the same ("stable") in different cell types. P-values from one-way ANOVA. **c.** GO enrichment of genes that change compartment status from A to B (top) and B to A (bottom) in CP to GZ. **d.** Heatmap of PC1 values of the genome that change compartment status in different cell types. **e.** Percentage of epigenetic marks for genomic regions that change compartment status between ES and CP. Note that B to A shift in ES to CP is associated with increased proportion of active transcribed regions (TssA and Tx) and enhancers (Enh, top), while A to B shift in ES to CP is associated with increased proportions of repressive marks (Het and ReprPCWk, bottom). P-values from Fisher's exact test. **f.** Epigenetic changes in topological associating domains (TADs) mediate gene expression changes during neuronal differentiation. Genes were divided by expression FC between ES and differentiated neurons, and epigenetic marks in the TADs containing genes in each group were counted and compared between ES and CP. Upregulated genes in neurons locate in TADs with more active epigenetic marks in CP than in ES, while downregulated genes in neurons locate in TADs with more repressive marks in CP than in ES. Epigenetic states associated with activation and transcription of the genes were marked as a red bar, while those associated with repression were marked as blue bars on the right. Annotation for epigenetic marks

described in http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Figure 3. Genetic architecture of human-gained enhancers. a.** Fraction of epigenetic states for regions interacting to human-gained enhancers in CP and GZ. **b.** Proportions of whether human-gained enhancers and interacting regions are within the same topological associating domain (TAD) vs. outside of the TAD. **c.** Overlap between human-gained enhancer interacting genes (Hi-$C_{evol}$ genes) in CP and GZ with closest genes to human-gained enhancers (left) and Hi-$C_{evol}$ genes in ES (right). **d.** Representative interaction map of a 10kb bin, in which human-gained enhancers reside, with the corresponding 1Mb flanking regions. This interactome map provides genes of action that interact with human-gained enhancers. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; Evol, genomic coordinate for a 10kb bin in which human-gained enhancers reside; -log10(P-value), P-value for the significance of the interaction between human-gained enhancers and each 10kb bin, grey dotted line for FDR=0.01; TAD, TAD borders in CP, GZ, and ES. **e.** GO enrichment for Hi-$C_{evol}$ genes in CP (left) and GZ (right). **f.** Number of primate-specific long non-coding RNAs (lncRNAs) interacting with human-gained enhancers in CP (red vertical lines in the graph) against a background control generated from 3,000 permutations, where the number of lncRNAs interacting with the same number of enhancers pooled from all fetal brain enhancers was counted. **g.** Overrepresentation of Hi-$C_{evol}$ genes in different tissues and closest genes with a curated set of intellectual disability (ID) risk genes. *P<0.05, **P<0.01, *** P<0.001. TSS, transcription start site; OR, odds ratio; GPCR, G-protein coupled receptor; Hi-C genes: GZ, CP, ES, Hi-$C_{evol}$ genes in each tissue; Hi-C genes: FB, union of Hi-$C_{evol}$ genes in GZ and CP; Hi-C genes: ES-specific, Hi-$C_{evol}$ genes in ES but not in fetal brain (FB); Hi-C genes: FB-specific, Hi-$C_{evol}$ genes in FB (union) but not in ES; Closest genes, closest genes to human-gained enhancers.

**Figure 4. Annotation of significant chromatin interactions for schizophrenia-associated loci. a.** Overlap between closest genes to index SNPs (Closest), genes locating in linkage disequilibrium (LD), and genes identified through SNP categorization and chromatin contacts in CP and GZ (Hi-$C_{SCZ}$ genes, diagram in **Extended Data Fig. 7**). **b.** Number of closest genes and LD genes that interact to credible SNPs (Hi-C supported) and those that do not interact to credible SNPs (Hi-C non-supported, top). Number of genes that interact to credible SNPs that are closest to or in LD with index SNPs (Hi-C genes), and not closest to or in LD with index SNPs (Hi-C genes not, bottom). Note that Hi-C genes here contain physically interacting genes, but not genes identified by functional SNPs or promoter SNPs. **c.** Distance between CAVIAR/index SNPs and their target genes for closest genes to index SNPs (Closest), genes locating in linkage disequilibrium (LD), and Hi-$C_{SCZ}$ genes in CP (CP) and GZ (GZ) **d.** GO enrichment for Hi-$C_{SCZ}$ genes in CP (left) and GZ (right). **e.** Representative interaction map of a 10kb bin, in which credible SNPs reside, to the corresponding 1Mb flanking regions. This interactome provides target genes interacting to credible SNPs-containing region. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; SNP, genomic coordinate for a 10kb bin in which credible SNPs locate; -log10(P-value), P-value for the significance of the interaction between credible SNPs and each 10kb bin, grey dotted line for FDR=0.01; GWAS loci, LD region for the index SNP; TAD, topological associating domain borders in CP, GZ, and ES.

## Acknowledgements

This work was supported by NIH grants to D.H.G. (5R01MH060233:05; 5R01MH100027) and J.L.S (K99MH102357), Glenn/AFAR Postdoctoral Fellowship Program for Translational Research on Aging (20145357) and Basic Science Research Program through the National Research Foundation of Korea (NRF) (2013024227) to H.W., the California Institute for Regenerative Medicine (CIRM)-BSCRC Training Grant to L.T.U. (TG2-01169). The Hi-C library was sequenced by Broad Stem Cell Research Center (BSCRC), and fetal tissue was collected from Center for Aids Research (CFAR, 5P30 AI028697). eQTL data was kindly provided by Mina Ryten and Adaiklavan Ramasamy. We thank Suhua Feng, Michael Gandal, Yuan Tian, Vivek Swarup, Grant Belgard for helpful discussions and critical reading of the manuscript.

## Author Contributions

H.W. designed and performed experiments, interpreted results, and co-wrote the manuscript. L.T.U. performed sample collection and experiments. J.L.S., N.N.P., and F.H. analyzed data. C.L. helped establishing Hi-C protocol. J.E. and E.E. participated in the discussion of the results. D.H.G. supervised the experimental design and analysis, interpreted results, provided funding, and co-wrote the manuscript.

## Author Information

*Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles*

Hyejung Won, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Changhoon Lee, Daniel H. Geschwind

*Department of Biological Chemistry, University of California California Los Angeles*

Jason Ernst

*Department of Computer Science, University of California Los Angeles*

Farhad Hormozdiari, Eleazar Eskin

*Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles*

Daniel H. Geschwind, Eleazar Eskin, Jason Ernst

## References

1 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

2 Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

3 Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).

4 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294, doi:10.1038/nature12644 (2013).

5 Network & Pathway Analysis Subgroup of Psychiatric Genomics, C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209, doi:10.1038/nn.3922 (2015).

6 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021, doi:10.1016/j.cell.2013.10.031 (2013).

7 Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007, doi:10.1016/j.cell.2013.10.020 (2013).

8 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159, doi:10.1126/science.1260943 (2015).

9 Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

10 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

11 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

12 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

13 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).

14 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64, doi:10.1038/nature12593 (2013).

15 Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206, doi:10.1038/nature13185 (2014).

16 Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-1065, doi:10.1016/j.cell.2015.07.048 (2015).

17 van de Leemput, J. *et al.* CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. *Neuron* **83**, 51-68, doi:10.1016/j.neuron.2014.05.013 (2014).

18 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

19 Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).

20 Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465-1470, doi:10.1126/science.aaa1975 (2015).

21 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).

22 Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635-640, doi:10.1038/nature12943 (2014).

23 Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647, doi:10.1016/j.neuron.2013.10.045 (2013).

24 Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428, doi:10.1038/nn.3801 (2014).

25 Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508, doi:10.1534/genetics.114.167908 (2014).

26 McCarthy, S. E. *et al.* De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658, doi:10.1038/mp.2014.29 (2014).

27 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221, doi:10.1038/nature13908 (2014).

28 Miller, J. A., Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12698-12703, doi:10.1073/pnas.0914257107 (2010).

29 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

30 Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic acids research* **42**, 87-96, doi:10.1093/nar/gkt857 (2014).

**Extended Data Figure 1. Basic characterization of Hi-C libary. a.** Hi-C library sequencing information. Percentage for double-stranded (DS) reads indicates percentage of DS reads to all reads, and percentage for valid pairs and filtered reads indicates percentage of valid pairs and filtered reads to DS reads. **b.** Frequency distribution of Hi-C contacts in GZ (left) and CP (right) **c.** Size distribution of topological associating domains (TADs) in GZ (left) and CP (right). **d.** Size distribution of genomic regions in between TADs that are less than 400kb (TAD boundaries) in GZ (left) and CP (right). **e.** Size distribution of genomic regions in between TADs that are bigger than 400kb (unorganized chromosome) in GZ (left) and CP (right). Cis ratio, ratio of cis (intra-chromosomal) reads to the total number of reads; chr, chromosome.

**Extended Data Figure 2. Chromosome conformation is associated with various genomic features.** a. Spearman correlation of principal components (PCs) of chromatin interaction profile of CP with GC content (GC), gene number, DNase I hypersensitivity (DHS), and gene expression level of fetal brains. **b.** GO enrichment of genes located in the top 1000 regions that gain inter-chromosomal interactions in CP compared to ES (upper left), ES compared to CP (upper right), CP compared to GZ (lower left), and GZ compared to CP (lower right). **c.** Top 5% (left) and 10% (middle) highest interacting regions both in GZ and CP (High) show positive correlation in gene expression, while low interacting regions (Low) and variant interacting regions (Variant) have no skew in distribution. (Right) Mean (top) and median (bottom) values for gene expression correlation for high, low, and variant interacting regions with different cutoffs, indicating that higher the interaction, higher the correlation of gene expression. **d.** Percentage of epigenetic marks for genomic regions that change compartment status between ES and GZ. Note that B to A shift in ES to GZ is associated with increased proportion of active transcribed regions (TssA and Tx) and enhancers (Enh, top), while A to B shift in ES to GZ is associated with increased proportions of repressive marks (Het and ReprPCWk, bottom). P-values from Fisher's exact test. Annotation for epigenetic marks described in a core 15-state model from
http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 3. Interacting regions share epigenetic states. a.** Epigenetic state combination in inter-chromosomal interacting regions in CP. Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1), transcriptional regulatory regions (TxReg), and transcribed regions (Tx) interact highly to each other as marked in red. **b-c.** Epigenetic state combination in intra-chromosomal interacting regions in GZ (**b**) and CP (**c**). Enhancers (TxEnh5', TxEnh3', TxEnhW, EnhA1) and transcriptional regulatory regions (TxReg) interact highly to promoters (PromD1, PromD2) and transcribed regions (Tx5', Tx) as marked in red. Inter- and intra-chromosomal contact frequency map is compared to epigenetic state combination matrix by Fisher's exact test to calculate the enrichment of shared epigenetic combinations in interacting regions. Colored bars on the left represent epigenetic marks associated with promoters and transcribed regions (orange), enhancers (red), and repressive marks (blue). Annotation for epigenetic marks described in a 25-state model from
http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 4. Characterization of chromatin interactome of human-gained enhancers. a.** Distribution fitting of normalized chromatin interaction frequency between human-gained enhancers with 1Mb upstream (top) and 100kb upstream (bottom) regions. Weibull distribution (red line) fits Hi-C interaction frequency the best for every distance range. **b.** Distribution of the number of significant interacting loci to human-gained enhancers in GZ (top), CP (middle), and ES (bottom). **c.** Fraction of histone states (left) and epigenetic mark enrichment (right) for regions interacting with

human-gained enhancers in GZ and CP. CDF, cumulative distribution function; Annotation for epigenetic marks described in http://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

**Extended Data Figure 5. Human-gained enhancers interact to evolutionary lineage-specific long non-coding RNAs (lncRNAs). a.** Protein-coding genes interacting with human-gained enhancers in CP (CP) and GZ (GZ) have lower non-synonymous substitutions (dN)/synonymous substitutions (dS) ratio compared to protein-coding genes non-interacting to human-gained enhancers (All) in mammals (mouse), primates (rhesus macaque), and great apes (chimpanzee), indicative of purifying selection. **b.** Number of lineage-specific lncRNAs interacting to human-gained enhancers (red vertical lines in the graph) in GZ (top) and CP (bottom). Null distribution generated from 3,000 permutations, where the number of lncRNAs interacting to the same number of enhancers pooled from all fetal brain enhancers was counted.

**Extended Data Figure 6. Association between eQTL and Hi-C interaction. a.** Overlap between eQTL transcripts and genes physically interacting to eQTL SNPs in CP and GZ. Significance of the overlap between eQTL transcripts and Hi-C interacting genes described in the upper right (Fisher's exact test). Background gene list for Fisher's exact test is all transcripts assessed in eQTL study within 1Mb from eQTL SNPs. **b-d.** Histone state enrichment for eQTL SNPs in adult frontal cortex (FCTX, **b**), fetal brain (FB, **c**), and IMR90 (**d**). **e.** Hi-C interaction frequency between eQTL SNPs and transcripts is greater than expected by chance in the relevant cell type. Lowess smooth curve plotted with actual data points. CP, chromatin contact frequency in CP; GZ, chromatin contact frequency in GZ; ES, chromatin contact frequency in ES; Exp, expected interaction frequency given the distance between two regions; Opp, opposite interaction frequency: interaction frequency of SNPs and transcripts when the position of genes was mirrored relative to the eQTL SNP. ***P<0.001, P-values from repeated measure of ANOVA.

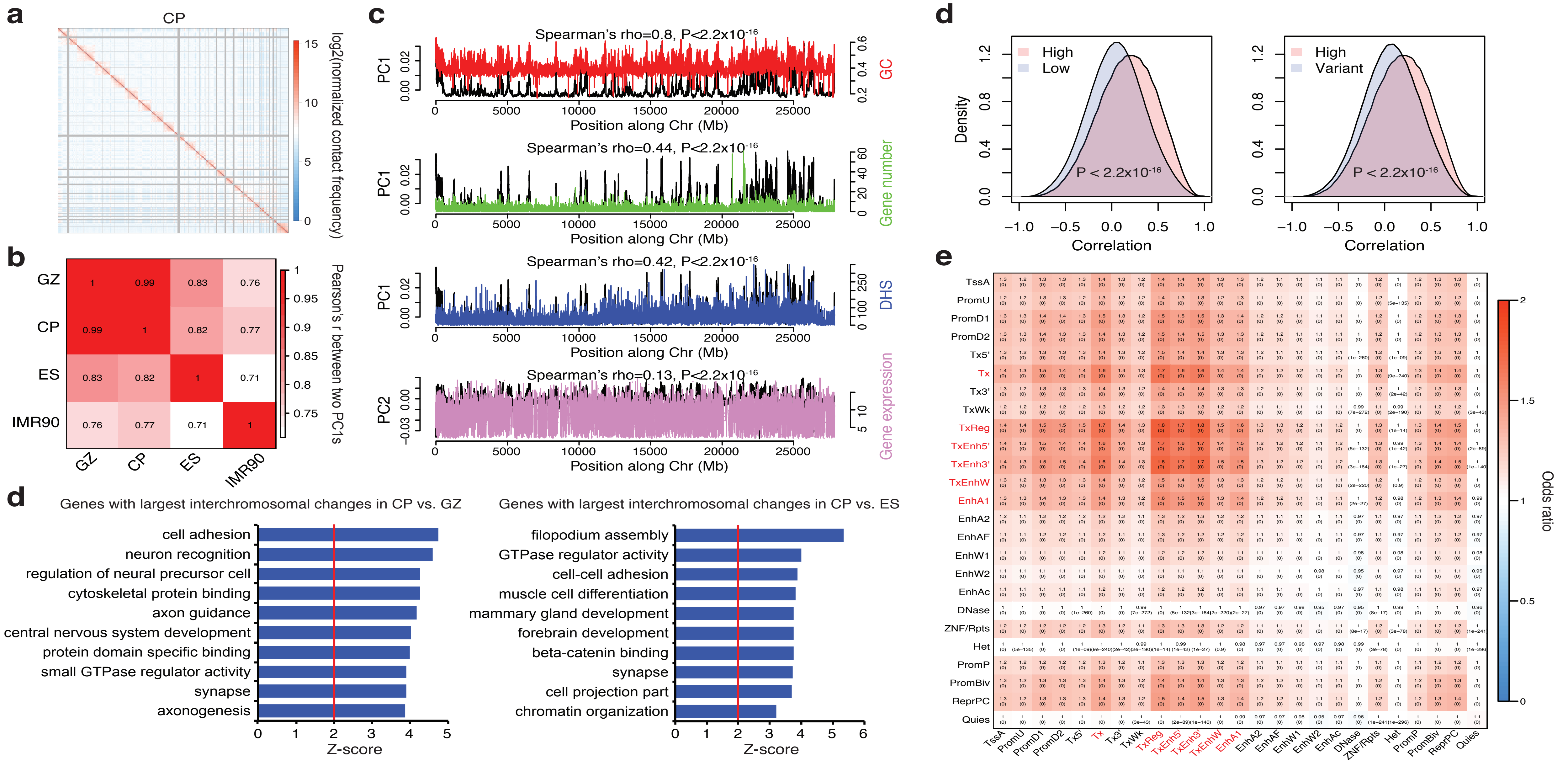
**Extended Data Figure 7. Defining schizophrenia risk genes based on functional annotation of credible SNPs.** Credible SNPs were selected using CAVIAR and categorized into functional SNPs, SNPs that fall onto gene promoters, and un-annotated SNPs. Histone state enrichment of credible SNPs was assessed in fetal brain (FB) and adult frontal cortex (FCTX). Functional SNPs and promoter SNPs were directly assigned to the target genes, while un-annotated SNPs were assigned to the target genes via Hi-C interactions in CP and GZ. GO enrichment for genes identified by each category is shown in the bottom. NMD, nonsense-mediated decay; TSS, transcription start site.

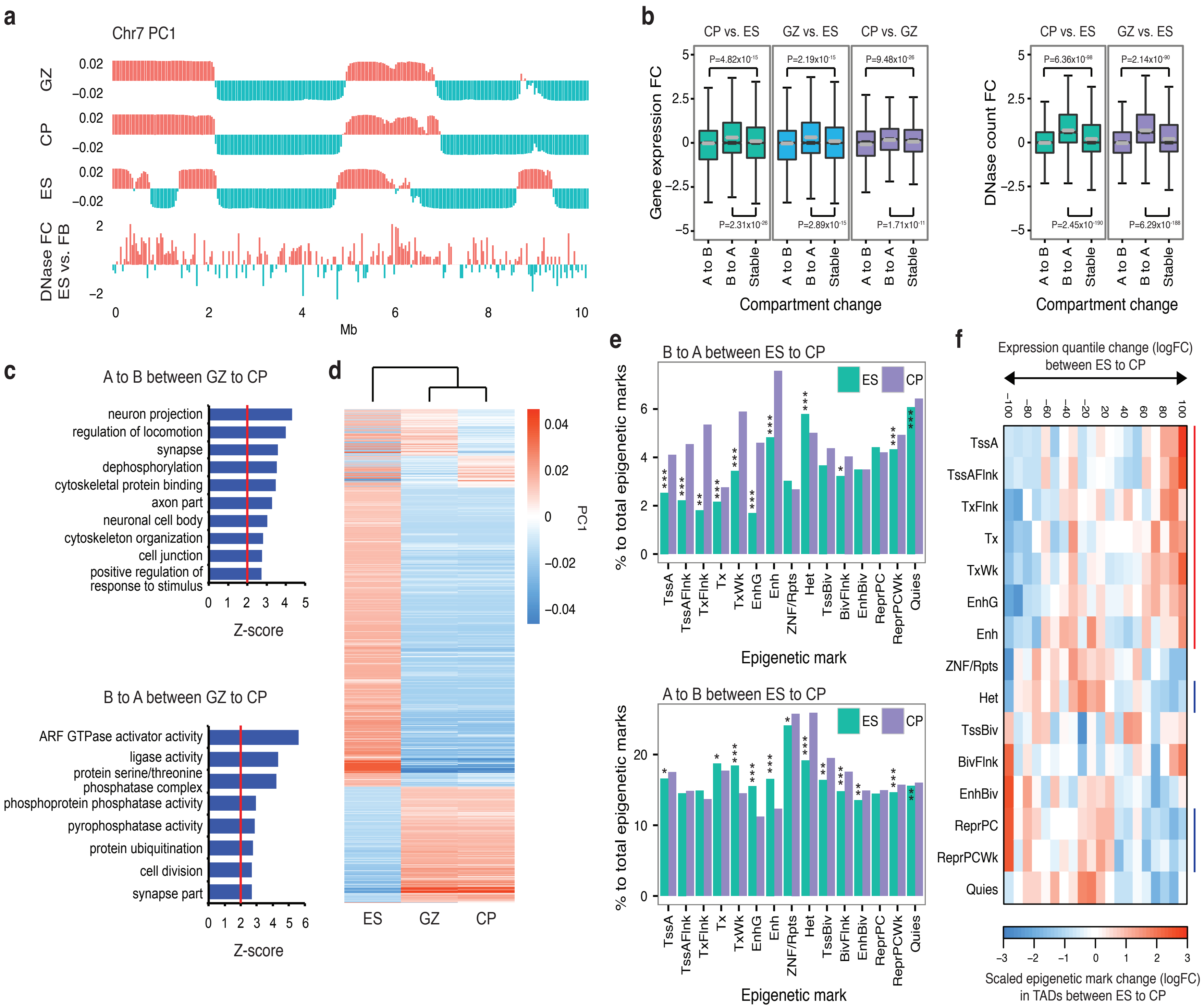
**Extended Data Figure 8. Representative interaction maps for credible SNPs to 1Mb flanking regions.** Interaction maps provide gene of actions for credible SNPs based on physical interaction. Chromosome ideogram and genomic axis on the top; Gene Model, gene model based on GENCODE19, possible target genes in red; SNP, genomic coordinate for a 10kb bin in which credible SNPs locate; -log10(P-value), P-value for the significance of the interaction between credible SNPs and each 10kb bin, grey dashed line for FDR=0.01; GWAS loci, linkage disequilibrium (LD) region with the index SNP; TAD, TAD borders in CP, GZ, and ES.

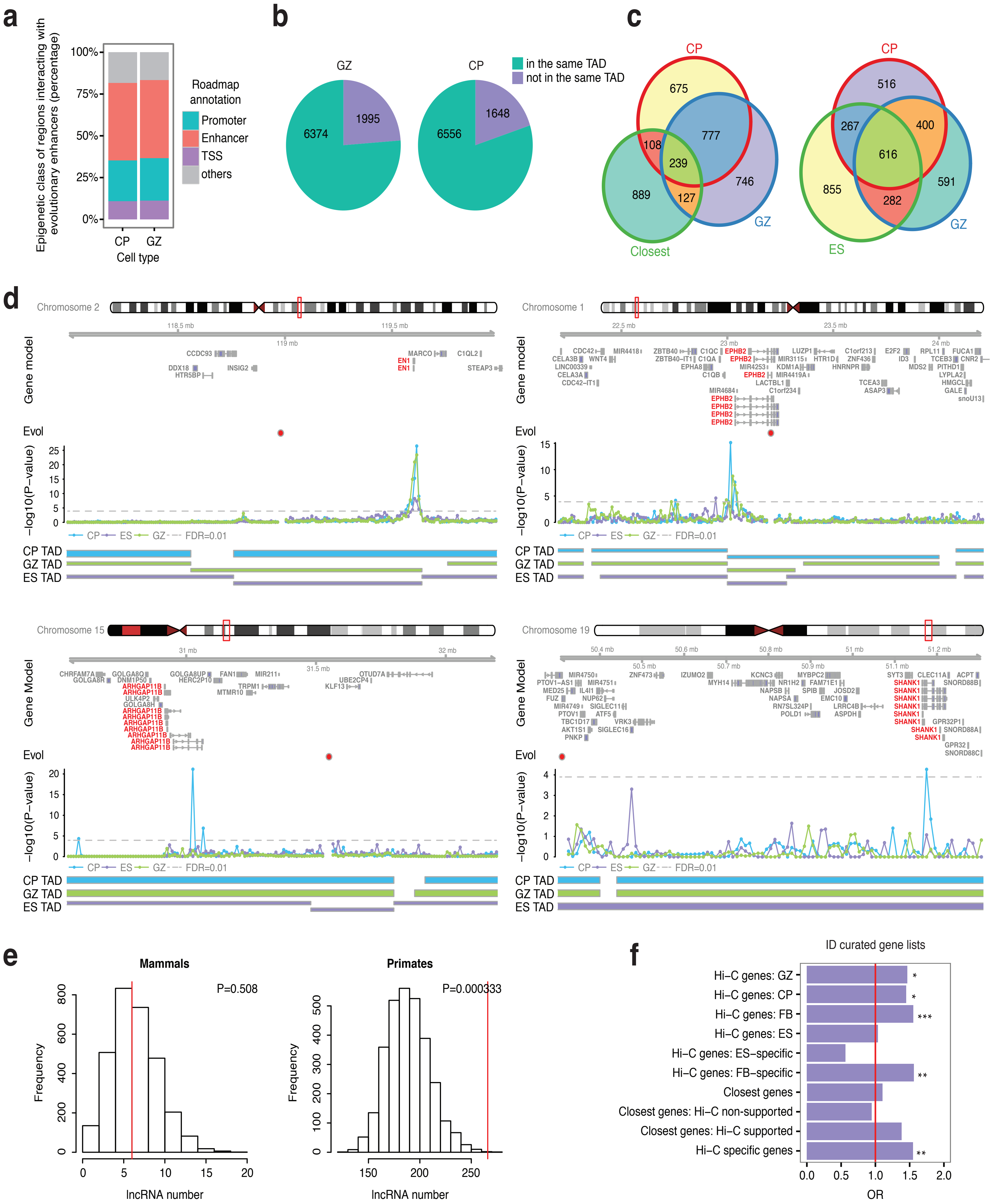
**Extended Data Figure 9. GO enrichment for schizophrenia risk genes curated by various methods. a-b.** GO enrichment for the closest genes to index SNPs (**a**) and genes in linkage disequilibrium (LD) with index SNPs (**b**) that are identified by a schizophrenia risk gene assessment pipeline in **Extended Data Fig. 7** (right) vs. not (left). **c.** GO enrichment for schizophrenia risk genes identified by a pipeline in **Extended Data Fig. 7** that are neither the closest genes nor in LD to index SNPs. Intersect and

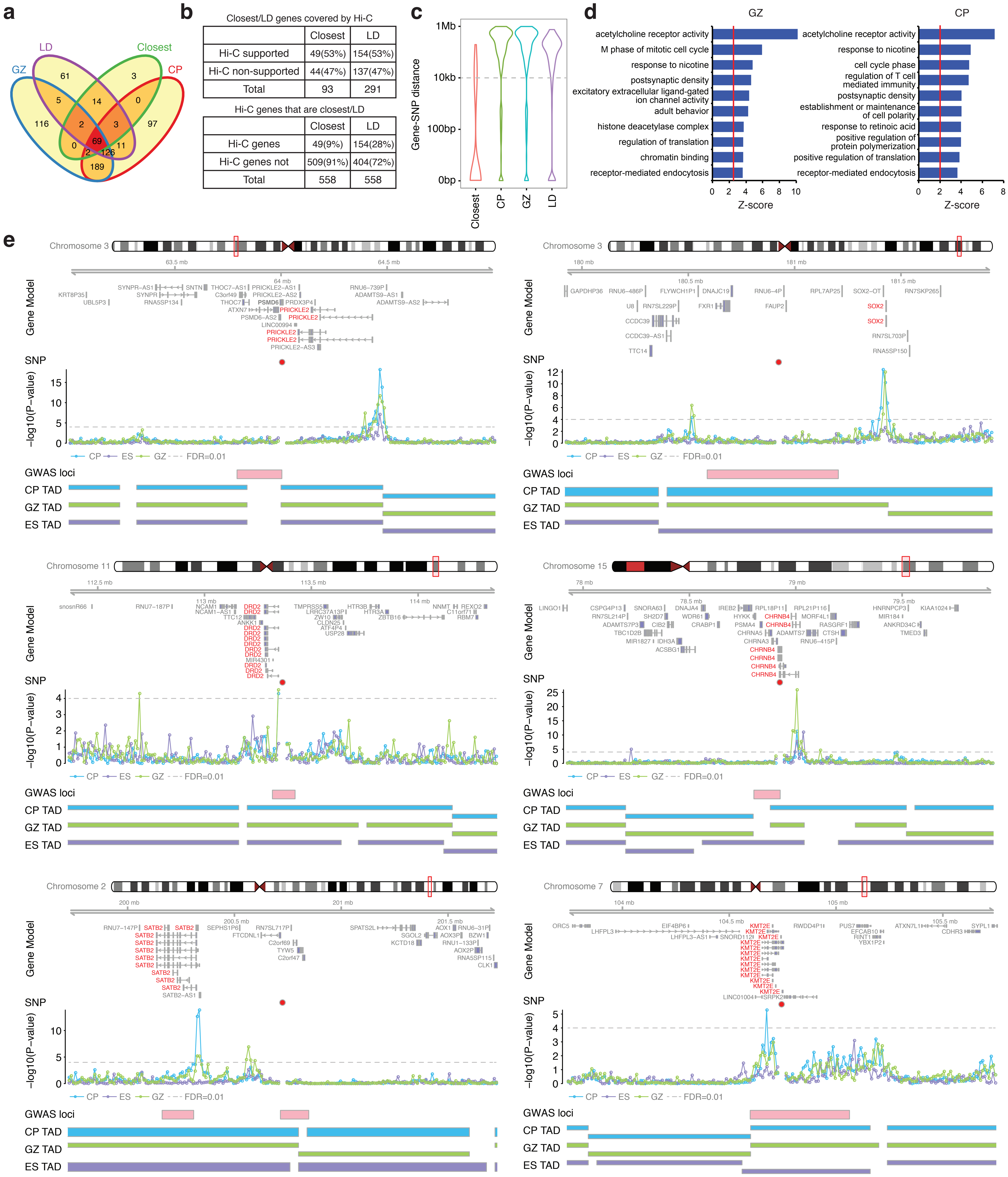
union between CP and GZ in left and right, respectively. Venn diagrams are marked in orange to depict the gene list assessed for GO enrichment.

**Extended Data Figure 10. Defining schizophrenia risk genes based on functional annotation of another set of credible SNPs.** Credible SNPs defined in the original study were categorized into functional SNPs, SNPs that fall onto gene promoters, and un-annotated SNPs. Overlap between credible SNPs identified by CAVIAR and credible SNPs originally identified indicates that two credible SNP lists overlap with each other. Histone state enrichment of credible SNPs in fetal brain (FB) and adult frontal cortex (FCTX). Functional SNPs and promoter SNPs were directly assigned to the target genes, while un-annotated SNPs were assigned to the target genes via Hi-C interactions in CP and GZ. GO enrichment for genes identified by each category and combined gene list is shown in the bottom. NMD, nonsense-mediated decay; TSS, transcription start site.

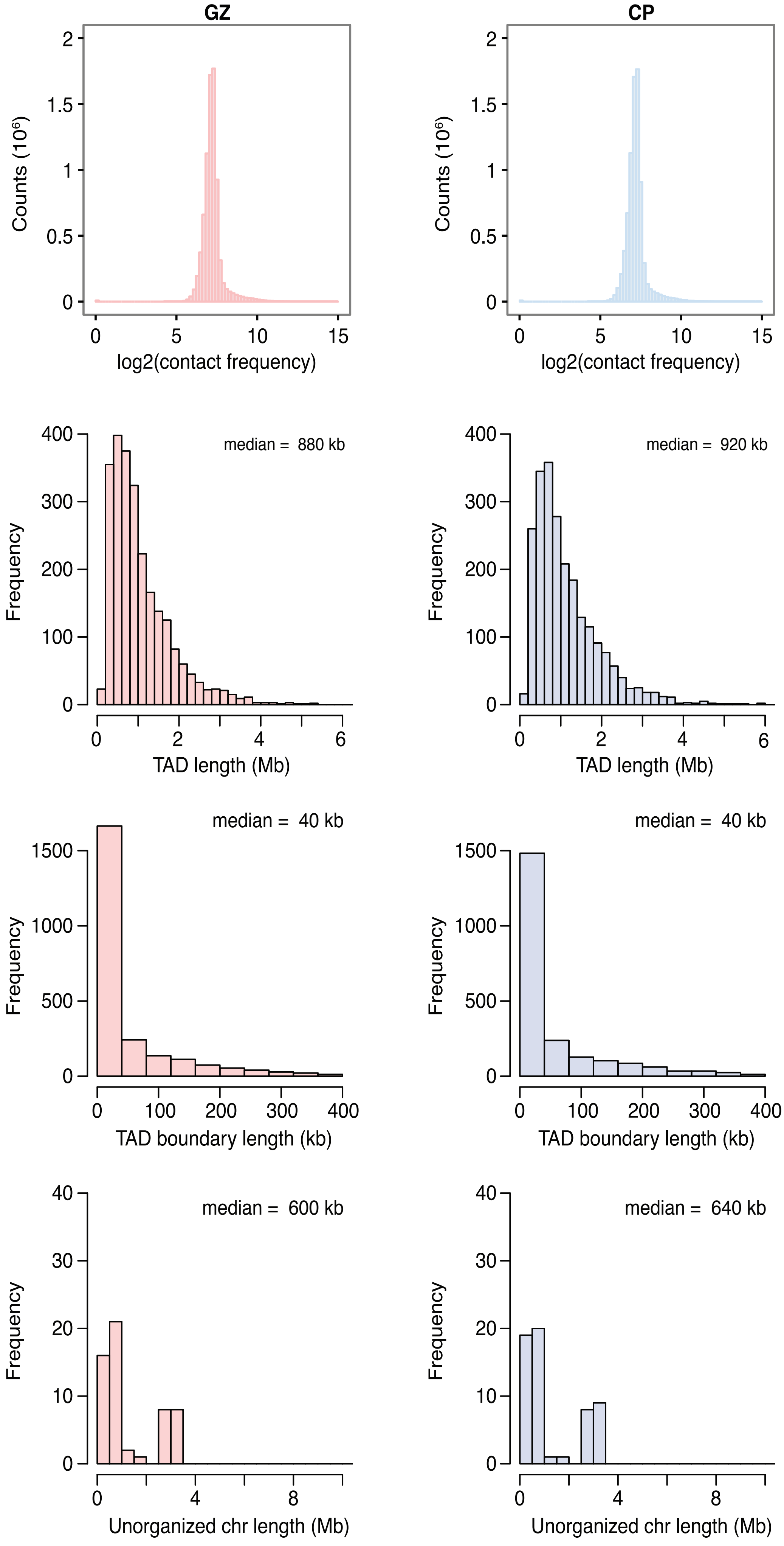

a
CP
log2(normalized contact frequency)
15
10
5
0
b
GZ
CP
ES
IMR90
1
0.99
0.83
0.76
0.99
1
0.82
0.77
0.83
0.82
1
0.71
0.76
0.77
0.71
1
Pearson's r between two PC1s
1
0.95
0.9
0.85
0.8
0.75
c
Spearman's rho=0.8, P<2.2x10-16
PC1
GC
Spearman's rho=0.44, P<2.2x10-16
Gene number
Spearman's rho=0.42, P<2.2x10-16
DHS
Spearman's rho=0.13, P<2.2x10-16
PC2
Gene expression
Position along Chr (Mb)
0
5000
10000
15000
20000
25000
d
High
Low
High
Variant
Density
P < 2.2x10-16
Correlation
−1.0
−0.5
0.0
0.5
1.0
d
Genes with largest interchromosomal changes in CP vs. GZ
cell adhesion
neuron recognition
regulation of neural precursor cell
cytoskeletal protein binding
axon guidance
central nervous system development
protein domain specific binding
small GTPase regulator activity
synapse
axonogenesis
Z-score
Genes with largest interchromosomal changes in CP vs. ES
filopodium assembly
GTPase regulator activity
cell-cell adhesion
muscle cell differentiation
mammary gland development
forebrain development
beta-catenin binding
synapse
cell projection part
chromatin organization
Z-score
e
TssA
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
Odds ratio
2
1.5
1
0.5
0

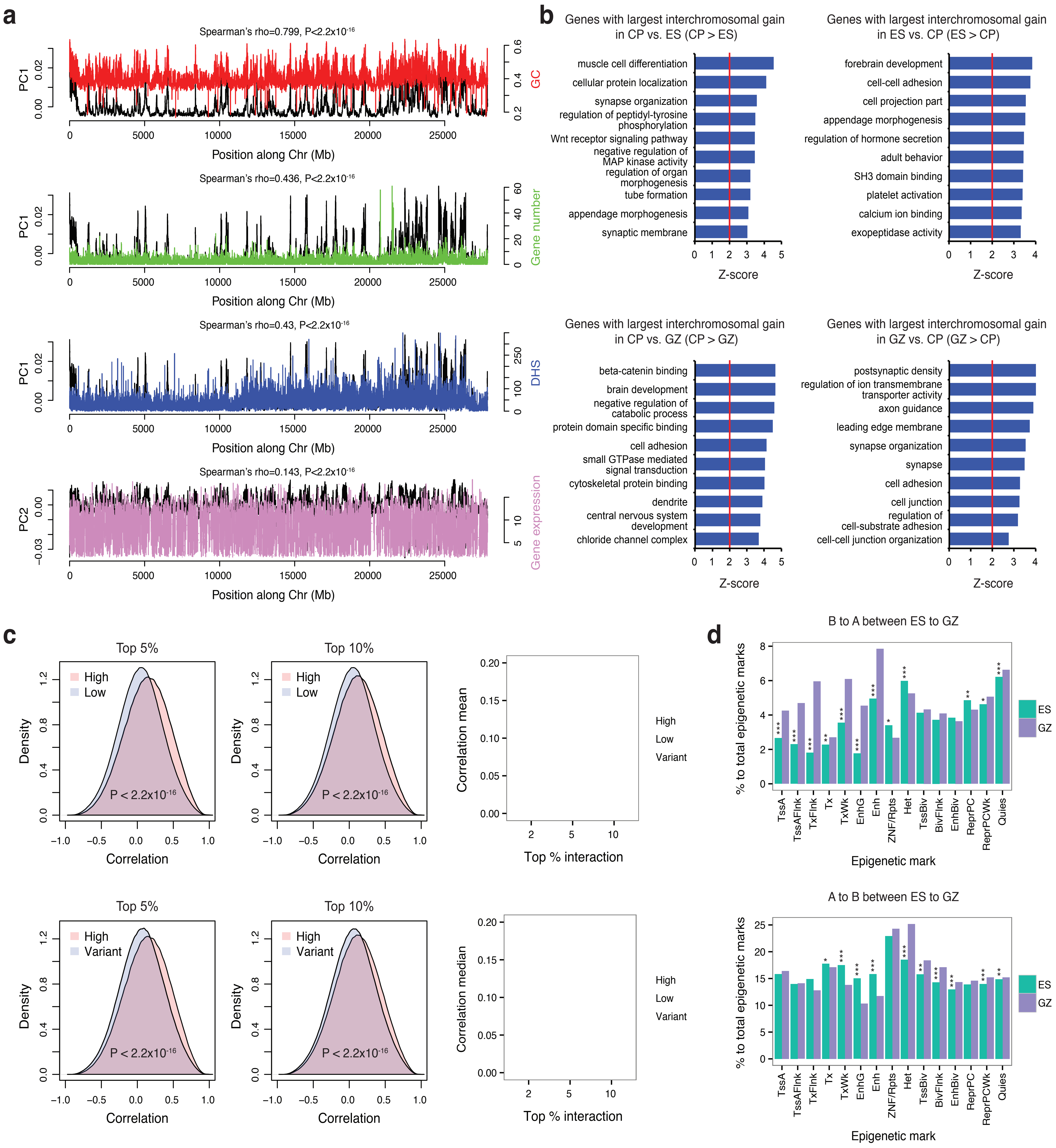

a
Chr7 PC1
GZ
CP
ES
0.02
−0.02
DNase FC ES vs. FB
2
−2
0
2
4
6
8
10
Mb
b
CP vs. ES
GZ vs. ES
CP vs. GZ
Gene expression FC
P=4.82x10$^{-15}$
P=2.19x10$^{-15}$
P=9.48x10$^{-26}$
P=2.31x10$^{-26}$
P=2.89x10$^{-15}$
P=1.71x10$^{-11}$
DNase count FC
P=6.36x10$^{-98}$
P=2.14x10$^{-90}$
P=2.45x10$^{-190}$
P=6.29x10$^{-188}$
5
2.5
0
−2.5
−5
A to B
B to A
Stable
Compartment change
c
A to B between GZ to CP
neuron projection
regulation of locomotion
synapse
dephosphorylation
cytoskeletal protein binding
axon part
neuronal cell body
cytoskeleton organization
cell junction
positive regulation of response to stimulus
Z-score
B to A between GZ to CP
ARF GTPase activator activity
ligase activity
protein serine/threonine phosphatase complex
phosphoprotein phosphatase activity
pyrophosphatase activity
protein ubiquitination
cell division
synapse part
Z-score
d
ES
GZ
CP
PC1
0.04
0.02
0
−0.02
−0.04
e
B to A between ES to CP
A to B between ES to CP
ES
CP
% to total epigenetic marks
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
Epigenetic mark
f
Expression quantile change (logFC) between ES to CP
−100
−80
−60
−40
−20
20
40
60
80
100
Scaled epigenetic mark change (logFC) in TADs between ES to CP
−3
−2
−1
0
1
2
3


a
Epigenetic class of regions interacting with evolutionary enhancers (percentage)
100%
75%
50%
25%
0%
CP
GZ
Cell type
Roadmap annotation
Promoter
Enhancer
TSS
others
b
GZ
CP
6374
1995
6556
1648
in the same TAD
not in the same TAD
c
CP
675
777
108
239
127
889
746
GZ
Closest
CP
516
267
400
616
855
282
591
GZ
ES
d
Chromosome 2
118.5 mb
119 mb
119.5 mb
Gene model
CCDC93
DDX18
HTR5BP
INSIG2
MARCO
EN1
EN1
C1QL2
STEAP3
Evol
-log10(P-value)
CP
ES
GZ
FDR=0.01
CP TAD
GZ TAD
ES TAD
Chromosome 1
22.5 mb
23 mb
23.5 mb
24 mb
EPHB2
Chromosome 15
31 mb
31.5 mb
32 mb
Gene Model
ARHGAP11B
Chromosome 19
50.4 mb
50.5 mb
50.6 mb
50.7 mb
50.8 mb
50.9 mb
51 mb
51.1 mb
51.2 mb
SHANK1
e
Mammals
P=0.508
Frequency
lncRNA number
Primates
P=0.000333
f
ID curated gene lists
Hi-C genes: GZ
Hi-C genes: CP
Hi-C genes: FB
Hi-C genes: ES
Hi-C genes: ES-specific
Hi-C genes: FB-specific
Closest genes
Closest genes: Hi-C non-supported
Closest genes: Hi-C supported
Hi-C specific genes
OR
0.0
0.5
1.0
1.5
2.0

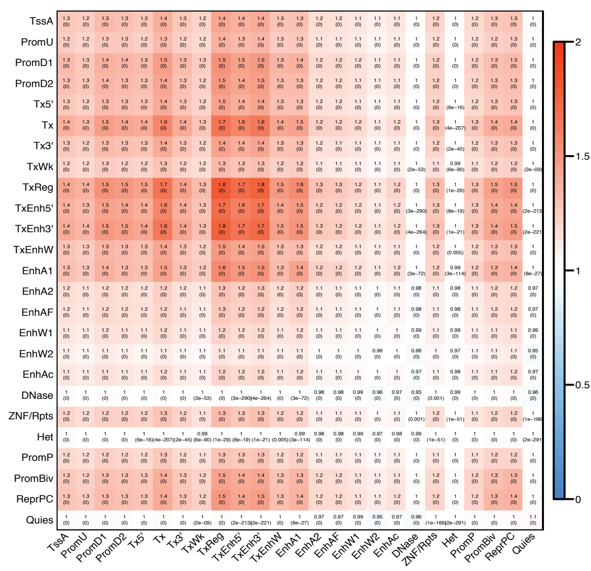


| Cell type | Cis ratio | All reads | DS mapped reads | Valid pairs | Filtered reads |
|---|---|---|---|---|---|
| GZ | 47.45% | 1,991,686,360 | 1,407,918,128<br>(70.69%) | 1,243,116,106<br>(88.29%) | 1,048,911,579<br>(74.50%) |
| CP | 46.40% | 1,958,637,304 | 1,352,951,087<br>(69.08%) | 1,225,315,488<br>(90.57%) | 1,022,593,960<br>(75.58%) |

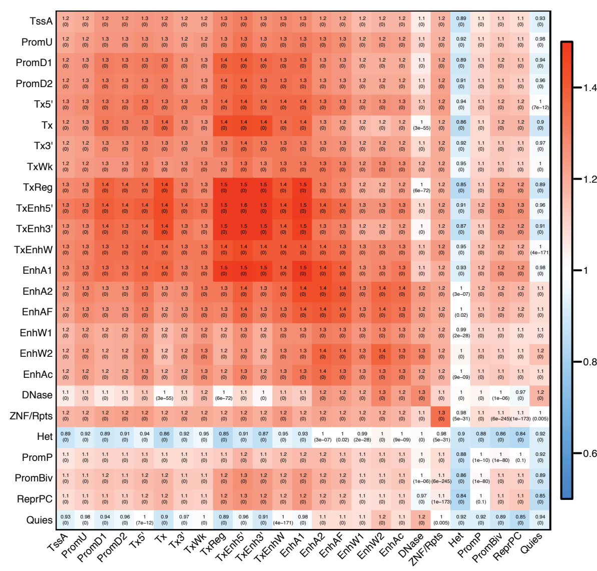



a
Spearman's rho=0.799, P<2.2x10-16
PC1
GC
Position along Chr (Mb)
Spearman's rho=0.436, P<2.2x10-16
Gene number
Spearman's rho=0.43, P<2.2x10-16
DHS
Spearman's rho=0.143, P<2.2x10-16
PC2
Gene expression
0 5000 10000 15000 20000 25000
b
Genes with largest interchromosomal gain in CP vs. ES (CP > ES)
muscle cell differentiation
cellular protein localization
synapse organization
regulation of peptidyl-tyrosine phosphorylation
Wnt receptor signaling pathway
negative regulation of MAP kinase activity
regulation of organ morphogenesis
tube formation
appendage morphogenesis
synaptic membrane
Genes with largest interchromosomal gain in ES vs. CP (ES > CP)
forebrain development
cell-cell adhesion
cell projection part
appendage morphogenesis
regulation of hormone secretion
adult behavior
SH3 domain binding
platelet activation
calcium ion binding
exopeptidase activity
Genes with largest interchromosomal gain in CP vs. GZ (CP > GZ)
beta-catenin binding
brain development
negative regulation of catabolic process
protein domain specific binding
cell adhesion
small GTPase mediated signal transduction
cytoskeletal protein binding
dendrite
central nervous system development
chloride channel complex
Genes with largest interchromosomal gain in GZ vs. CP (GZ > CP)
postsynaptic density
regulation of ion transmembrane transporter activity
axon guidance
leading edge membrane
synapse organization
synapse
cell adhesion
cell junction
regulation of cell-substrate adhesion
cell-cell junction organization
Z-score
c
Top 5%
Top 10%
High
Low
Variant
P < 2.2x10-16
Density
Correlation
Correlation mean
Correlation median
Top % interaction
d
B to A between ES to GZ
A to B between ES to GZ
% to total epigenetic marks
ES
GZ
TssA TssAFlnk TxFlnk Tx TxWk EnhG Enh ZNF/Rpts Het TssBiv BivFlnk EnhBiv ReprPC ReprPCWk Quies
Epigenetic mark

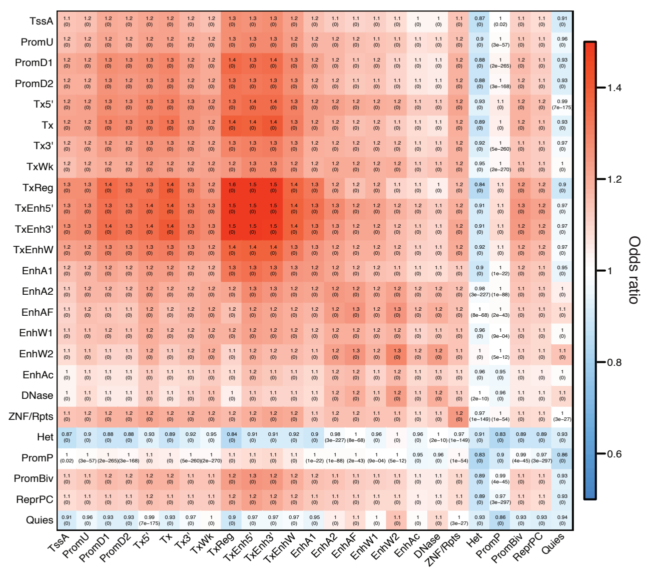
CP: interchromosomal

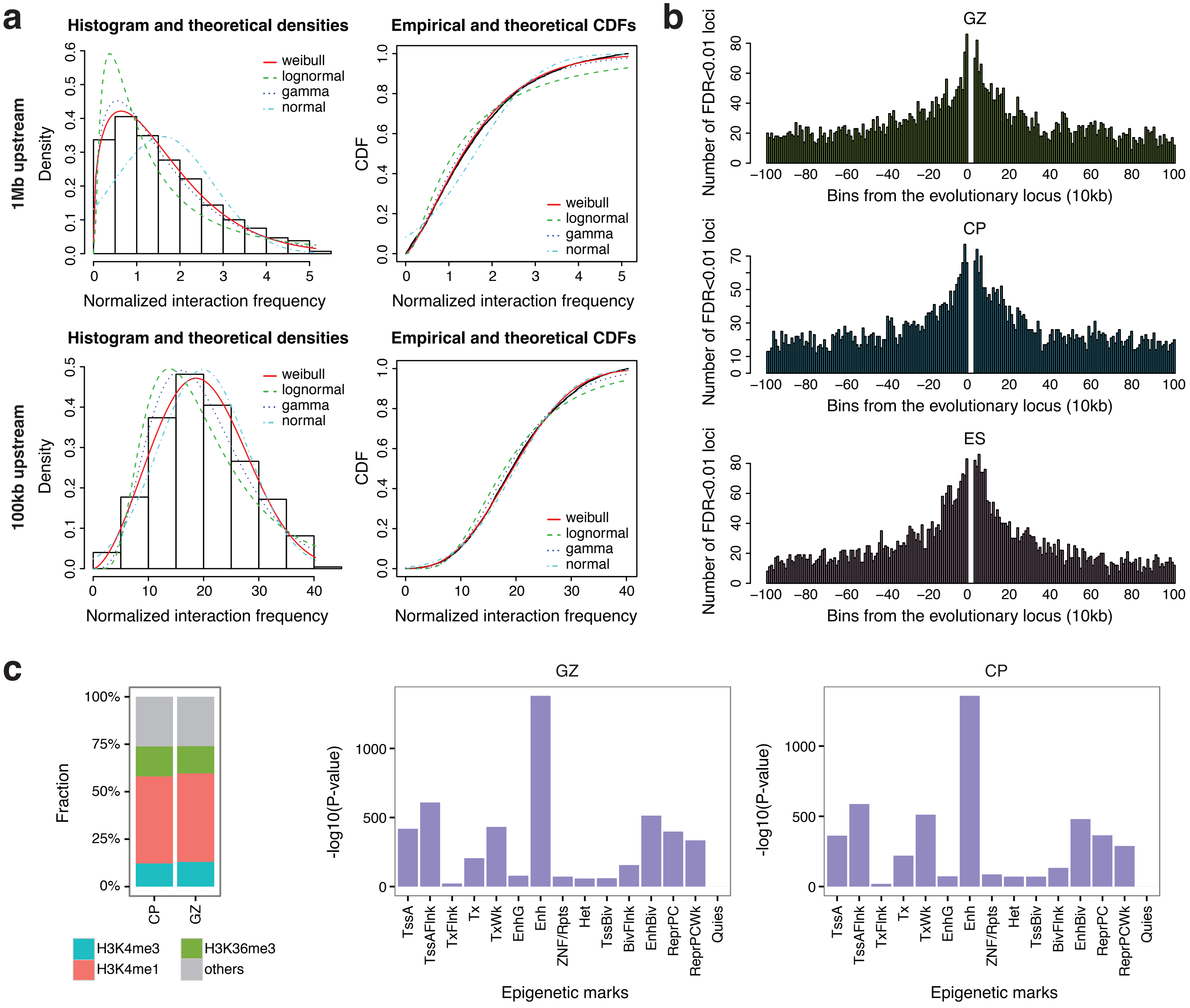


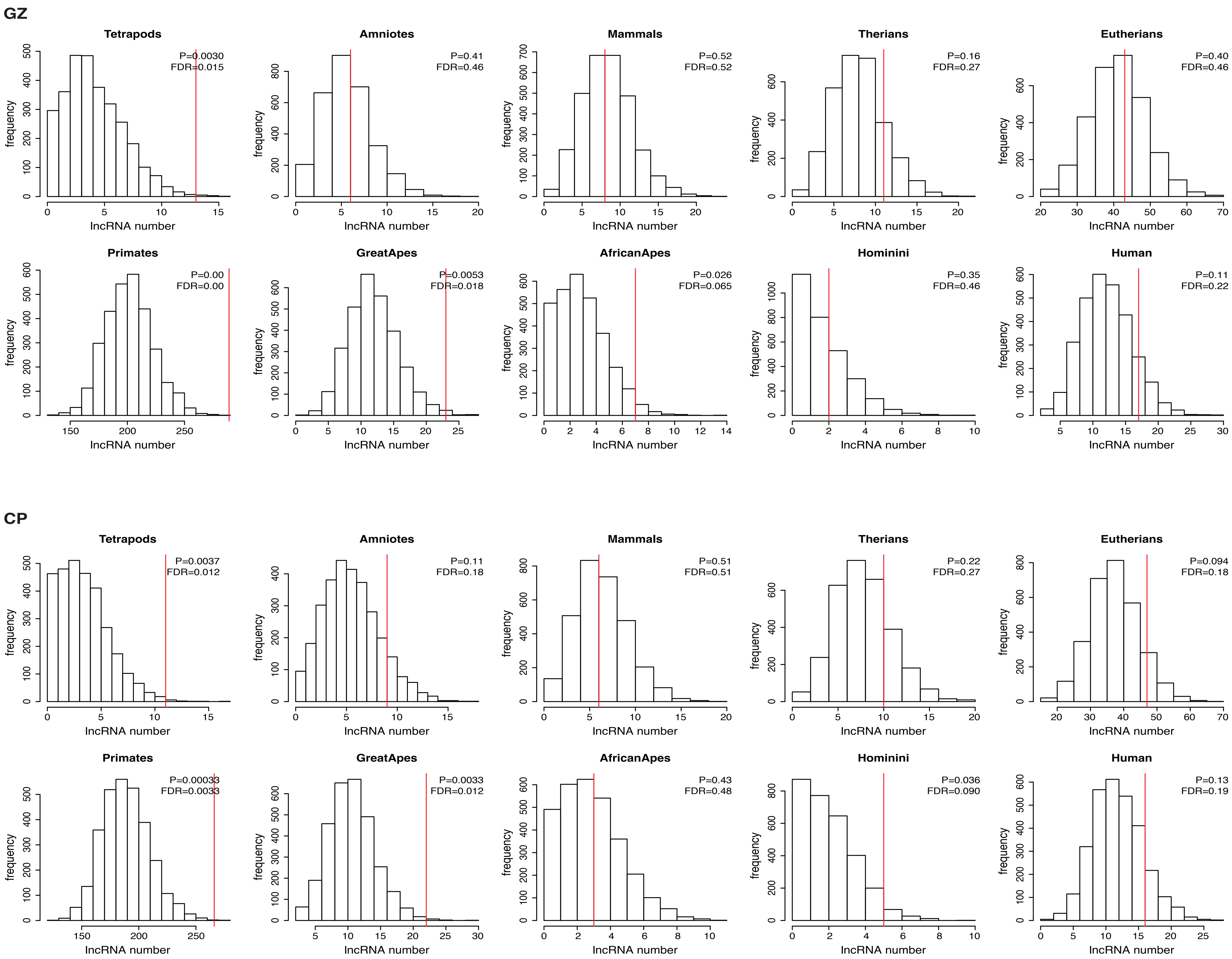
GZ: intrachromosomal

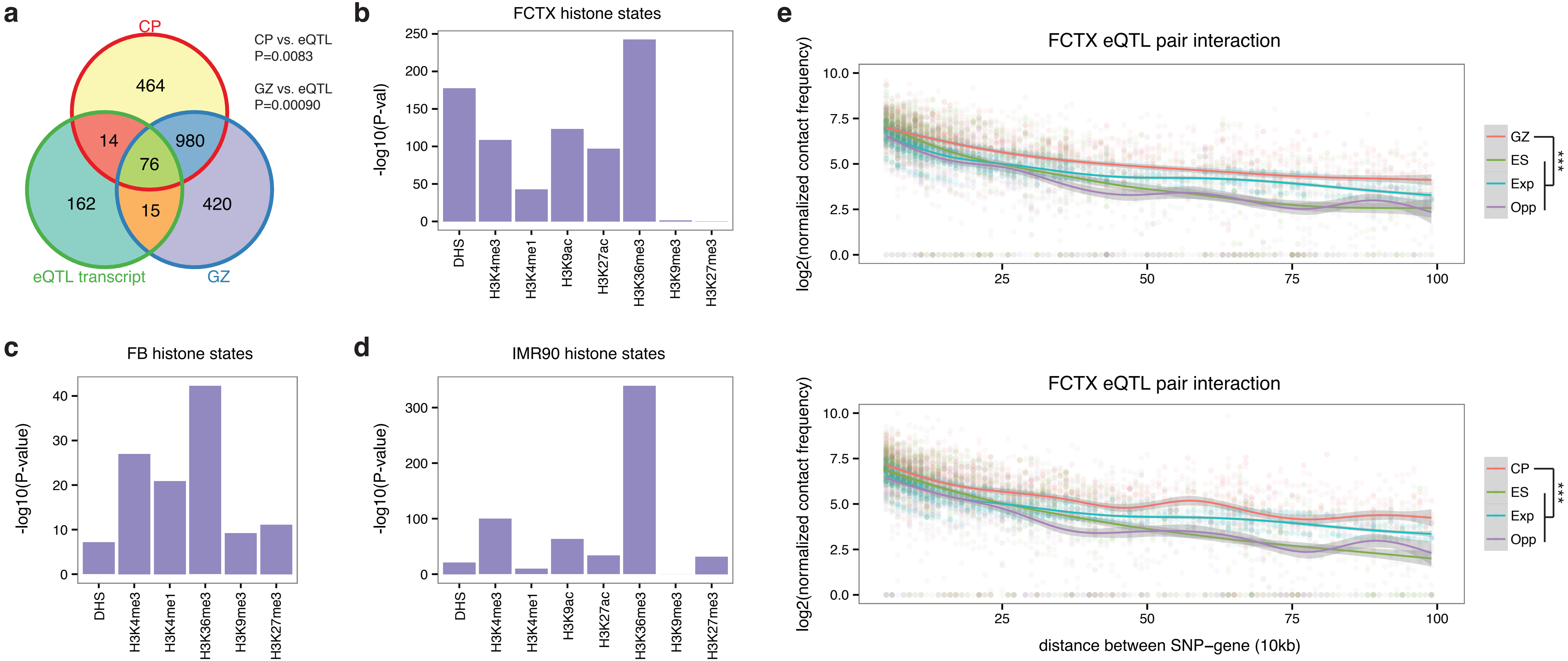


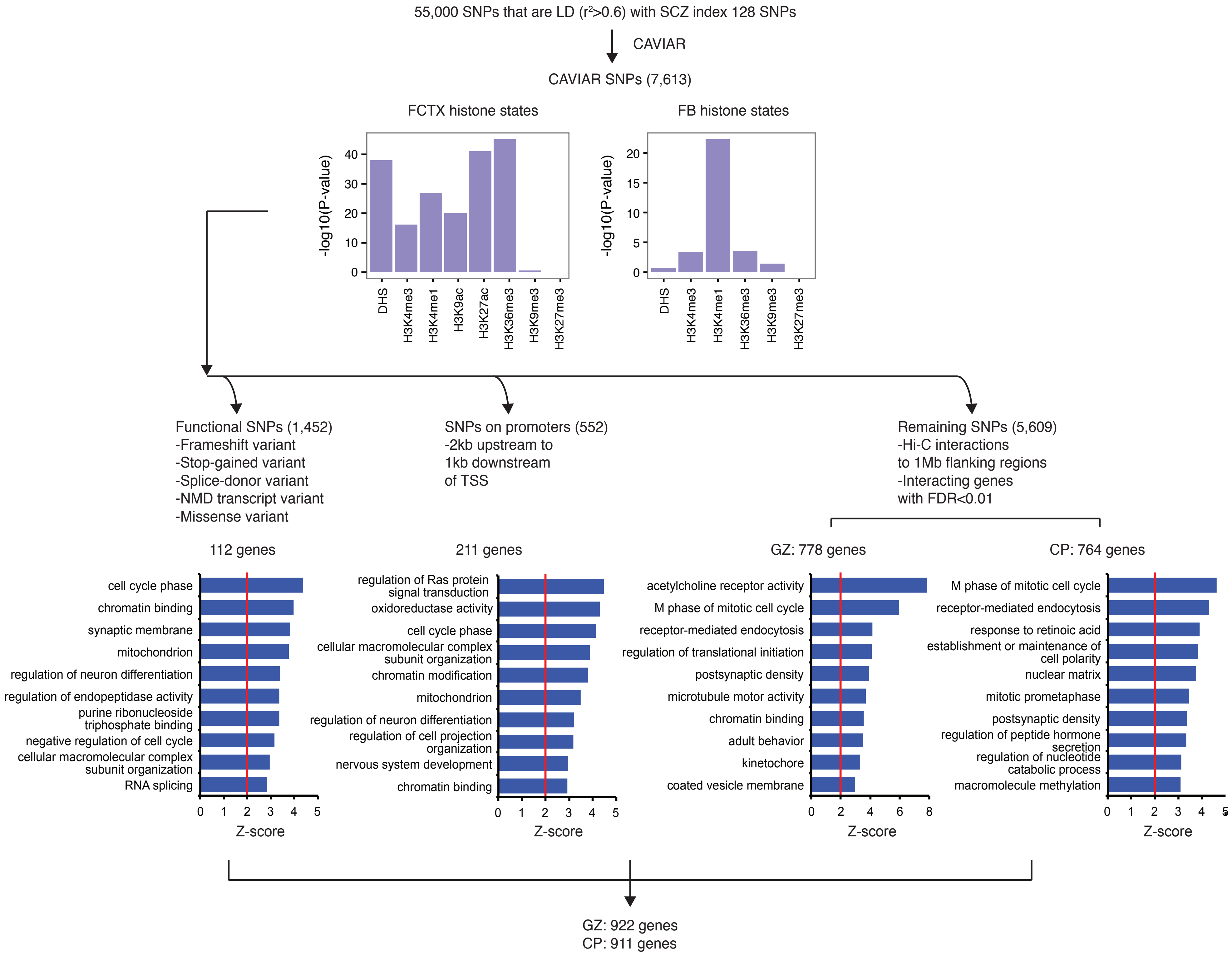
CP: intrachromosomal



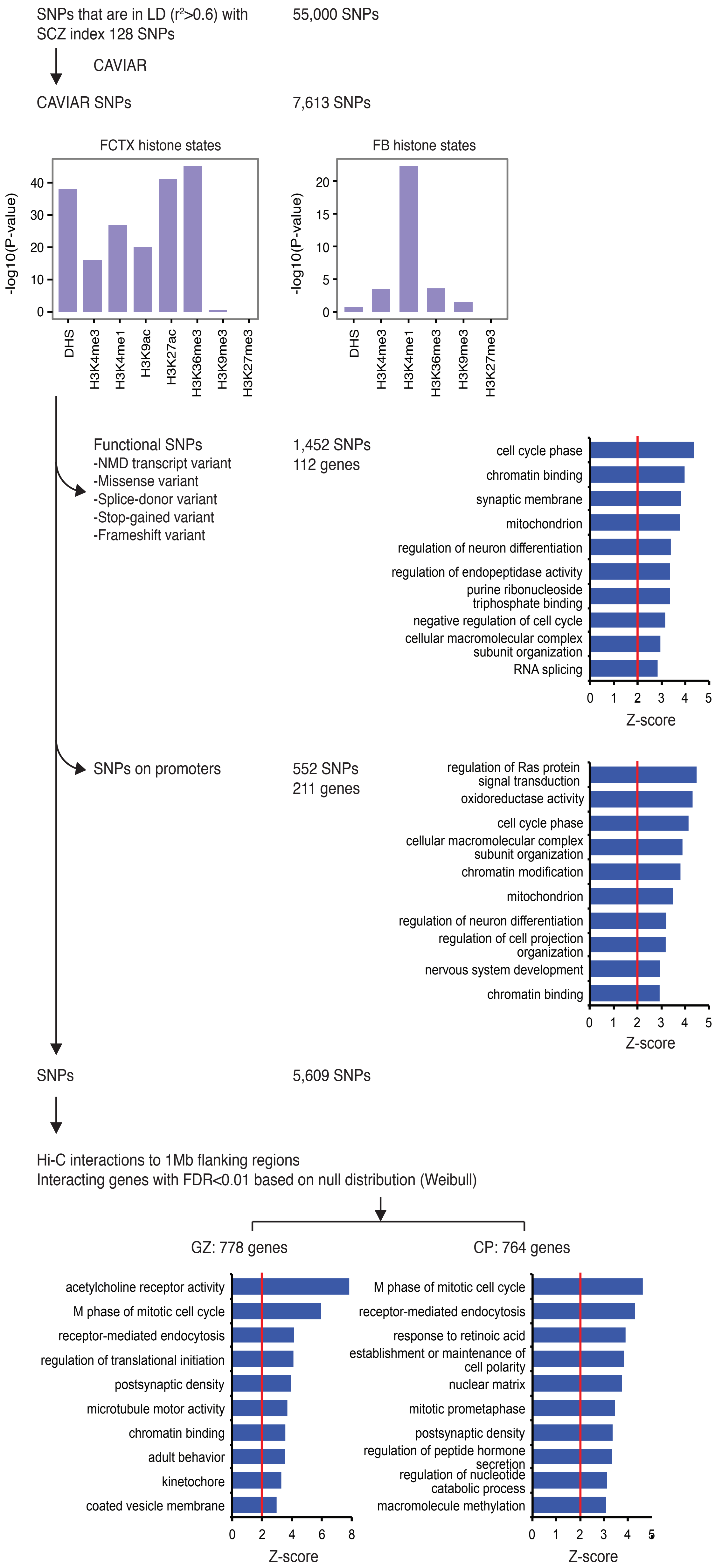

a
Histogram and theoretical densities
1Mb upstream
Density
weibull
lognormal
gamma
normal
Normalized interaction frequency
Empirical and theoretical CDFs
CDF
Normalized interaction frequency
Histogram and theoretical densities
100kb upstream
Density
Normalized interaction frequency
Empirical and theoretical CDFs
CDF
Normalized interaction frequency
b
GZ
CP
ES
Number of FDR<0.01 loci
Bins from the evolutionary locus (10kb)
c
Fraction
100%
75%
50%
25%
0%
CP
GZ
H3K4me3
H3K4me1
H3K36me3
others
GZ
CP
-log10(P-value)
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
Epigenetic marks

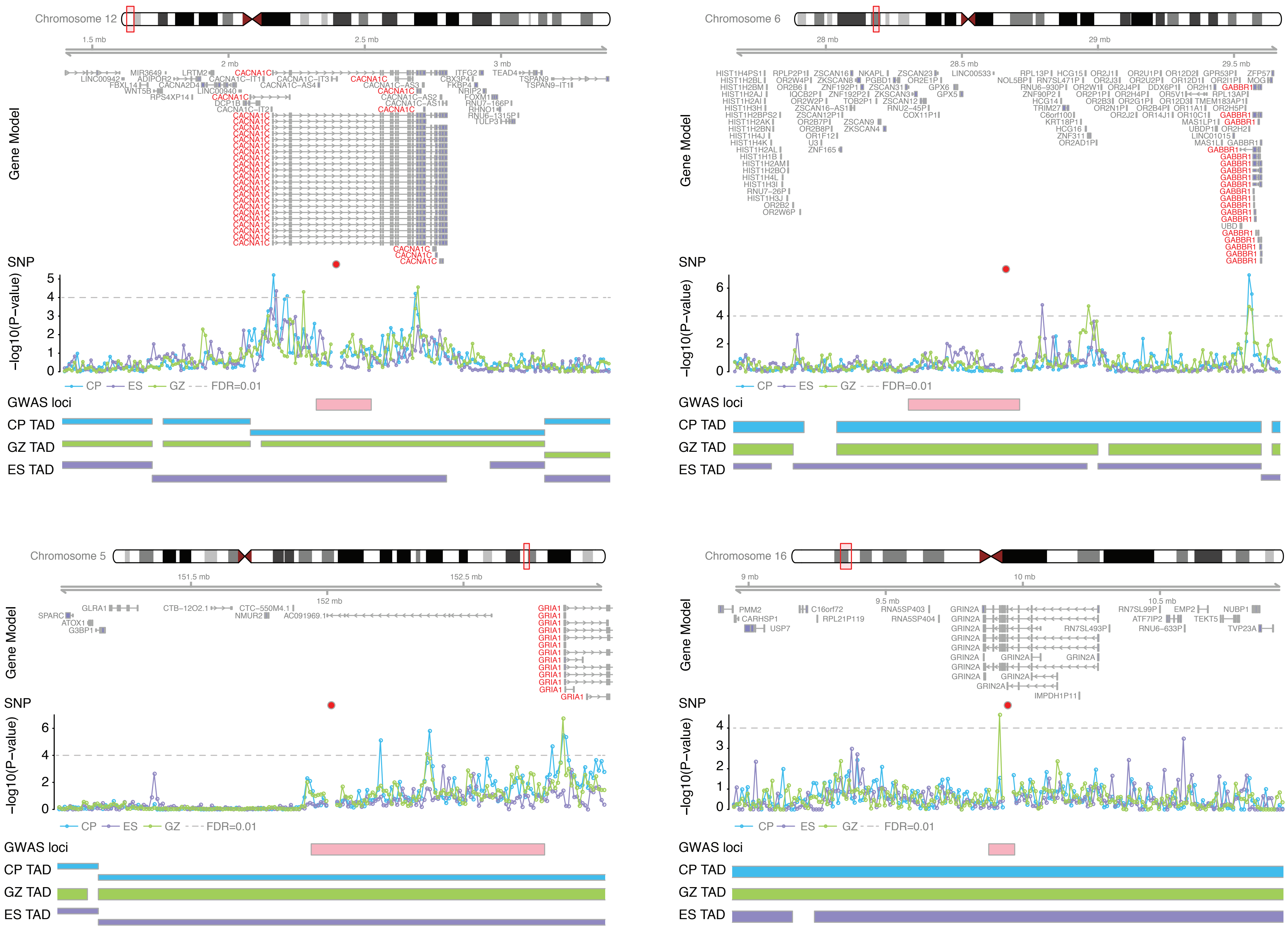

GZ
Tetrapods
P=0.0030
FDR=0.015
Amniotes
P=0.41
FDR=0.46
Mammals
P=0.52
FDR=0.52
Therians
P=0.16
FDR=0.27
Eutherians
P=0.40
FDR=0.46
Primates
P=0.00
FDR=0.00
GreatApes
P=0.0053
FDR=0.018
AfricanApes
P=0.026
FDR=0.065
Hominini
P=0.35
FDR=0.46
Human
P=0.11
FDR=0.22
CP
Tetrapods
P=0.0037
FDR=0.012
Amniotes
P=0.11
FDR=0.18
Mammals
P=0.51
FDR=0.51
Therians
P=0.22
FDR=0.27
Eutherians
P=0.094
FDR=0.18
Primates
P=0.00033
FDR=0.0033
GreatApes
P=0.0033
FDR=0.012
AfricanApes
P=0.43
FDR=0.48
Hominini
P=0.036
FDR=0.090
Human
P=0.13
FDR=0.19
frequency
lncRNA number

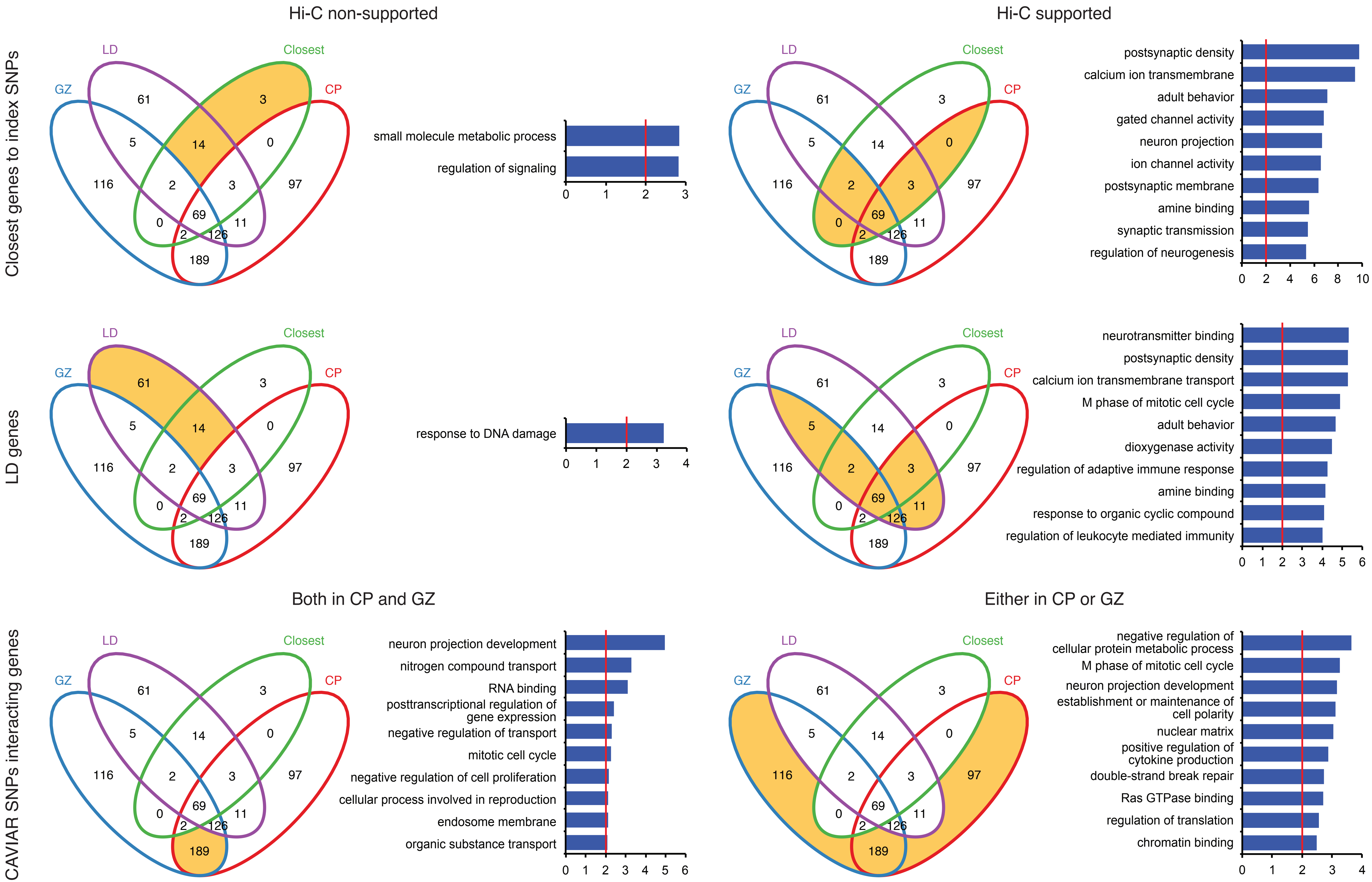

a
CP
464
14
980
76
162
15
420
eQTL transcript
GZ
CP vs. eQTL
P=0.0083
GZ vs. eQTL
P=0.00090
b
FCTX histone states
-log10(P-val)
250
200
150
100
50
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
c
FB histone states
-log10(P-value)
40
30
20
10
0
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
d
IMR90 histone states
-log10(P-value)
300
200
100
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
e
FCTX eQTL pair interaction
log2(normalized contact frequency)
10.0
7.5
5.0
2.5
0.0
25
50
75
100
GZ
ES
Exp
Opp
***
FCTX eQTL pair interaction
log2(normalized contact frequency)
10.0
7.5
5.0
2.5
0.0
25
50
75
100
CP
ES
Exp
Opp
***
distance between SNP–gene (10kb)

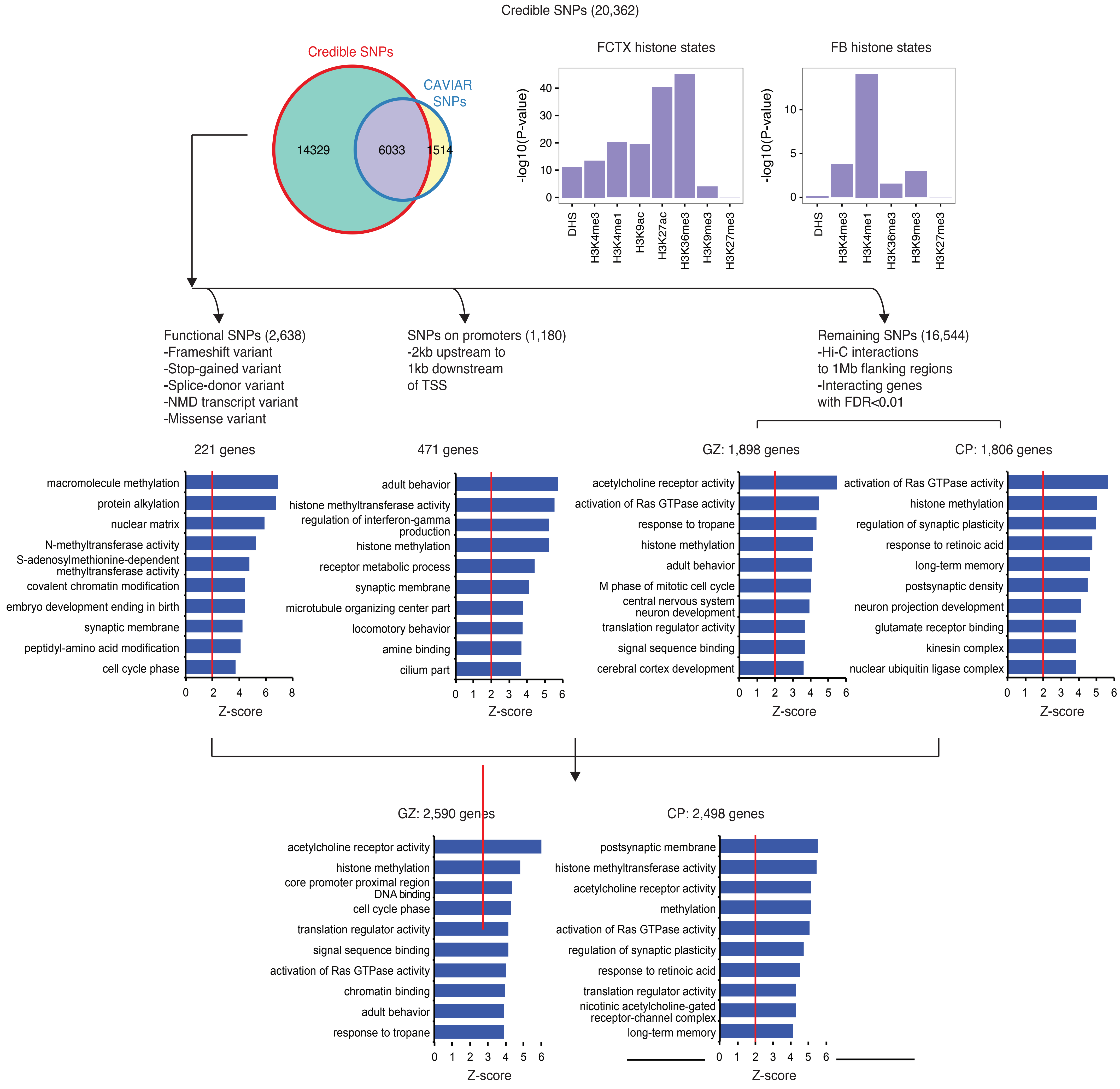


55,000 SNPs that are LD (r2>0.6) with SCZ index 128 SNPs
CAVIAR
CAVIAR SNPs (7,613)
FCTX histone states
FB histone states
-log10(P-value)
0
10
20
30
40
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
-log10(P-value)
0
5
10
15
20
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs (1,452)
-Frameshift variant
-Stop-gained variant
-Splice-donor variant
-NMD transcript variant
-Missense variant
SNPs on promoters (552)
-2kb upstream to
1kb downstream
of TSS
Remaining SNPs (5,609)
-Hi-C interactions
to 1Mb flanking regions
-Interacting genes
with FDR<0.01
112 genes
cell cycle phase
chromatin binding
synaptic membrane
mitochondrion
regulation of neuron differentiation
regulation of endopeptidase activity
purine ribonucleoside triphosphate binding
negative regulation of cell cycle
cellular macromolecular complex subunit organization
RNA splicing
0 1 2 3 4 5
Z-score
211 genes
regulation of Ras protein signal transduction
oxidoreductase activity
cell cycle phase
cellular macromolecular complex subunit organization
chromatin modification
mitochondrion
regulation of neuron differentiation
regulation of cell projection organization
nervous system development
chromatin binding
0 1 2 3 4 5
Z-score
GZ: 778 genes
acetylcholine receptor activity
M phase of mitotic cell cycle
receptor-mediated endocytosis
regulation of translational initiation
postsynaptic density
microtubule motor activity
chromatin binding
adult behavior
kinetochore
coated vesicle membrane
0 2 4 6 8
Z-score
CP: 764 genes
M phase of mitotic cell cycle
receptor-mediated endocytosis
response to retinoic acid
establishment or maintenance of cell polarity
nuclear matrix
mitotic prometaphase
postsynaptic density
regulation of peptide hormone secretion
regulation of nucleotide catabolic process
macromolecule methylation
0 1 2 3 4 5
Z-score
GZ: 922 genes
CP: 911 genes


SNPs that are in LD ($r^2$>0.6) with SCZ index 128 SNPs
55,000 SNPs
CAVIAR
CAVIAR SNPs
7,613 SNPs
FCTX histone states
-log10(P-value)
40
30
20
10
0
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
FB histone states
-log10(P-value)
20
15
10
5
0
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs
-NMD transcript variant
-Missense variant
-Splice-donor variant
-Stop-gained variant
-Frameshift variant
1,452 SNPs
112 genes
cell cycle phase
chromatin binding
synaptic membrane
mitochondrion
regulation of neuron differentiation
regulation of endopeptidase activity
purine ribonucleoside triphosphate binding
negative regulation of cell cycle
cellular macromolecular complex subunit organization
RNA splicing
0
1
2
3
4
5
Z-score
SNPs on promoters
552 SNPs
211 genes
regulation of Ras protein signal transduction
oxidoreductase activity
cell cycle phase
cellular macromolecular complex subunit organization
chromatin modification
mitochondrion
regulation of neuron differentiation
regulation of cell projection organization
nervous system development
chromatin binding
0
1
2
3
4
5
Z-score
SNPs
5,609 SNPs
Hi-C interactions to 1Mb flanking regions
Interacting genes with FDR<0.01 based on null distribution (Weibull)
GZ: 778 genes
acetylcholine receptor activity
M phase of mitotic cell cycle
receptor-mediated endocytosis
regulation of translational initiation
postsynaptic density
microtubule motor activity
chromatin binding
adult behavior
kinetochore
coated vesicle membrane
0
2
4
6
8
Z-score
CP: 764 genes
M phase of mitotic cell cycle
receptor-mediated endocytosis
response to retinoic acid
establishment or maintenance of cell polarity
nuclear matrix
mitotic prometaphase
postsynaptic density
regulation of peptide hormone secretion
regulation of nucleotide catabolic process
macromolecule methylation
0
1
2
3
4
5
Z-score


Chromosome 12
1.5 mb
2 mb
2.5 mb
3 mb
Gene Model
CACNA1C
SNP
−log10(P−value)
CP
ES
GZ
FDR=0.01
GWAS loci
CP TAD
GZ TAD
ES TAD
Chromosome 6
28 mb
28.5 mb
29 mb
29.5 mb
GABBR1
Chromosome 5
151.5 mb
152 mb
152.5 mb
GRIA1
Chromosome 16
9 mb
9.5 mb
10 mb
10.5 mb
GRIN2A


Hi-C non-supported
Hi-C supported
Both in CP and GZ
Either in CP or GZ
Closest genes to index SNPs
LD genes
CAVIAR SNPs interacting genes
LD
Closest
GZ
CP
61
3
5
14
0
116
2
3
97
0
69
11
2
126
189
small molecule metabolic process
regulation of signaling
0
1
2
3
postsynaptic density
calcium ion transmembrane
adult behavior
gated channel activity
neuron projection
ion channel activity
postsynaptic membrane
amine binding
synaptic transmission
regulation of neurogenesis
0
2
4
6
8
10
response to DNA damage
0
1
2
3
4
neurotransmitter binding
postsynaptic density
calcium ion transmembrane transport
M phase of mitotic cell cycle
adult behavior
dioxygenase activity
regulation of adaptive immune response
amine binding
response to organic cyclic compound
regulation of leukocyte mediated immunity
0
1
2
3
4
5
6
neuron projection development
nitrogen compound transport
RNA binding
posttranscriptional regulation of gene expression
negative regulation of transport
mitotic cell cycle
negative regulation of cell proliferation
cellular process involved in reproduction
endosome membrane
organic substance transport
0
1
2
3
4
5
6
negative regulation of cellular protein metabolic process
M phase of mitotic cell cycle
neuron projection development
establishment or maintenance of cell polarity
nuclear matrix
positive regulation of cytokine production
double-strand break repair
Ras GTPase binding
regulation of translation
chromatin binding
0
1
2
3
4


Credible SNPs (20,362)
Credible SNPs
CAVIAR SNPs
14329
6033
1514
FCTX histone states
-log10(P-value)
0
10
20
30
40
DHS
H3K4me3
H3K4me1
H3K9ac
H3K27ac
H3K36me3
H3K9me3
H3K27me3
FB histone states
-log10(P-value)
0
5
10
DHS
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
Functional SNPs (2,638)
-Frameshift variant
-Stop-gained variant
-Splice-donor variant
-NMD transcript variant
-Missense variant
SNPs on promoters (1,180)
-2kb upstream to
1kb downstream
of TSS
Remaining SNPs (16,544)
-Hi-C interactions
to 1Mb flanking regions
-Interacting genes
with FDR<0.01
221 genes
macromolecule methylation
protein alkylation
nuclear matrix
N-methyltransferase activity
S-adenosylmethionine-dependent methyltransferase activity
covalent chromatin modification
embryo development ending in birth
synaptic membrane
peptidyl-amino acid modification
cell cycle phase
0
2
4
6
8
Z-score
471 genes
adult behavior
histone methyltransferase activity
regulation of interferon-gamma production
histone methylation
receptor metabolic process
synaptic membrane
microtubule organizing center part
locomotory behavior
amine binding
cilium part
0
1
2
3
4
5
6
Z-score
GZ: 1,898 genes
acetylcholine receptor activity
activation of Ras GTPase activity
response to tropane
histone methylation
adult behavior
M phase of mitotic cell cycle
central nervous system neuron development
translation regulator activity
signal sequence binding
cerebral cortex development
0
1
2
3
4
5
6
Z-score
CP: 1,806 genes
activation of Ras GTPase activity
histone methylation
regulation of synaptic plasticity
response to retinoic acid
long-term memory
postsynaptic density
neuron projection development
glutamate receptor binding
kinesin complex
nuclear ubiquitin ligase complex
0
1
2
3
4
5
6
Z-score
GZ: 2,590 genes
acetylcholine receptor activity
histone methylation
core promoter proximal region DNA binding
cell cycle phase
translation regulator activity
signal sequence binding
activation of Ras GTPase activity
chromatin binding
adult behavior
response to tropane
0
1
2
3
4
5
6
Z-score
CP: 2,498 genes
postsynaptic membrane
histone methyltransferase activity
acetylcholine receptor activity
methylation
activation of Ras GTPase activity
regulation of synaptic plasticity
response to retinoic acid
translation regulator activity
nicotinic acetylcholine-gated receptor-channel complex
long-term memory
0
1
2
3
4
5
6
Z-score

Nenad Sestan, MD, PhD
Departments of Neuroscience, Genetics and Psychiatry
Program in Cellular Neuroscience, Neurodegeneration and Repair
Section of Comparative Medicine, Child Study Center
Kavli Institute for Neuroscience
Yale School of Medicine
333 Cedar Street, New Haven, CT  06510

Email: *nenad.sestan@yale.edu*
Web: *www.sestanlab.org*

October 30, 2015

**Collaborative applications:**

Nenad Sestan, MD, PhD (contact PI)
Yale School of Medicine
Title: *1/3 Integrative Genomic Analysis of Human Brain Development and Autism*

Daniel Geschwind, MD, PhD
University of California, Los Angeles
Title: *2/3 Integrative Genomic Analysis of Human Brain Development and Autism*

Matthew W. State, MD, PhD
University of California, San Francisco
Title: 3/*3 Integrative Genomic Analysis of Human Brain Development and Autism*

**Funding Opportunity:**
RFA-MH-16-230: Multi-scale Molecular Profiling of Brains from Psychiatric Cohorts (Collaborative R01)

**Disciplines involved:**
Autism, Genomics, Genetics, Human brain development

**Assignment of application:**
Institutes/Centers: National Institute of Mental Health- NIMH
Scientific Review Groups: As appropriate within the assigned institute

**Individuals who should NOT review this application:** None

The proposed research will perform time, region and cell type specific molecular profiling of control and ASD brains. We will analyze and integrate these datasets to identify regional and developmental, and ASD related processes to gain insight into underlying mechanisms. Finally, we will perform integrated analysis of germ-line ASD variations to characterize causal enrichments in developmental periods, brain regions, and cell types to better characterize the mechanisms by which genetic variation in humans alters brain development and function in health and disease.

We have received permission from Dr. Geetha Senthil to submit this application with a budget exceeding $500,000 per year.

Thank you for your consideration. Please contact me should any questions arise.

Yours sincerely,

Daniel Geschwind                Matthew W. State                Nenad Sestan

## DATA-RESOURCE SHARING PLAN

### 1) Data Sharing Plan:

### Data

RNA-seq, ChIP-seq and Hi-C data will be generated using next-generation sequencing (NGS) on human postmortem brain samples as part of Aim 1. NGS will be conducted on the Illumina HiSeq 2500 platform. FASTQ/BAM files, and gene- and isoform level expression files will be generated. Standard quality control measures, as well as fastqc analysis prior to sequence alignment, will be implemented. The RNA-seq, ChIP-seq and Hi-C data will be deposited into an NIH approved repository such as the GEO database (*www.ncbi.nlm.nih.gov/geo/*).

Overview and descriptions of each dataset will be provided alongside the relevant dataset. Links to all of these resources will be provided in all publications arising from this project. In addition, we will follow policies adopted by the psychENCODE consortium (see *www.psychencode.org*).

### Data dissemination

We will comply with the NIH Grants Policy Statement, revised 10/01/2011, SECTION: 8.2.3 on Sharing Research Resources http://grants.nih.gov/grants/policy/nihgps_2011/nihgps_ch8.htm#_Toc271264950, relating to the distribution of unique research resources produced with DHHS funding. In particular, we will endeavor to publish all findings in peer-reviewed journals as soon as possible and provide all raw NGS data to the Gene Expression Omnibus (GEO) database. In addition, we will follow policies adopted by the psychENCODE consortium (see *www.psychencode.org*). Because the community standard for early data release is evolving, this data sharing may be reassessed at the end of each funding period. The data and renewable reagents generated will be submitted and released to the public in accordance with the following schedule.

### Data sharing within the collaborative project

All data generated by the Geschwind lab and Sestan lab as part of this project, as well as the data analysis performed by the psychENCODE DAC members (M. Gerstein and Z. Weng), will be immediately shared among the groups.

### Computational analysis tools

All computer programs implementing the algorithms generated in this project will be available to the scientific community. The Park lab has previously released several popular tools for the community with full documentation and source code, either on the lab website or in the centralized locations such as the Bioconductor or CRAN repositories for R packages.

### Sharing of all research data through scientific meetings and journals

Research data will be presented at the yearly Society for Neuroscience and Genetics meetings and submitted to peer-reviewed journals. Over the course of the data collection period, we aim to publish multiple papers reporting selected results from the analyses described in this application.

### 2) Sharing Model Organisms and Reagents:

### Renewable reagents

No model organisms or renewable reagents are expected to be generated. Research resources generated with funds from this grant will be freely distributed, as available, to qualified academic investigators for non-commercial research. Material transfers would be made with no more restrictive terms than in the Simple Letter Agreement or the UBMTA and without reach through requirements.

**Timelines**

| Resource | **Deadlines for Resource Submission** *[Please include hard release dates]* | **Deadlines for Releasing Resources for Distribution** *[Please include hard release dates]* |
|---|---|---|
| Freshly drawn biospecimens (e.g., blood, skin punch, olfactory biospy) | Not applicable. | Not applicable. |
| Established source cell lines for iPSC reprogramming (e.g., fibroblast cell lines) | Not applicable. | Not applicable. |
| Reprogrammed cell lines (e.g., iPSCs, iNSC, iGPC, iOPC) | Not applicable. | Not applicable. |
| Phenotypic/clinical assessments | Not applicable. | Not applicable. |
| GWAS data with accompanying study documents and phenotype data | Not applicable. | Not applicable. |
| Sequence data (e.g., WGS, exome) | Not applicable. No WGS/exome data will be generated under this grant and the data used is covered by data sharing plans associated with their proposal. | Not applicable. |
| Other genetic/genomic data (e.g., methylome, transcriptome, CNV) | RNA-seq data<br>Approximately 200 samples (5 developmental periods, 6 brains/ period, 3 regions, 2 cell types) from control brains, 240 from control and ASD (20 brains/condition, 3 regions, 2 cell types) will be profiled with RNA-seq using the Illumina HiSeq 2500 platform.<br>Data will be submitted at 6 month intervals after data cleaning, to be completed by the project year 4 end date of 6/30/2020.<br>ChIP-seq data<br>Approximately 400 samples (5 developmental periods, 6 brains/ period, 3 regions, 2 cell types, 2 histone marks) from control brains, 480 from control and ASD (20 brains/condition, 3 regions, 2 cell types,2 histone marks) will be profiled by ChIP-seq using the Illumina HiSeq 2500 platform.<br>Data will be submitted at 6 month intervals after | Data will be released, upon publication but not later than 6 months after last data set is received, and not later than 6 months after project year 4 ends (6/30/2020). |

| | | |
|---|---|---|
| | data cleaning, to be completed by the project year 4 end date of 6/30/2020.<br>Hi-C data<br>For the developmental series, we will analyze 5 stages, 6 samples per stage, for 3 regions (frontal, temporal, and basal ganglia), a total of 90 tissue samples, each sorted into neurons and non-neuronal nuclei for a total of 180 Hi-C libraries. For ASD analysis, we will perform Hi-C on 2 cortical regions and basal ganglia from 20 matched control and 20 ASD individuals as well as 5 dup15q subjects. In total, 330 Hi-C libraries will be made and analyzed, as we leverage multiple controls from the lifespan analysis to also serve as controls for the ASD comparisons.<br>Data will be submitted at 6 month intervals after data cleaning, to be completed by the project year 4 end date of 6/30/2020. | |
| Analyzed data | RNA-seq, ChIP-seq and Hi-C data<br>Gene- and isoform level expression values, ChIP-seq peaks and chromosomal interactions as well as any other relevant files, such as a list of differentially expressed genes.<br>Data will be submitted at 6 month intervals after data cleaning, to be completed by the project year 4 end date of 6/30/2020. | Analyzed data is released with associated biospecimens, upon publication but not later than 6 months after last data set is received, and not later than 6 months after project year 4 ends (6/30/2020). |

**MAJOR EQUIPMENT**

**Equipment available within the Sestan laboratory**

Major equipment for genomics, molecular biology, histology, tissue culture, mouse transgenesis and imaging, are available including: a Bio-Rad QX100 digital droplet PCR system; four MJ Research PTC-200 PCR machines; a Thermo Scientific Nanodrop 1000 UV-Vis spectrophotometer and 3300 fluorospectrometer; an Agilent TapeStation, a Zeiss Stemi SV6 dissecting microscope; an inverted Zeiss AxioVert 25 microscope; a Leica CM-3050S cryostat system, a Leica VT1000 S vibrating blade microtome; three -80ºC freezers, of which one is used exclusively for storing human and non-human primate brain specimens; four -20ºC freezers; three SterilGard laminar flow hoods; two tissue culture rooms with four Heraeus HeraCell $CO_2$ tissue culture incubators; a room with a BTX ECM 830 electroporator for *in utero* electroporation and rodent surgery table; and electrophoresis equipment. In addition, the Sestan laboratory has a Turner Designs TD-20/20 luminometer, a Corning pH meter, a Speed-Vac, four Eppendorf microcentrifuges, two water baths, four Fisher refrigerators, a Zeiss 135/Eppendorf Transman mouse oocyte and blastocyst microinjection set-up for making transgenic mice, an inverted microscope with micromanipulators for cell collection, a Sutter Instrument P-87 micropipette puller, a Sorvall low speed centrifuge, an Eppendorf high speed centrifuge, a 37ºC shaker, a Beckman-Coulter ultracentrifuge, a scintillation counter, and access to a darkroom equipped with an X-ray film processor. The Sestan laboratory has access to and shares maintenance and service responsibilities for the multiphoton and electron microscopy facility within the department.

Laboratory computer resource: The lab's computing infrastructure is partitioned into a private and a public network. The entire infrastructure is fully gigabit capable and is connected to the Yale backbone and Yale High-Performance-Computer (HPC) via gigabit optic fibre; the network architecture was designed with computing efficiency and network security in mind. The private network consists of individual laptops, desktops and workstations, as well as communal computational servers, dumb terminals, a central fileserver, a consolidated NAS, and printers. The public network consists of numerous production webservers that are either real or virtual machines, including the *www.humanbraintranscriptome.org* database. The department has a full-time ITS administrator maintaining the network and computers.

**Common equipment within the Department**

Zeiss LSM multi-photon microscope, a JOEL electron microscope, a Sorvall low speed centrifuge, a 37 ºC shaker, a Beckman-Coulter ultracentrifuge, a scintillation counter, and a darkroom equipped with a X-ray film processor.

**Equipment available within the Geschwind laboratory**

The Geschwind laboratory is equipped with an Ion Torrent Personal Genome Machine for high throughput automated sequencing (with Ion One Touch System for template preparation), a Roche LightCycler 480 Real-Time PCR Station, a Genetics MicroSystems/Affymetrix 418 microarray reader, an Agilent 2100 Bioanalyzer, a NanoDrop spectrophotometer, a Bio-Rad GS Gene Linker UV chamber, an Amaxa Nucelofector II transfection device, a UVP BioSpectrum AC imaging system, a Barnstead NANOpure DIamond ultrapure water system, a Qiagen QIAcube robotic workstation, a BioTek Synergy 2 multi-mode microplate reader, a Leica 1950 cryostat, and a Tecan 4800 automated microarray hybridization station. For dedicated tissue culture work, there are two biosafety cabinets and six CO2 incubators within two dedicated tissue culture rooms adjacent to the Geschwind lab. For high throughput assays, the laboratory shares a Victor/PE fluorescent plate reader and an Illumina Bead Array system. The Center for Neurobehavioral Genetics (Geschwind, co-director) in the Gonda building provides access to an Illumina BeadLab SNP genotyping system capable of performing over 800,000 genotypes per day (as part of the Southern California Genotyping Consortium, see Resources and Facilities page) and to a Sequenom MassARRAY Compact System with Server and RT workstation, including the MassARRAY Analyzer Compact MALDI-TOF Mass Spectrometer, the MassARRAY RT Workstation for real-time data analysis, Nanodispenser RS 1000, and Liquid Handler Station. Also available within the UNGC are three Illumina Hi Seq 2000 sequencers. Core equipment in the Human Genetics Department in the Gonda building that is used regularly by the Geschwind laboratory includes an Affymetrix Chip station for array processing and analysis. Also available for high throughput experimentation are a Beckman Biomek 1000 robot, a Catalyst PCR robot, a Packard robot, and a 96 channel pipetting robot with motorized stage, four MJ

Research Quadra thermal cyclers, a 96-well Pyrosequencer, and a Packard Multiprobe II robot, all connected to a data management system. The adjacent Mental Retardation Research Center and the Brain Research Institute both have sophisticated microscopy and imaging cores including a Zeiss 4-laser confocal microscope.

**Equipment available within the State, Sanders and Willsey laboratories**

Major equipment for genomics, molecular biology, and tissue culture are available including: An Eppendorf epMotion 5075LH automated pipetting system; An Eppendorf epMotion 5070 automated pipetting system with the ability to pick single wells in a preprogrammed pattern; An Eppendorf New Brunswick U725 Innova Ultra-low Temperature Freezer for sample storage; An Eppendorf Model 5810R refrigerated centrifuge; An Eppendorf Model 5804 centrifuge; A Micronic bar code scanner including a BioMicrolab Tracker Table Model Scanner (3 racks), a Wireless Handheld Scanner, and 1D & 2D barcode reader w/ USB connection; An BioMicroLab XL20 Tube Handler for automated sample re-array, sample weighing, volume checking, barcode reading, and data collection for Micronic tubes in a 96 tube rack; 2 Tetrad 2 PCR instruments, each with a (swappable) 384-well and 96 well-reaction module; 9 BioRad icycler PCR instruments, each with a (swappable) 384-well and 96 well-reaction module; An ABI-7900HT fast real time PCR system with 96 and 384 well capacity; A Covaris S2 DNA Sonicator; A BioTek Synergy HT Fluorometer for PicoGreen DNA concentration assessment; A Thermo Scientific (ND8000) 8-well Nanodrop 8000 Spectrophotometer; BioTek Elx800 Plate Reader; An Agilent 2100 Bioanalyzer; 2 Invitrogen E-Gel iBase electronic gel readers; A BioRad ChemiDoc MP System for gel imaging.

Additional equipment and supplies can be found on the Resources and Facilities page.

## PROTECTION OF HUMAN SUBJECTS

This study will use human postmortem specimens and data obtained in the following forms:

1. Post-mortem brain tissue from clinically unremarkable individuals (controls) or individuals who had ASD.
2. Publicly available genomic datasets on de-identified normal control samples from Gene Expression Omnibus (GEO), dbGaP, the National Database for Autism Research (AGRE) and the Simons Simplex Collection (SSC).
3. De-identified DNA from SSC and AGRE samples available through the Rutgers University Cell and DNA Repository (RUCDR)).

According to the U.S. Department of Health & Human Service Code of Federal Regulations Section 46.102(f):

> Human subject means a *living* individual about whom an investigator (whether professional or student) conducting research obtains
>
> (1) Data through intervention or interaction with the individual, or
> (2) Identifiable private information.

Therefore, as the post-mortem tissue is not from living individuals, and the genomic data and DNA samples have no identifiable information, these specimens are not considered human subjects.

### Human post-mortem tissue

The research outlined in this grant application will be conducted using control post-mortem human brain specimens available in Daniel Geschwind's lab at UCLA and Nenad Sestan's lab at Yale. Additional control and ASD postmortem brains may be obtained through arrangements with the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, the Harvard Brain Tissue Resource Center and the Oxford Brain Bank (Autism Tissue Program). A table of available brain specimens is provided in the Resources and Facilities section.

All specimens have or will be collected after parental or next of kin consent and with approval of the institutional review boards (IRBs). Appropriate written informed consent has or will be obtained and all available non-identifying information recorded. The tissue, linked information, and consent forms obtained by these institutions has been properly deposited and will continue to be deposited into brain banks according to IRB guidelines or similar protocols approved by each institution's medical ethical committee for the studies proposed in our application. This information can only be accessed by the brain bank/tissue repository coordinator. Furthermore, de-identifiable data is doubly encoded and saved on separate computer systems before release to the Geschwind and Sestan laboratories. The handling of tissue has been performed in accordance with ethical guidelines and regulations for the research use of human brain tissue set forth by the NIH (http://bioethics.od.nih.gov/humantissue.html), the UK Health Department's Human Tissue Authority code of practice (http://www.hta.gov.uk/legislationpoliciesandcodesofpractice/codesofpractice/code9research.cfm) and the WMA Declaration of Helsinki (www.wma.net/en/30publications/10policies/b3/index.html). To ensure the highest standard for data protection, no personal identifying information will be collected by the Geschwind and Sestan laboratories nor will it be accessible to any member of the laboratory. All of the work will be done in accordance with the new Health Insurance Portability and Accountability Act (HIPAA), which safeguards the health information of individuals obtaining healthcare in the United States. We are in full compliance with these regulations.

The acquisition of human fetal brain tissue is in accordance with NIH guidelines in that (1) the tissue is human fetal tissue obtained in a spontaneous or induced abortion or pursuant to a stillbirth, (2) the tissue was donated anonymously for research purposes and that the identity of the individual who donated the tissues can never be determined, (3) no investigator part of the research has any decisions as to the timing, method, or procedures used to terminate the pregnancy, and (4) no investigator part of the research is the donor's attending physician.

For quality control measure, we will use this non-identifying medical history of the subject from which the brain tissue will be obtained, or the mother's medical history in the case of pre- and neonatal specimens. We will review available ante mortem information, including: gender, ethnicity, weight, cause of death, medications, Apgar score, and relevant medical conditions. This information will only be used to exclude some postmortem specimens from the study, such as those from individuals with a known history of drug or alcohol abuse. Information showing specific agonal conditions, including coma, hypoxia, pyrexia, seizures, severe dehydration, hypoglycemia, multiple organ failure, head injury, and ingestion of neurotoxic substances at time of death will also be grounds for exclusion of the postmortem tissue.

Mount Sinai

**Schahram Akbarian,**
M.D., Ph.D.

Chief
Psychiatric Epigenomics Division
Department of Psychiatry

Friedman Brain Institute
Hess CSM, 9-105
1470 Madison Avenue, 9-
New York, NY  10029-6574

Tel : 646 627 5529
Email: Schahram.akbarian@mssm.edu

October 28, 2015

Daniel Geschwind, MD, PhD
Nenad Sestan, MD, PhD
Matthew State, MD, PhD

Dear Dan, Matt and Nenad,

This letter confirms my willingness to serve as a consultant on your application, "*Integrative Genomic Analysis of Human Brain Development and Autism*".

I have enjoyed working closely with you on the PsychENCODE project.  The current project aims you outline will deliver important data and are an important continuation of your prior work and will interdigitate nicely with my own psychENCODE studies.

I am a Professor of Psychiatry and of Neuroscience at the Icahn School of Medicine at Mount Sinai. My lab studies histone modifications and variants, chromosomal loopings and other building blocks of the epigenome in postmortem human brain tissue collected across the lifespan, including potential alterations in psychiatric diseases such as schizophrenia and depression.

My lab was one of the first groups to FACS neuronal and non-neuronal nuclei from the postmortem human brain tissue and perform ChIP-seq. As part of this work we have improved our methodology to efficiently collect these cell types from the human brain tissue.  I would be happy to provide any advice on the methodology.

I commit to the following general obligations for the grant:

- Serve as a consultant with reimbursement of $2000/year
- Make myself available to provide advice and assistance
- Attend teleconferences as needed and occasional face-to-face meetings
- Participate in project video conferences that could benefit from my expertise

As noted above, the specific expertise I will provide is in the interpreting your findings to our neurodevelopmental work as well as on FACS sorting of different cell types.

Sincerely,

## PROJECT NARRATIVE

Autism Spectrum Disorder (ASD) is a group of complex disorders of brain development, characterized by impairments in social communication and restricted or repetitive behavior or interests. For most patients the genetic causes and molecular underpinnings of ASD are not known. The work proposed in this application will help determine which parts of the developing brain and molecular processes are involved in ASD, thus improving our understanding of the disease and contributing to the development of diagnostic tests to detect such changes.

## PROJECT MANAGEMENT PLAN

### Rationale

This proposal combines the expertise of three sites into a single, highly integrated effort aimed at integrating multidimensional datasets to better understand the etiology of neurodevelopmental disorders. The proposal includes three groups led by Nenad Sestan at the Yale School of Medicine (Application 1/3), Daniel Geschwind at the University of California, Los Angeles (UCLA), and Matthew State at the University of California, San Francisco (UCSF).

Given the complexity of human neurodevelopment and the genetics of autism spectrum disorders (ASD) we believe that integrating the respective expertise of these groups offers the best opportunity to utilize the plethora of existing multidimensional genomic and neurobiological data to help elucidate the origin of neurodevelopmental disorders. This organizational structure elaborated below combines the expertise and capabilities of the PIs, Daniel Geschwind, Nenad Sestan, and Matthew State and the key investigators, Jason Ernst (UCLA), Mark Gerstein (Yale), Stephan Sanders (UCSF) and Zhiping Weng (University of Massachusetts), and Jeremy Willsey (UCSF), to create a project well beyond the capacity of each individual group. An assurance of effectiveness in these interactions is that Drs. Gerstein, Geschwind, State, Sanders, Sestan, Weng and Willsey have a substantial history of working collaboratively in various combinations and have several joint publications. Additionally, each collaborator has significant experience in large multi-institutional grants of this nature (see Project Management Plan for further details). These efforts have never required a formal decision-making or grievance structure and we anticipate that the current work will continue in this manner. However, in case of unresolved conflict between the PIs, a decision-making and conflict resolution plan is detailed below. We recognize, given the novel nature of the work proposed in this grant, that we will need to continually re-evaluate our data and approaches. Given our history, we do not anticipate that this will present substantial difficulties. We have included this multiple PI management plan to clarify the relationships and role for each group.

### Integration of scientific research procedures across different elements of the project

*Expertise*:

These groups bring together a range of expertise including human and mouse genetics, functional genomics, bioinformatics, biostatistics, human neuroanatomy, developmental neurobiology, and gene discovery in ASD. These combined skills, which are needed to computationally identify and functionally analyze the role of ASD-associated mutations, would be impossible to find in a single site, making a multi-site collaborate approach a necessity.

The Yale group (PI, Nenad Sestan) has been at the forefront of creating and analyzing spatial and temporal maps of coding and non-coding elements in developing and adult human brain using high-throughput genomics; and in leveraging these to identify and study specific molecular mechanisms critical to the development of the human brain and pathogenesis of ASD and intellectual disabilities. The Sestan lab has a collection of almost 200 fresh-frozen post-mortem human brain specimens. In addition, they have generated RNA-seq, ChIP-seq and DNA methylation datasets using this tissue as a part of the BrainSpan project (*www.brainspan.org*) and psychENCODE project (*www.psychencode.org*). Furthermore, the Sestan lab has been characterizing the evolution and function of coding and non-coding elements, and regulatory networks influencing neurodevelopmental processes, bringing together data generation and bioinformatics, and model systems approaches.

The UCLA group (PI, Daniel Geschwind) has been at the forefront at functional genomics and transcriptional profiling in human brain, including ASD, where they identified a core pattern of transcriptional dysregulation. He will continue his close collaborations with co-Investigators, Jason Ernst, UCLA and Shayam Prabhakar, GIS Singapore, as part of the UCLA site contributions. In PsychEncode I, Geschwind and Prabhakar have been working collaboratively to define chromatin states using Chip-seq to uncover the regulatory factors underlying transcriptome changes in ASD and in normal controls. Dr. Prabhakar has significant experience in genomic data analysis, specifically in epigenetic mechanisms of gene regulation, and has developed powerful new methods for histone QTL analysis (hQTL). Jason Ernst, who has played an important analytical role in ENCODE, defining chromatin states based on histone marks, is collaborating on the integrative analysis of the Hi-C data with histone marks and methylation data. Geschwind has also worked extensively with Sestan and State labs to integrate transcriptional profiling in control (Brainspan) and ASD brains with genetic and epigenetic data, including several published studies, and contributes a collection of >50 cases and 45 controls with good quality RNA, and tissue level RNAseq data. His laboratory will continue this productive collaboration by conducting Hi-C on the sorted neuronal and non-neuronal nuclei from the core set of tissues compiled by Sestan and his labs as outlined in the Research Plan. Geschwind has also characterized gene regulatory networks, and in conjunction with Ernst and Prabhakar (and the Yale group led by Gerstein) will integrate transcriptional networks, with Hi-C and epigenetic data.

The UCSF group (PI, Matthew State) has extensive expertise in the identification of ASD risk variants through the genomic analysis of large ASD cohorts including the Simons Simplex Collection, and the use of these variants to identity points of spatiotemporal convergence in the human brain. Dr. State has pioneered the detection of rare genetic variation through genomic analysis as a means to identify the genes involved in neurodevelopmental disorders. He has led a number of multi-site collaborations applying this approach to large cohorts of ASD and Tourette Syndrome cases. His lab's research into the association between ASD and de novo copy number variants (CNVs) in SNP genotyping data and de novo loss of function (LoF) mutations in exome sequencing data has provided considerable insight into the genomic architecture of ASD. This insight has enabled specific genomic loci (e.g. 7q11.23 duplications), and specific genes (e.g. *SCN2A*) to be associated with ASD. Extending this to larger cohorts, and applying the statistical methods developed in collaboration with the Roeder lab, has led to the identification of over 50 ASD associated genes that form the basis of aim 1 of this proposal. In collaboration with the other PIs, Dr. State's lab developed the methods of spatiotemporal co-expression analysis that led to the identification of mid-fetal, prefrontal cortex as a key point of convergence in ASD etiology. Recently Matt has been a lead PI on a project performing whole-genome sequencing (WGS) on over 500 ASD families from the SSC. In addition his lab is leading a project to identify further mutations in 250 genes with moderate ASD association in 17,000 samples using molecular inversion probes (MIPs). Dr. State will co-ordinate this proposal between the collaborating sites.

Dr. Sanders trained as a graduate student and postdoc in Dr. State's lab. He led the analysis of both the CNV and exome work described above, including development of genomic analysis methods to identify de novo mutations (e.g. https://sourceforge.net/projects/cnvision/). His own lab has continued to work on genomic analysis methods including detecting de novo insertion/deletions in exome data and the integration of CNV and exome data to maximize gene discovery in ASD. In addition, analysis of gene expression data, in particular it's relationship to sex bias in ASD, is a major interest of his lab and he is leading the analysis of sexual dimorphism in human brain development as part of the BrainSpan Consortium in collaboration with Dr. Sestan. Finally, Dr. Sanders is the Director of the UCSF Psychiatry Department Bioinformatics Core that has developed cloud-based analysis pipelines for genomic data, including the detection of de novo mutations in whole genome sequencing data. His lab will contribute towards the detection and analysis of non-coding mutations in Aim 3.

Dr. Willsey also trained as a graduate student and postdoc in Dr. State's lab. He led the analysis of the spatiotemporal co-expression networks that form a foundation for Aim 3 of this proposal and has continued to develop these approaches. Dr. Willsey will lead the spatiotemporal analysis components of Aim 3.

*Division of labor:*
The groups will undertake the following roles:

- *Specific Aim 1*: Drs. Geschwind and Sestan will lead the molecular profiling using RNA-seq, ChIP-seq and Hi-C of developmental control and ASD brains.
- *Specific Aim 2*: Drs. Ernst, Gerstein, Geschwind, Prabhakar, Sestan and Weng will lead the integration and analysis of multidimensional data.
- *Specific Aim 3*: Drs. Sanders, State and Willsey will lead integrated analysis of germ-line ASD variations leveraging Aims 1 and 2 data

In addition, Drs. Geschwind, Sestan and State will co-ordinate across the three aims to ensure consistency in experimental approach and objectives.

*Integration of data:*
While each group will be responsible for their own set of specific aims, the success of the project depends on careful integration of the work being done by each group, as well as open communication and sharing of data. For example, the information needed to complete specific aim 3 will come from the data derived in earlier aims. And while data generated by the three groups in this project will be very valuable individually, they have the potential to be paradigm shifting when integrated together. Therefore sequence data as well as material and

techniques being generated by one group will be made available to the other group, and the analysis and interpretation of the data freely discussed at the monthly meetings described below.

## Administrative Structure

*Leadership:*

All PIs will have equal scientific leadership responsibilities. They will all be responsible for implementing the scientific and leadership plans at their respective institutions, and ensuring compliance with applicable federal and state laws, regulations and policies, biosafety, and data security. Nenad Sestan will serve as the contact PI for the project and have overall responsibility for coordinating the effort between participating laboratories and institutions. He will also be responsible for fiscal and administrative management, including all communication with the NIH, and submission of required progress reports to the NIH. Nenad Sestan has extensive experience in leading and participating multi-institutional projects. Dr. Sestan already has ongoing collaborations with all the labs involved in this proposal and has experience leading collaborative studies, including BrainSpan and psychENCODE. Daniel Geschwind and Matthew State also have considerable experience leading collaborative studies of this nature.

*Scientific Coordination:*

Implementation of this project will also take advantage of ongoing collaborations and meetings between the UCLA, UCSF, and Yale groups. *Monthly* joint conference calls will occur between participating labs on the first Monday of the month at 10am. Each group will present their progress on research project related to the genetics and neurobiology of autism, and human brain development. *Yearly* face-to-face meetings will be held at the beginning of each budget period. PIs and senior investigators will share their research results with other PIs and investigators, and publication authorship will be based on the relative scientific contributions of individual PIs and investigators. The efforts of each laboratory will need to be tightly integrated in order to communicate progress and results, design and implement analytical tools, and to transfer data. These proceedings of these meetings will be documented and disseminated.

*Decision-making and conflict resolution:*

In the execution of the scientific plan, each overall project PI will have separate responsibilities as delineated in the Approach and Research Timeline. The three laboratories have a substantial history of working collaboratively in various combinations (see biosketches). These efforts have never required a formal decision-making or grievance structure. We have co-authored collaborative papers and have not had any controversies regarding scientific content or order of authorship. We anticipate that the current work will continue in this manner. We recognize, given the novel nature of the work proposed in this grant, that we will need to continually re-evaluate our data and approaches. Given our history and the established monthly conference calls, we do not anticipate that this will present substantial difficulties. Nonetheless, if conflicts arise during the course of the project, the PIs will make every effort to resolve the dispute and reach a position of compromise. If a conflict cannot be resolved, the matter shall be referred to an Arbitration Committee for final resolution. The Arbitration Committee will consist of an equal number of impartial senior executives from each institution who are not directly involved in the conflict but who have appropriate technical credentials, experience and knowledge, and ongoing familiarity with the project, to assist in reaching final resolution.

*Loss of key member of group:*

If a key member of one of the research teams leaves the project, then it will be up to the PI of that group to ensure that an appropriate person is found the replace him or her. If a PI leaves the group, then the remaining PIs will work with the NIMH Project Scientist and Program Officer to find a suitable replacement. However, this will not be an easy task, as the PIs for the project have been specially chosen based on their expertise, productivity, and ability to work collaboratively with each other.

**Data management**

*Comprehensive transparency of data reporting/sharing:*

As describe above, data will be shared freely among the three groups. And while every effort will be made to publish results and release data in a timely manner, we will ensure that the PI's on the project have a chance to fully review and edit all data generated as part of the collaborative project prior to release.

*Reliability and quality control:*

Each of the experiments proposed in this study had built in quality control steps to ensure that accurate data is generated, and a validation step to ensure the data is reliable. The quality of the data is further checked during bioinformatics analysis (which is usually conducted by a different researcher than the one that generated the data), where it is checked for unexpected results or outliers. If the data analysis reveals unpredicted results, then the data will be regenerated, or checked with a second, independent method, to ensure the results are reliable. For consistency in generating the data, for most experiments all of the data will be generated at one site, however, for quality control and reproducibly, the other site might try reproducing the results on a subset of samples.

*Experimental rigor and control of bias:*

The sample sizes for the experiments in this study have been chosen to be large enough to provide sufficient power to provide significant results when compared to normal controls, as described in the proposal. To avoid bias, most of the experiments take a genome-wide prospective, which will allow an unbiased detection of somatic mutations across the genome. Where bias exists because it is not feasible to avoid them (i.e., the high coverage targeted sequencing of blood derived DNA), the nature of the bias, and its possible influence on the results is discussed in the proposal.

## FACILITIES AND RESOURCES

***Statement on how the project will benefit from our collaboration and the integration of resources and expertise unique to each site.*** This proposal combines the expertise and resources of three sites into a single, highly integrated effort aimed at integrating multidimensional datasets to better understand the etiology of neurodevelopmental disorders. The proposal includes three groups led by Nenad Sestan at Yale University (*Application 1/3*), Daniel Geschwind at the University of California, Los Angeles (UCLA; *Application 2/3*) and Matthew State at the University of California, San Francisco (UCSF; *Application 3/3*). Given the complexity of human neurodevelopment and the genetics of autism spectrum disorders (ASD) we believe that integrating the respective expertise and resources of these groups offers the best opportunity to generate new datasets and utilize the plethora of existing multidimensional genomic data to help elucidate the origin of neurodevelopmental disorders. This organizational structure elaborated below combines the expertise and capabilities of the PIs, Daniel Geschwind, Nenad Sestan, and Matthew State and the key investigators, Jason Ernst (UCLA), Mark Gerstein (Yale), Stephan Sanders (UCSF) and Zhiping Weng (University of Massachusetts), and Jeremy Willsey (UCSF), to create a project well beyond the capacity of each individual group. An assurance of effectiveness in these interactions is that Drs. Gerstein, Geschwind, State, Sanders, Sestan, Weng and Willsey have a substantial history of working collaboratively in various combinations and have several joint publications. Additionally, each collaborator has significant experience in large multi-institutional grants of this nature (see Project Management Plan for further details).

### Application 1/3 (Yale and University of Massachusetts)

### Sestan Laboratory (Yale School of Medicine; *www.sestanlab.org*)

***Laboratory*:** A significant part of the research proposed here will take place in the PI's main laboratory of 2,500 sq. ft. located in the Department of Neurobiology on the third floor of the Sterling Hall of Medicine (SHM C316 and 338). This consists of 25 carrel desks, cell culture room, a cold room, two equipment rooms, and two offices with computer workstations. The PI has two tissue culture rooms (one reserved for iPS cell work), a room equipped with a transgenic microinjection set-up, a room with two RNA/DNA workstations and an ABI Genetic Analyzer, *in utero* electroporation equipment, and a room for rodent survival surgeries. The Sestan laboratory has access to and shares maintenance and service responsibilities for the multi-photon and electron microscopy facility within the department.

### Laboratory biological materials resources:

***The Sestan laboratory collection of human and non-human primate brain tissues*:**
Human tissue: The collection has around 200 high-quality de-identified fresh frozen human brain specimens from clinically unremarkable (neurotypical) control donors and donors affected with neurodevelopmental disorders that have passed our internal neuropathological assessment. These specimens range in age from 5 weeks post-conception (PCW) to over 80 years old. Among the archived brains are two pairs of mid-fetal monozygotic twins with a very low post-mortem interval and for which parental DNA was collected, thereby providing a unique opportunity to study transcription and genomic imprinting in human monozygotic twins. In addition, the Sestan lab has an extensive collection of fixed and cryoprotected human brain specimens for immunohistochemistry and in situ hybridization.

Human tissue was collected after parental or next of kin consent and with approval by the institutional review boards at the Yale University School of Medicine and of each institution from which tissue specimens were obtained. Appropriate written informed consent was obtained and all available non-identifying information was recorded for each specimen. Because none of these human sources provided any linkable identifiers to the Sestan lab, and the proposed use of the post-mortem tissue has been reviewed and exempt by the Human Investigation Committee at the Yale School of Medicine. This is because the research does not involve any living human subjects or information about any identifiable living human subjects. Importantly, all brain banks require donors to sign appropriate consent forms. The tissue, linked information, and consent forms obtained by these institutions has been properly deposited into brain banks under IRB or similar protocol approved by each institution's medical ethical committee for the studies proposed in our application. This information can only be accessed by the brain bank/tissue repository coordinator. Furthermore, identifiable data at brain banks/tissue repositories is doubly encoded and saved on separate computer systems. The handling of tissue was performed in accordance with ethical guidelines and regulations for the research use of human brain tissue set forth by the NIH (http://bioethics.od.nih.gov/humantissue.html) and the WMA Declaration of Helsinki (www.wma.net/en/30publications/10policies/b3/index.html).

**List of healthy (neurotypical) and ASD human postmortem available for this project is presented at the end of this section.**

Information on human post-mortem tissue quality: The tissue collection described above was used to generate RNA-seq and ChIP-seq data for the BrainSpan project (*www.brainspan.org*). The high overall quality of the human post-mortem specimens in this collection can be illustrated by the following information on the tissue samples used for the BrainSpan project: post-mortem interval (PMI), 12.11±8.63 (mean±s.d.) hours; pH, 6.45±0.34 (mean±s.d.); and RNA integrity number, 8.83±0.93 (mean±s.d.).

Tissue storage system: All human and non-human primate tissue specimens are stored in four -80 C freezers within the Sestan lab. These frezeers are connected to a SCADA -80C freezer alarm system and a liquid carbon dioxide ($CO_2$) backup system with 24-hour battery back-up time.

***Human brain transcriptome and epigenome datasets*****:** The Sestan lab has been involved in the generation of human brain transcriptome and epigenome datasets for the BrainSpan project (*www.brainspan.org*). This project aims to provide a comprehensive assessment of epigenomic, transcriptional and post-transcriptional events in the developing and adult human brain. The central part of the project consists of transcriptome profiling of 16 cortical and subcortical brain regions of 57 clinically unremarkable post-mortem individuals of multiple ethnicities, with their ages ranging from 5.7 weeks post-conception (PCW) to 82 years of age. We have generated and analyzed a total of 1,340 samples with Exon Microarray transcriptome data (Kang et al., 2011; see also *www.BrainSpan.org* and *www.humanbraintranscriptome.org*) and a total of 578 samples (16 brain regions from one hemisphere of the above brain collection with less frequent developmental coverage) have been subjected to both mRNA-seq and small RNA-Seq analyses. To further illuminate underlying regulatory mechanisms we profiled the same human brain samples for their DNA methylation status, for which data has also been uploaded to www.brainspan.org. The availability of this dataset puts us in the very unique position to be able to integrate this knowledge with data acquired in this project, facilitating both the annotation of brain-specific transcripts and functional elements as well as providing a point of reference for analyses in iPSCs and iPSC-derived neural cells. **Table 1** lists relevant neurogenomic resources currently available in the lab.

***Computational and bioinformatics resources*****:** The laboratory's computing infrastructure is partitioned into a private and a public network. The entire infrastructure is fully gigabit capable and is connected to the Yale backbone and Yale High-Performance-Computer (HPC) via gigabit optic fiber; the network architecture was designed with computing efficiency and network security in mind. The private network consists of individual laptops, desktops and workstations, as well as communal computational servers, dumb terminals, a central fileserver, a consolidated NAS, and printers. The public network consists of numerous production webservers that are either real or virtual machines, including the *www.humanbraintranscriptome.org* database. The department has a full-time ITS administrator maintaining the network and computers.

***Cell culture room and equipment for iPS cell work*****:** The main cell culture room (~500 sq. ft.) located in the Sestan laboratory is dedicated exclusively to iPSC work and is equipped with new and up-to-date equipment necessary for advanced cell culturing techniques, iPSC generation and differentiation. There are two cell culture hoods, a -20 ºC freezer and a refrigerator, one Thermo Scientific Water-Jacketed CO2 incubator and one Thermo Scientific Water-Jacketed CO2 incubator with O2 level control, two liquid nitrogen storage units (Locator 6 plus) each with 6,000 vials storage capacity, Sorvall legend X1R centrifuge, Zeiss Primo Vert inverted microscope, Zeiss SteREO Discovery V8 stereomicroscope with fluorescence module and AxioCam camera connected to a computer, barcode reader for labeled vials, Eppendorf microcentrifuge and a water bath.

The second cell culture room (~80 sq.ft) is shared with the rest of the Sestan lab for general cell culture work. It's equipped with Heraeus HeraCell CO2 tissue culture incubator, one cell culture hood, refrigerator/freezer, water bath and a microcentrifuge. The cold room (~80 sq. ft.) provides additional space for manipulation of temperature sensitive samples.

***Additional equipment in the Sestan lab***: In addition to the above equipment and facilities, the Sestan laboratory has other available equipment relevant to characterization of generated cells: two RNA/DNA workstations, Agilent TapeStation 2200 for automated RNA, DNA and protein sample.

***Transgenic micro-injection set-up***: The transgenic micro-injection room (~80 sq.ft.) located adjacent to the main laboratory is equipped with Zeiss Stemi 2000 and Zeiss Stemi SV6 microscopes, Eppendorf TransferMan NK2 micromanipulator, Eppendorf CellTram Air vario holding pipette control, Eppendorf FemtoJet microinjector, Micro- Forge Filament Technical products International micropipette processor, Sutter instrument P-87 micropipette puller, Zeiss Axiovert 135 Inverted Microscope, Heraeus HeraCell CO2 tissue culture incubator, refrigerator/freezer and a chemical hood. In June of 2015, all this equipment and room have become a part of the newly formed Yale Genome Editing Center (YGEC) with Nenad Sestan as the Executive Director. As evident for the attached letter of support from managing Co-Directors, Timothy Nottoli and Xiaojun Xing, the YGEC provide traditional transgenic and gene targeting services and CRISPR/Cas genome editing.

**Gerstein Laboratory (Yale University; *www.gersteinlab.org*)**

**Laboratory (Bass Central/Main campus):** The Gerstein laboratory is found in two connected buildings. The laboratory consists of 6 rooms and comprises a total of ~1,900 sq. ft. In addition, three conference rooms that have projectors provide venues for interaction. There are 40 gigabit-ready desks, equipped with one or two 23" and 30" LCD screens. The space is properly air conditioned for supporting a large number of computers.

**Office:** Mark Gerstein's office space is 178 sq. ft.

**Gerstein Lab Computer Infrastructure:** Laboratory Network and Storage. The lab's computing infrastructure is partitioned into a private and a public network. The entire infrastructure is fully gigabit capable and is connected to the Yale backbone via gigabit optic fibre; the network architecture was designed with computing efficiency and network security in mind. The private network consists of individual laptops, desktops and workstations, as well as communal computational servers, dumb terminals, a central fileserver, a consolidated NAS, and printers. There are also servers that provide essential network services such as NIS, NFS, SMB, DHCP, monitoring and backups. The public network consists of numerous production webservers that are either real or virtual machines. The laboratory maintains its own public subnets of 128 public IP addresses and manages many of its own domains (e.g. gersteinlab.org, molmovdb.org, pseudogenes.org, and partslist.org). The lab has a full-time administrator maintaining the network.

The private and public networks obtain gigabit connectivity through four HP Procurve 5300xl switches that are mutually connected via fibre. The private network is behind a Cisco PIX 525, which is concurrently used as an IPSec VPN gateway into the private network. Within the private network are two NetApp storage appliances with 43Tb of raw space, which is configured with 27.5Tb of working space, thirty custom made 4Tb network disks with a total 120Tb capacity, a Dell NAS with a total of 30TB capacity; the NetApp appliances and Dell NAS are used for live user file space, backups of user files and backups of public production webservers. A seven-day incremental backup and a twelve-month incremental backup are currently being implemented in the lab.

Wireless access is available all throughout the lab. Wireless access connects computers directly to the public network.

Available Computers. There are about forty-seven working laptops in the lab, in which eighteen are recent Macbook Pro models.

In total, the lab has 315u of rack space spread over eight racks. Residing in these racks are a dual CPU twelve core Opteron server with 256GB of memory, a dual CPU six core Opteron server with 128GB of memory, a dual CPU four core Opteron server with 64GB of memory, three Intel blade enclosures with 10 dual CPU Intel blades each, fourteen dual cpu 64 bit Xeons servers and six dual cpu 64 bit Opteron servers; these rack servers are in addition to the NetApp storage appliances and the Dell NAS mentioned above. The rack servers have various uses. The dual CPU Opteron servers are for hosting virtual machines, which function as web hosts. In the private network, five rack servers are for essential network services, four are storage head nodes for the Dell

SAN and a few are network support or experimental machines. The rest of the rack servers are in the public network acting as webservers. The private network has seven business class color laserjet printers.

**Software:** A number of open source software, programs created in-house, and proprietary software is used by the lab researchers for their needs. The lab maintains a set of wiki servers for the documentation of internal information and the public dissemination of information. The lab also manages mailman servers for its mailing lists. The compute nodes are mainly used to develop and run Java and Perl code and to perform Matlab and Gromacs calculations. The public webservers are used to deploy Java, Perl, PHP and Python applications.

Individual tasks are coordinated by a web group calendar. Web applications and servers are continually being monitored by a Nagios monitoring system.

**Yale Life Sciences Supercomputer**: The Gerstein laboratory has priority access to two of the Yale supercomputers, namely Louise and BulldogI, and regular access to six other Yale supercomputers. There are two full-time administrators maintaining the supercomputer.

Louise is a cluster with 112 Dell PowerEdge R610 with (2) quad core E5620 nodes, each with 2.4 Ghz cpu cores and 48 GB RAM. They are interconnected with a Force10 network switch. There is therefore a total of 112*8 cores = 896 cores. Louise has 300 TB (raw) of BlueArc parallel file storage.

BulldogI is a cluster consisting of a head node and 170 Dell PowerEdge 1955 nodes, each containing 2 dual core 3.0 Ghz Xeon 64 bit EM64T Intel cpus, for a total of 680 cores. Each node has 16 GB RAM. The network is Gigabit ethernet. Bulldogi runs a high performance Lustre filesystem. It is managed via PBS. Three 20Tb Dell Power Vault with storage arrays are attached to BulldogI and are dedicated for Gerstein laboratory use. The laboratory also has priority access to a SGI F1240 system. This system has 12 Xeon E5345 Quad-Core 2.33GHz CPUs (for a total of 48 processor cores), with 2 x 4M L2 cache per CPU, a 1333MHz front side bus, 96GB of memory, and 6 Raptor 150GB, 10K rpm SATA drives. It runs SUSE Linux Enterprise Server 10 as a system single image. That is, all 48 cores are managed by a single process scheduler, and the 96 GB memory is, in principle, addressable by a single process. In practice, system caches and buffers reduce the maximum amount of memory available to any given process to about 70 GB. In many ways then, the system can be thought of as an SMP, but in terms of hardware architecture it is closer to an infiniband-connected cluster.

**Core Lab:** The Gerstein Lab is adjacent to the Yale Center for Structural Biology (CSB) Core laboratory. The Core laboratory resources are available to members of the Gerstein lab. The Core laboratory supports the work of all the people associated with the CSB, in total about 200 users and >200 computers. These computers include a number of high-performance graphics workstations for visualizing macromolecular structures and complex data sets. The CSB Core staff of 2 FTE provides support to the associated CSB laboratories as well as the Core computers.

**Oracle Server:** Yale University has an institutional site license for the Oracle database management system. As a result, many major administrative computing systems at Yale are being developed using Oracle, and Yale's ITS staff has extensive Oracle experience. Yale ITS maintains and operates several Oracle database systems at the School of Medicine, and provides access to these machines to many different projects. There are several advantages to using institutional servers. The ITS staff backs up each database on a regular schedule, typically with full backups weekly and partial backups several times a day. The ITS staff maintains the hardware of the database machine, the system software, and the Oracle software. They perform periodic upgrades when new versions of the software become available. They also handle any systems problems that occur, and are available to help troubleshoot any application problems that arise.

### Relevant facilities and other resources at the Yale University

Core facilities, available on a fee-for-service basis, are available at Yale School of Medicine for flow cytometry and cell sorting, oligonucleotide synthesis, DNA and protein sequencing, and mouse transgenic and gene-targeting services.

***Animal facilities*****:** Yale School of Medicine's Division of Animal Care has full facilities for housing and caring for all animals, including contagion-free rooms and veterinary services. In addition, surgical suites are available on a fee per use basis. The macaque breeding colony is housed at the main campus one floor above the Sestan lab, providing quick access.

***Yale Center for Genome Analysis (www.ycga.yale.edu; Director, Shrikant Mane, Ph.D.):***
High-throughput DNA sequencing is carried out by the Yale Center for Genome Analysis (YCGA). The YCGA is housed in a dedicated building with over 7,000 square feet of laboratory and office space. It is providing microarray and high-throughput DNA sequence analysis services using various technologies including Affymetrix, Illumina, Pacific Biosciences, Sequenom, Ion torrent and Nimblegen. The recent addition of automated equipment including the Caliper LabChip GX and the Caliper Sciclone liquid handling system ensure the timely completion of the proposed DNA sequencing. The Center has 20 full time staff including three Ph.D. and three MS level staff appointments. Shrikant Mane, Ph.D., who has extensive experience in high-throughput DNA sequencing technology, is the founding director of the YCGA. The senior staff of the YCGA has over three years of experience in high-throughput sequencing and other staff is well trained in sample processing as well as operation and maintenance of the equipment. Additional infrastructure includes sample tracking using WikiLIMS, DNA quality control and high-performance computation and bioinformatics support (see 'High-performance computing (HPC)', below).

We use this resource for all our sequencing prosed in this application at the YCGA. Even though Matthew State has accepted a position as Chairman of the Department of Psychiatry at UCSF during preparation of this application, his group will also perform all their targeted re-sequencing experiments proposed in Aim 3 at the YCGA due to our long standing interactions with the YCGA.

***High Performance Computing (HPC) and Bioinformatics*****:** The genomic data generated at the YCGA is transferred to a dedicated high performance cluster (HPC) for further analysis. Data is retained on the cluster for at least 18 months, after which it is transferred to tape backup system managed by Yale's Information Technology Services (ITS). The HPC consists of 140 nodes with approximately 1200 cores/CPUs and approximately 2.5 Petabytes of high performance parallel storage (Panasas Inc.); it runs a Linux operating system. All machines are connected via gigabit Ethernet. Hardware and software support for the HPC is provided by ITS and is housed in a secure climate controlled room. Two Ph.D. computer scientists and one M.S. level staff support the IT and High Performance Computational needs of the Center.

***Biostatistics and Bioinformatics*****:** Technical assistance in analysis of data generated at YCGA is provided by four Ph.D.-level scientists and the Keck laboratory's bioinformatics and biostatistics section. The bioinformatics staff has extensive experience developing new programs for the analysis of data generated by microarray as well as by next gene sequencing platforms and are currently actively involved not only in conducting high level analyses but are also in developing new algorithms and data analysis tools.

***Keck Foundation Biotechnology Resource Laboratory at Yale University*****:** While most of the Keck Resources are at 300 George St., the Microarray Resource is located within the Yale Center for Genome Analysis (YCGA, http://ycga.yale.edu/index.aspx) at West Campus. Keck Genomic resources at Yale provide a wide range of genomic analyses including Sanger sequencing, oilgonucleotide synthesis, microarray analysis and next-generation sequencing. The Keck Sanger DNA sequencing Resource provides competitive and timely DNA sequencing in an efficient and cost effective manner for > 300,000 templates each year. In addition, they also provide high throughput DNA sequencing using Ion Torrent next generation sequencing platform. The Keck Oligonucleotide facility provides > 30,000 high quality, timely DNA and RNA syntheses. In addition to standard synthesis, they routinely perform a wide variety of complex specialty syntheses, emphasizing quality and efficiency. Keck Microarray Resource is very closely associated with the Yale Center for Genome Analysis (YCGA).

***The Yale Stem Cell Center***: The Yale Stem Cell Center (http://stemcell.yale.edu/) provides core facilities for stem cell research at Yale and throughout Connecticut. Major pieces of equipment purchased or supported with funds from the Connecticut Stem Cell Research Fund are available to stem cell researchers, including human ES Cell and iPS cell culture core laboratories, a Confocal Microscopy core, a Genomics and Bioinformatics core, and a Flow Cytometric Analysis and Cell Sorting (FACS) core.

***A Cell Sorter Facility*** (http://info.med.yale.edu/immuno/cytometry/): Mario Skarica, a research associate scientist in the Sestan lab has been trained and authorized to use one of the sorters in the core facility. In this facility there are 3 high-speed sorters: a BD FACSVantage SE, a BD Aria, and a DakoCytomation MoFlo. Cells can be sorted at the rate of 20K/s at 99+% purity, using a variety of commonly used fluorochromes, eg. FITC, PE, PE-Cy5, APC, APC-CY7, PE-CY7, CY5, Alexafluor 350, Hoechst 33342, I ndo-1, and PI. The MoFlo and Vantage SE have 6-color and the Aria has 8-color capability. Cells can be sorted into 5 or 15ml tubes or into various plates as single or multiple cells/well or onto microscope slides for analysis. In addition, the SE and Aria are capable of simultaneous 4-way sorting. The SE and MoFlo have dedicated operators (Tom Taylor and Gouzel Tokmoulina) who will help with experimental design and running samples. For FACS analysis, there are 6 user-operated analyzers (Four BD FACSCaliburs, one FACScan and LSR II). The Caliburs can analyze up to 4 colors with 488 nm and 633 nm excitation lasers while the FACScan can analyze 3 colors from a single 488 nm laser. LSR II has a capability of 9 colors. Also some of our machines are equipped with the FACSFlow fluidics stabilization system and the FACSLoader for automatic processing of samples. Machines are checked daily for optimal performance and Geoff Lyon is available during working hours to troubleshoot and can arrange training by appointment.

**Weng Lab (University of Massachusetts Medical School Subcontract; *http://www.umassmed.edu/zlab/*)**

The Zhiping Weng lab at UMass Medical School (UMMS) has all the equipment needed to efficiently carry out the UMass portion of this project. Her office and her lab (8 offices, housing 15 people in total) are on the 5th floor of the Albert Sherman Building, a state-of-the-art research building opened in Jan 2013. In her lab, everyone has his or her own Mac or PC to connect to Linux clusters and University supercomputers. The lab also uses Amazon Elastic Compute Cloud services. The Weng lab has twelve file servers, with a total of 640 cores, 4.4 TB of RAM, and 1206 TB of disk space. Two servers house a local installation of the UCSC genome browser, as well as a local installation of the Galaxy Web Server.

UMMS is part of the Mass Green High Performance Computing Center (GHPCC) at Holyoke. The computer cluster has 10264 cores available, and 400TBs of high performance EMC Isilon X series storage. The GHPCC consists of the following hardware: an FDR based Infiniband (IB) network and a 10GE network for the storage environment, qty three (6) GPU nodes (Intel with 256GB RAM) with two NVIDIA Tesla C2075 - GPU computing processor - Tesla C2075 - 6 GB GDDR5 - PCI Express 2.0 x16 units or K80 GPUs, qty six (6) AMD Opteron(tm) Processor 6380 based Dell chassis with 64 cores / 512GB RAM per blade (48 blades), qty seven (7) AMD (2x AMD Opteron 6278, 2.4GHz, 16C, Turbo CORE, 16M L2/16M L3, 1600Mhz ) based Dell chassis with 64 cores / 512GB RAM per blade (42 blades), qty (3) Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz, QPI, Turbo, 20c, qty two (2) Intel (Xeon E5-2650 2.00GHz, 20M Cache, 8.0GT/s QPI, Turbo, 8C, 95W, Max Mem 1600MHz) based chassis with 16 cores / 196GB RAM per blade (16 blades), qty two (2) SGI UV200 with 512 Intel (Intel® Xeon® processor E5-4600) cores and 4TBs of fully addressable memory, qty one (1) AMD based Dell chassis with 128 cores Quad-Core AMD Opteron(tm) Processor 2376 and 256GB RAM, qty three (3) Intel (six-core Intel(R) Xeon(R) CPU X5650 @ 2.67GHz ) based Dell chassis with 12 cores / 48GB RAM per blade (16 blades). The HPC environment runs the IBM LSF scheduling software for job management.

**Application 2/3 (UCLA)**

**Geschwind Laboratory (UCLA David Geffen School of Medicine; *http://geschwindlab.neurology.ucla.edu/*)**

**Laboratory:**
Located in the Gonda (Goldschmied) Neuroscience and Genetics Research Center, Dr. Daniel Geschwind's laboratory comprises approximately 2600 square feet of laboratory space on two floors plus a cold room, a radioactive use room, two dark rooms, and over 200 square feet of dedicated bioinformatics space with computer workstations. The laboratory contains all equipment needed for modern molecular biology and genetics research. There are three MJ Research PTC-100 thermal cyclers, five MJ Mini thermal cyclers, five Applied Biosystems GeneAmp 9700 thermal cyclers, an Applied Biosystems Veriti thermal cycler, one Sorvall floor model high-speed centrifuge, one Sorvall table-top model high-speed centrifuge, fifteen table top centrifuges, one full size refrigerator, two under-bench refrigerators, two refrigerator/freezer combos, five full sized -20° C freezers,

eighteen under-bench -20° C freezers, four ultra-low temperature freezers, three liquid nitrogen storage dewars, a Leica CM1850 cryostat, two Eppendorf thermo-mixers, a Benchmark shaking incubator, a New Brunswick shaking incubator, four incubators, five water baths, a hybridization oven, a microwave, a Bio-rad Gene Linker UV chamber, a speed-vac concentrator, a gel dryer, eight high voltage power supplies, a UVP digital gel image capture and analysis systems, a Bio-rad Transblot Turbo transfer system, and vertical and horizontal gel electrophoresis equipment. In addition, there are five fume hoods (one with a dedicated radioactive use zone), two laminar flow hoods for tissue culture, and four CO2 incubators. Other equipment includes two Nikon microscopes, one with dark-field and bright-field optics and another inverted microscope with fluorescent objectives, both equipped with a digital imaging system using a cooled CCD camera. There is also a Zeiss confocal microscope with fluorescent and Nomarski DIC optics. Common use equipment adjacent to the laboratory includes a dark room with film developer, a scintillation counter, two autoclaves, two washing machines, and several high-speed centrifuges and ultracentrifuges. The Geschwind laboratory has an Ion Torrent genome sequencer from Life Technologies, which is capable of 200Mbp in two hours for confirmatory runs and checking library quality prior to Illumina sequencing.

The NIH-funded Informatics Center for Neurogenetics and Neurogenomics (ICNN) (co-directed by Dr. Giovanni Coppola and Dr. Eleazar Eskin, PI: Dr. Freimer, Dr. Geschwind is a co-investigator) was initiated in mid-2009 to augment the informatics needs of neurogenetics program members including Dr. Geschwind, and currently occupies approximately 300 square feet of dedicated office space in the Gonda research facility on the UCLA campus, adjacent to the UNGC genomics core and in close proximity to the laboratory of Drs. Coppola and Geschwind. Some ICNN staff members also have private/shared offices elsewhere in the Gonda building. The ICNN is funded by a Center Core grant from NINDS (5P30NS062691) which supports the provision of bioinformatics services (statistical genetics, sequence analysis, and gene expression analysis) to campus neuroscientists. ICNN staff has routine access to 400 nodes in the large genomics computing cluster of UCLA (Hoffman2 UCLA Cluster, http://www.ats.ucla.edu/clusters/hoffman2/, overseen by Dr. Eleazar Eskin), housed in dedicated space managed by the Institute of Digital Research and Education within the California Nanosystems Institute building, located in close proximity to the Gonda building. The ICNN office hosts up to eight computer workstations. One computing server (8 core, 64-bit, 3.1 GHz, 32GB RAM) and one data server (50 TB RAID) is in a separate server room also within the Gonda building. The workstations and servers are connected through a private subnet within the larger Gonda network. The ICNN has already established the analysis pipelines for gene-expression and statistical genetic study, as well as analysis pipelines for targeted and exome resequencing, ChIP-seq and RNA-seq data. The next-gen sequencing pipelines can analyze 1 billion (1000M) reads in few hours, using the full capacity of the above-mentioned sub-cluster. Databases include an online gene expression database (REPAIR), and a sequence variant database (AWEXOME), both of which were developed in the Geschwind lab and run on ICNN servers (see below, computer and database resources as well).

**Clinical:**

The UCLA Department of Neurology Neurogenetics Program, directed by Dr. Geschwind, is nationally and internationally recognized for excellence in the clinical diagnosis and management of neurodevelopmental and neurogenetic disease. UCLA is a tertiary referral center for the State of California's Genetic Handicapped Persons Program (GHPP) and receives referrals from across the United States. In addition, UCLA's Center for Autism Research and Treatment (CART), co-directed by Dr. Geschwind, is one of eight centers in the Studies to Advance Autism Research and Treatment (STAART) Network. With funding from the NIH, including two Autism Center of Excellence Grants, the center performs innovative studies in autism genetics, phenotyping, neuroimaging, and treatment.

**Animal:**

The UCLA Department of Laboratory Animal Medicine is an animal care and housing facility at UCLA fully accredited by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) International. DLAM consists of 141,218 sq. ft. total vivarium space supporting a population of approximately 115,000 animals (92% mice). DLAM is staffed by eight full-time veterinarians, ten veterinary technicians, 110 animal technicians (including their supervisors), and 16 support staff (clinical and diagnostic labs; administrative personnel). All veterinary and technical staff is located on site and available at all times. Animal technicians and veterinary technicians monitor animals daily including weekends and holidays. Animals are housed in UCLA

IACUC-approved animal facilities under temperature and ventilation controlled conditions. Facility staff ensure animals receive fresh food, water and clean bedding on a regular basis (weekly, bi-weekly or more often). Animal care staff carry out routine husbandry procedures including changing cages, feeding and watering. Animals are checked daily by animal care staff to assess their health condition. Veterinarians examine animals that are reported sick or injured and conduct routine facility rounds at least weekly. If animals exhibit any indications of illness, injury, or distress, veterinary staff confers with research personnel to recommend and agree on appropriate treatment or euthanasia. In emergency situations, decisions are made immediately. In non-emergency cases, the customary timeline for attention is 24-48 hr after a report is made. In all cases, the attending veterinarian has ultimate and final decision-making authority over the treatment or euthanasia of the animal.

**Computer:**
The Geschwind laboratory contains over 30 networked Intel Pentium based computers and three Pentium or Unix-based servers with associated peripherals, such as multiple printers including a Dell color laser printer, Dell multi-function printer, and a document scanner. All computers are fully loaded with software and are connected via Ethernet to the campus network and the Gonda Genetics Bioinformatics Core. For more intensive computational analysis, such as RNAseq alignments and exome sequencing, there is high-level daily access to the Hoffman2 computer cluster in the Nanosystems Institute for genomic and genetic analysis, which comprises over 1600 nodes and > 13000 cores for parallel computing, including specifically allocated memory to permit efficient genomic alignments and RNAseq data storage. This cluster is part of the infrastructure used by the ICNN (see above), but is also accessed directly by lab members as needed. Network resources include unrestricted access to full text electronic journals.

An internal data analysis cluster is available for computationally intensive analytical work, which is a shared resource with Coppola lab. The UNIX-based cluster named ‘Orion’ is made up of 1 login node, 1 management node, 12 compute nodes and 4 network attached storage (NAS). Each compute node consists of two 12-core high-performance Intel Xeon processors with 128GB of DDR3 ECC RAM. Geschwind lab currently has 172TB of space in Orion. The cluster includes a job scheduler, compilers for C, C++, Fortran 77,90 and 95 and softwares like R, Perl, Bioconductor, and MySQL. A storage cluster, IDRE cloud archival storage (CASS) is used for storing and archiving data, geschiwnd lab currently has over 50TB of high-performance off-site cloud archival storage space. The current gene expression database and sequence is hosted on an AMD Opteron machine with two eight- core processors, 64GB of RAM and 24TB of disk space. The database currently holds phenotype information on 8,315 individuals, and whole exome sequence data for 512 individuals. The patient phenotypic information is stored in a PostgreSQL relational database management system (http://www.postgresql.org/). The sequence data of the individuals is managed by HDF5 technology suite (http://www.hdfgroup.org/HDF5/). HDF5 makes possible to manage extremely large and complex data collections. In the frontend there is a web interface to assist retrieving information from both the phenotype database as well as the sequence data. The server code is written in Python (http://www.pytables.org/moin) and the client code in Javascript, a combination of ExtJS javascript framework (http://www.sencha.com/products/extjs/) and Google Web Toolkit (http://code.google.com/webtoolkit/). An embedded implementation of JBrowse is used to aid in visualization of individual and population variant data.

**Office:**
Dr. Geschwind has a total of 491 square feet of office space in the Gonda Building, including space for project administration.

**Other:**
The UNGC currently occupies approximately 1500s sq/ft of dedicated laboratory space in the Gonda research facility on the UCLA campus. Installed capital equipment includes two Illumina HiSeq_2500 sequencers, one cBot cluster station, one Illumina LIMS capable iScan confocal laser scanner with Autoloader II automatic loading support capable of scanning all Illumina beadchip formats. One Tecan Genesis 150 robotic liquid handling platform with Illumina GTS and Infinium robot control software installed, One Tecan Evo 150 robotic liquid handling platform with Illumina GTS and Infinium robot control software installed one Tecan Evo 100 robotic liquid handling platform and two 48 place temperature controlled beadchip processing racks, one SciGene Little Dipper microchip processing robot, one Tomtec autosealer, one Sequenom Massarray compact mass spec with

associated nanodispener chip spotter and one MJ Research tetrad 2 thermocycling system. Additional equipment includes one Covaris M220 nucleic acid shearing system, one Covaris E210 high throughput nucleic acid shearing platform one VisonMate SR 2D barcode plate scanner, one Agilent 2200 Tapestation, one Caplier XT nucleic acid size selection system, six programmable incubation ovens, six microplate heat blocks, two tabletop Jouan centrifuges, two Molecular Dynamics fluorescent microplate readers, one speedvac and four high capacity microplate shakers and  -20 and -80 freezer storage. Computer resources include ten networked workstations and one 8 core 64 bit mini tower running Mac OSX and Windows XP-64. The UNGC has 60TB of network storage space with RAID backup. All computer resources are connected through a private subnet within the larger Gonda network.

The UNGC is equipped to provide sequencing services, including library preparation and QC, using all current Illumina and compatible third party chemistries and kits on our HiSeq-2500 instrumentation. The UNGC supports all versions of Illumina's  whole genome  and custom iSelect Infinium genotyping assays, including methylation analysis and all versions of Illumina's Gene expression chips. All of the Geschwind lab microarray processing for SNPs and gene expression is performed via this core.

The facility is currently capable of processing over 480 million genotypes per week at full capacity and is equipped to provide sequencing services, including library preparation and QC, using all current Illumina and compatible third party chemistries and kits on Illumina HiSeq-2500 instrumentation (n= 2). Services: A)RNA purification: The Tecan Freedom Evo workstation has been configured to process RNA purification protocols using the Qiagen RNeasy plus columns in 96 well plate format. Purified RNA will be quantified using the Tecan Evo to set up RiboGreen fluorescent quantitation assays, which will then be read on the Molecular Dynamics fluorescent microplate readers. Aliquots of RNA preps will be analyzed for degradation and scored for overall quality using the Agilent Bioanalyzer service available in the MicroArray Core. B) Library construction: DNA/RNA fractionation, first and second strand cDNA synthesis (RNA), adapter ligation, library pre-aplification and size selection. C) Cluster generation and sequencing: Completed libraries will be accessioned into the UNGC sequencing production pipeline. Cluster preparation for the HiSeq_2500 pipeline takes place on dedicated cBot platform. The cBot is a stand-alone, software-controlled system for the automated generation of clonal clusters from single molecule fragments on Illumina HiSeq flow cells. The automated cBot can generate greater than 190 million clusters per channel of an eight-channel flow cell. The Geschwind lab performs NextGen sequencing via this core, which is adjacent to their labs in the Gonda building, or on their own dedicated sequencers, depending on sample flow and efficiency.

## Application 3/3 (UCSF)

### The State, Sanders, and Willsey Laboratories (UCSF)

The bioinformatics requirements of this proposal make specific use of the facilities offered by the State, Sanders, and Willsey labs and UCSF. UCSF has extensive high-performance computing resources to allow thorough analysis of high-throughput data. The combination of a dedicated genomics cluster (IHG-1) and a cluster with very large computational resources (QB3) which are necessary for simulations will be ideal in integrating the extensive genomics data and systems analysis.

### University of California, San Francisco (UCSF) Environment

The University of California, San Francisco (UCSF), one of the ten campuses of the University of California, is devoted solely to graduate education and research in the health sciences. UCSF is composed of the Schools of Medicine, Dentistry, Pharmacy, and Nursing, and the Graduate Division. In both size and number of students, UC San Francisco is the smallest of the UC campuses. Nevertheless, its relative size belies its distinction as one of the leading biomedical research and health science education centers in the world. In addition, UCSF is a major health care delivery center in northern California with a high volume of regional, national, and international patient referrals.

Over the last century, the original nucleus of academic schools and divisions has grown to include a School of Nursing (1939); the Langley Porter Psychiatric Institute (1942), which contains the city's first psychiatric hospital; and a Graduate Division (1961). The Graduate Division functions as the administrative and quality control unit for more than 854 PhD students, 593 academic master's students, and 1,100 academic postdoctoral fellows, representing 94 countries. UCSF also is home to 11 research institutes, 1,500 laboratories, more than 5,000 ongoing research projects, and a library with a state-of-the-art computing and communications infrastructure. In 2006, UCSF applied for and was successfully awarded an NIH Center for Translational Sciences (CTSI), which is dedicated to research and education in clinical and translational science at UCSF, at affiliated institutions, and in participating communities (described below). UCSF's four professional schools (Dentistry, Medicine, Nursing, and Pharmacy) are ranked in the top tier nationally and internationally (measured by academic quality, publication citations of faculty, and amount of extramural support for research and education) as centers for education and research in the various disciplines. UCSF's graduate academic PhD programs are also ranked in the top tiers of programs in the biomedical bio-psychosocial disciplines. There are over 35 academic departments, 17 multidisciplinary research centers, and many NIH-funded multidisciplinary research grants including 20 Research Program Projects (P01), 9 Center Core Grants (P30), 12 Specialized Center Grants (P50), and 3 Comprehensive Center Grants (P60). The Graduate Division offers 19 degree programs to students pursuing masters and doctoral degrees in disciplines ranging from bioengineering to chemical biology, from biopharmaceutical sciences and pharmacogenomics to nursing, and from global health to sociology. Graduate programs are organized around several interdisciplinary research areas that often contain members from several departments. UCSF also offers a K-30 supported Advanced Training in Clinical Research certificate program and a Master's Degree in Clinical Research. The number of postdoctoral scholars appointed at the UCSF campus is larger than the number appointed at any of the other University of California campuses. UCSF has taken national leadership in the establishment of quality standards for the selection, appointment, compensation, and education of postdoctoral scholars. Of UCSF staff, 52.5% are minorities and 68% are women. Of faculty, 27.5% are minorities and 43% are women. Of the student body, 54% are minorities and 64% are women.

UCSF continues as one of the leading biomedical research and graduate education centers in the world, and it ranks in the top group of institutions of higher learning in total federal funding for research and training. In the past three decades, UCSF has evolved into a world-renowned biomedical research center with an annual budget of over $3.3 billion to support its various research, teaching, and patient care activities. A large portion of the extramural funds received is allocated for biomedical research. Research funding primarily is obtained on a competitive basis from the federal government. Additional research funding is received annually from the State of California, the University of California Office of the President, private research foundations, state and local government agencies, private philanthropy, and industry. UCSF was awarded $532.8M in NIH funding in 2011, which was first among public institutions and second among all institutions nationwide. Fourteen UCSF scientific departments ranked among the top 10 in 2011, with five leading their fields in funding: anesthesiology ($8.8 million), internal medicine ($162 million), neurology ($36.4 million), neurosurgery ($16.2 million), and obstetrics and gynecology ($23.3 million). In 2010, the UCSF School of Dentistry received $19.5M, the School of Nursing received $9.9M, and the School of Pharmacy received $22.6M. The School of Medicine ranked first nationally in 2012, with $445.7M in NIH funding. Among faculty members are five Nobel laureates, 36 National Academy of Sciences members, 54 American Academy of Arts and Sciences members, 79 Institute of Medicine members, and 16 Howard Hughes Medical Institute investigators.

Over the past decade, UCSF's capacity for basic science and clinical research in the context of world class graduate education has been redoubled by the construction of academic facilities at the new UCSF Mission Bay Campus. Currently, UCSF has over 1.5 million assignable square feet (ASF) of research space: ~62,000 ASF in the School of Dentistry, ~1.3 million ASF in the School of Medicine, ~32,000 ASF in the School of Nursing, and ~126,000 ASF in the School of Pharmacy. With the opening of the Neurosciences Building at Mission Bay, another 152,000 ASF of research space became available in the Spring of 2012. This space supports approximately 2,300 principal investigators with active sponsored awards. A UCSF shuttle bus service (running every 20 minutes) allows for efficient staff, reagent, and mail travel between all main campus facilities.

**UCSF Library**

The UCSF Library provides access for UCSF researchers to an exceptionally large range of research journals. The extensive list is in part due to the coordination and consolidation of library services across the entire 10-campus University of California system.

GALEN is the Digital Library at UCSF. PubMed@UCSF is publicly available, but access to full text articles is limited to computers on the UCSF network or to approved offsite computers. It provides access to the MEDLINE database as well as other NLM databases, and is strong in clinical and basic sciences, nursing, dentistry, and health care planning and administration from 1966 to the present. References published between 1958 and 1965 can be viewed through OLDMEDLINE. The MELVYL Catalog is used to locate books at all UC libraries, and California Periodicals to find journals/titles at other University of California, California State University, and California libraries. Many other important databases are available, including Current Contents, BIOSIS, and PsycINFO.

## UCSF School of Medicine

The UCSF School of Medicine, established in 1864, is the oldest continuously operating medical school in the western states. Ranked as one of the top 10 medical schools in the country, it operates facilities at seven campuses in San Francisco and Fresno. It was ranked number one in NIH funding, receiving over $427M in total awards and over $14.7M in training awards in FY2012. With 26 departments, nine organized research units and six interdisciplinary centers, medical school faculty and staff reach beyond the neighborhood to bring cutting-edge scientific research and complex clinical care to the nation and the world.

## UCSF Department of Psychiatry

The UCSF Department of Psychiatry is among the nation's foremost resources in the field of child, adolescent and adult mental health. Psychiatry faculty members are recognized for their leadership roles in state-of-the-art, comprehensive and compassionate patient care, pioneering research, excellence in training the next generation of leaders, advancing public policy to advance mental health, and commitment to diversity. Department programs are active at all major UCSF campuses.

## UCSF High Performance Computing

UCSF has multiple high-performance computing (HPC) clusters. The two that are most relevant to this proposal are:

- **IHG-1:** This cluster is dedicated to high-throughput sequence analysis and maintained by the Institute for Human Genetics (IHG). The equipment is stored at the newest UCSF datacenter at 654 Minnesota Street, which was awarded a LEED Silver rating for sustainability and energy efficiency and is manned by a 24/7 365 operations center group. Access to the facility is restricted to technology staff only and the facility offers redundant power and onsite emergency generators. At present the cluster consists of: 48 Dell R620 nodes, each with 16 cores and 128GB of RAM; and 2 R820 nodes, each with 32 cores and 1TB of RAM. In total 800 cores are available for computations. Each node has a 1GB, 10GB and FDR inifiniband connection. The 1GB is used for node-to-node communication, the 10GB is for NFS data stores while the FDR is used for MPIO and connection to our GPFS data store. The GPFS store is a DDN SFA-12k with 60TB of high-performance storage. Our NFS store acts as tier 3 storage for data at rest and home folders, this is powered by a internally built cluster and currently has 300TB of capacity. All controls are redundant and the ingress and egress point from the cluster is over a dedicated 10GB link.

- **QB3:** This Linux-based shared cluster is located in a dedicated server room on the 1st floor of Byers Hall at Mission Bay. It consists of over 4,500 processor cores, a high quality network attached storage device with 36 TB of storage capacity, and a networking infrastructure, housed in a total of 19 racks.

## UCSF Genomics Core Services

UCSF has multiple genomics-focused cores, however, our labs primarily utilize the Institute for Human Genetics (IHG) Genomics Core at UCSF.

### DNA Banking and Extraction Services

The IHG DNA Banking and Extraction Services Lab (DNA Bank) is located at HSW901A on the Parnassus campus. Staffed by experts in the field, the core facility prepares high quality DNA samples from blood, saliva, and other tissues at appropriate yields. The high quality DNA samples obtained can be directed to other core facilities in microfuge tubes or other standard formats (including SBS 96-well PCR plates) for DNA analysis. If desired, the DNA samples can be stored in their secure -80°C freezers on site at nominal cost.

### High-Throughput Sequencing

Equipment located in the IHG Next Generation Sequencing (NGS) core includes 3 Illumina HiSeq 2500 systems, a Covaris S2 sonicator, 1 Advanced Analytical Fragment Analyzer, 1 NanoDrop ND-1000 spectrophotometer, 2 Beckman Coulter Biomek FXP robotic liquid handling workstations, 2 Affymetrix GeneTitan Systems, 2 Binder hybridization ovens, 1 Beckman Coulter SNP Stream system, 1 PE Evolution P3 robotic liquid handling workstation, 1 Nanodrop Express low volume precision liquid handling workstation, 1 PE EnVision plate reader, multilabel plate reader, over 20 PCR machines, 1 Illumina BeadStation system with an iScan, an ABI Taqman genotyping platform, an Illumina BeadXpress system, and ABI 3730xl DNA sequencer. Additional sequencers, including the Illumina HiSeq 2500 system, the Illumina MySeq system, and the Life Technologies Ion Proton sequencer are found elsewhere on campus and are available to UCSF Researchers.

**UCSF Psychiatry Department Bioinformatics Core (PsychCore) Services**

PsychCore was created to provide bioinformatics support for the psychiatry department, especially in the fields of genomics and imaging analysis. Stephan Sanders MD PhD is the director and the core currently employs three full time professional software engineers. The core is located in Rock Hall at the Mission Bay campus, in close proximity to Sanders Lab.

With in increasing scale of genomic datasets it has become necessary to transfer existing pipelines from local high-performance clusters to cloud based clusters such as Amazon Web Services (AWS) and Google Genomics (GG). At present PsychCore offers the following services:

- SNP genotyping analysis for copy number variants
- Targeted sequencing analysis (GATK best practices)
- Exome sequencing analysis (GATK best practices)
- Whole genome sequencing analysis (GATK best practices)
- Family-based detection of *de novo* mutations in CNV and sequencing data
- Genomic variant annotation against coding and non-coding datasets
- RNA-Seq analysis
- Data storage in secure redundant cloud-based providers

Analysis pipelines are optimized for rapid and efficient analysis using multi-threading and parallelization and have quality control metrics built in. In the future we aim to include ChIP-Seq and imaging pipelines that are currently under development.

**State Laboratory (UCSF; http://statelab.ucsf.edu)**

The State lab includes wet (bench work) and dry (computational work) space in Rock Hall at UCSF Mission Bay Campus. The State lab is fully equipped for computational, molecular biological, and molecular genetics research.

**Office:** Matthew State has an office of adjacent to the wet and dry laboratories.

High-performance computing: The State Lab has access to two high performance clusters, both at UCSF and described in detail above. Specifically:

- IHG-1 at UCSF (832 CPUs, 560Tb Storage)
- QB3 at UCSF (>4500 CPUs, 36Tb Storage)

Bioinformatic analysis:

- 4 Mac Pro desktop computers with Intel Xeon 64-bit systems with 8-16 CPUS, 8-32GB of SDRAM, and 2-6TB of RAID HDD. These have dual-boot ability for Windows and Mac operating systems allowing the use of tools written in either environment.
- 13 iMac desktop computers with dual- or quad-core Intel processors, 8-32 GB of SDRAM, and 0.5-3TB of storage.
- 5 MacBook Pro laptop computers with dual-core Intel processors, 8-16 GB of SDRAM, and 0.5-1TB of storage
- 1 Dell Precision T7400 with 8 CPUs, 20Gb SDRAM, 6TB of RAID HDD.
- Two QNAP Turbo NAS servers, one TS-809 and one TS-859 Pro+.

Cloud computing: See UCSF Psychiatry Department Bioinformatics Core (PsychCore) Services.

UCSF State Lab Equipment:

In addition to the standard equipment required for molecular genetics research, the State lab owns the following specialized equipment:

- An Eppendorf epMotion 5075LH automated pipetting system
- An Eppendorf epMotion 5070 automated pipetting system with the ability to pick single wells in a preprogrammed pattern
- An Eppendorf New Brunswick U725 Innova Ultra-low Temperature Freezer for sample storage
- An Eppendorf Model 5810R refrigerated centrifuge
- An Eppendorf Model 5804 centrifuge
- A Micronic bar code scanner including a BioMicrolab Tracker Table Model Scanner (3 racks), a Wireless Handheld Scanner, and 1D & 2D barcode reader w/ USB connection
- An BioMicroLab XL20 Tube Handler for automated sample re-array, sample weighing, volume checking, barcode reading, and data collection for Micronic tubes in a 96 tube rack
- 2 Tetrad 2 PCR instruments, each with a (swappable) 384-well and 96 well-reaction module
- 9 BioRad icycler PCR instruments, each with a (swappable) 384-well and 96 well-reaction module
- An ABI-7900HT fast real time PCR system with 96 and 384 well capacity
- A Covaris S2 DNA Sonicator
- A BioTek Synergy HT Fluorometer for PicoGreen DNA concentration assessment
- A Thermo Scientific (ND8000) 8-well Nanodrop 8000 Spectrophotometer
- BioTek Elx800 Plate Reader
- An Agilent 2100 Bioanalyzer
- 2 Invitrogen E-Gel iBase electronic gel readers
- A BioRad ChemiDoc MP System for gel imaging

**Sanders Laboratory (UCSF; http://sanderslab.ucsf.edu)**

The Sanders Lab includes both dry (computational work) and wet (bench work) lab space in Rock Hall on the same floor as the State Lab. The wet lab space is within an open lab shared with neighboring PIs and has access to all the equipment necessary for molecular genetic analysis. The dry lab space is housed in two dedicated bioinformatics rooms in close proximity to the wet lab. It is fully equipped for a wide range of computational analyses, including high-throughput sequencing analysis, genomic annotation, and statistical analysis.

**Office:** Stephan Sanders has an office adjacent to the wet and dry laboratories.

Computational resources:

High-performance computing: The Sanders Lab has access to the two high performance clusters. Specifically:

- IHG-1 at UCSF (832 CPUs, 560Tb Storage)
- QB3 at UCSF (>4500 CPUs, 36Tb Storage)

Bioinformatic analysis: Every member of the Sanders Lab has a dedicated Mac computer, however the majority of analysis is performed on the high-performance or cloud-based computing clusters.

Cloud computing: See UCSF Psychiatry Department Bioinformatics Core (PsychCore) Services.

**Willsey Laboratory**

The Willsey Lab has dry lab space at Rock Hall on the same floor as the State Lab.

Computational resources:

High-performance computing: The Sanders Lab has access to the two high performance clusters. Specifically:

- IHG-1 at UCSF (832 CPUs, 560Tb Storage)
- QB3 at UCSF (>4500 CPUs, 36Tb Storage)

Bioinformatic analysis: Every member of the Willsey Lab has a dedicated Mac computer, however the majority of analysis is performed on the high-performance computing clusters.

Cloud computing: See UCSF Psychiatry Department Bioinformatics Core (PsychCore) Services.

**Office:** Jeremy Willsey has an office adjacent to the dry laboratories.

**State, Sanders, and Willsey Lab Software**

Processing the large volumes of data generated by high-throughput sequencing and genotyping microarrays requires custom-designed, automated bioinformatic tools. Alongside pre-installed open source and proprietary software we have written and implemented the following tools:

- **Genomic pipeline:** This is a pipeline for processing and analyzing high-throughput sequencing data in a Unix-based high-performance cluster environment. It makes use of parallelization to rapidly and efficiently process unaligned sequence to annotated single nucleotide variant (SNV) and insertion/deltion (indel) predictions using BWA for alignment and GATK for variant prediction. In addition it can predict *de novo* SNVs in family-based data (Sanders *et al. Nature* 2012).
- **Targeted sequencing:** The UCSF genomic pipeline has been modified to manage pooled data to allow analysis of Molecular inversion probes (MIPs), or capture arrays. Once the target is defined the pipeline can run automatically on all forms of targeted sequencing.
- **Indel detection:** To accurately predict *de novo* insertion/deltion (indels) we have designed a specific indel analysis pipeline (Dong *et al. Cell Reports* 2014). The aligned and sorted BAM files produced as an intermediary in UCSF genomic pipeline (above) are passed into Dindel. Putative indels are identified then local realignment is used to refine the call. By performing local realignment for indels in all family members and accurate assessment of *de novo* status is made.
- **Annotation:** Cross-referencing predicted variants with other sources of data (e.g. genes, conservation scores, brain-expression, frequency data) is key to downstream analysis. To allow accurate and fast annotation of region-based and single nucleotide data we have designed customizable annotators that use multi-level indexing and binary search algorithms for rapid processing.
- **CNVision (http://sanderslab.ucsf.edu/article/software):** This is a pipeline for processing and analyzing genotyping data to predict copy number variation (CNV). The pipeline is designed to work in a UNIX environment and to make use of parallelization for rapid and efficient throughput. It uses three CNV prediction tools to maximize accuracy: PennCNV, QuantiSNP and GNOSIS (in-built) and assigns a p-value to every CNV based on per SNP variability in the underlying data (Sanders *et al. Neuron* 2011).
- **Primer prediction:** To allow simple and consistent confirmation of predicted variants we have written an automated pipeline to retrieve reference sequence around a variant, generate primers using Primer3, testing with in-silico PCR, and cross referencing against dbSNPv135; in addition a Sequencher input file is generated for simple analysis of Sanger sequencing results.
- **Identity (http://sanderslab.ucsf.edu/article/software):** Ensuring the correct samples have been sequenced and compared is essential for accurate analysis. An in-house script generates a barcode from exome BAM files, genome BAM files, or SNP genotyping data which can be compared across samples to confirm identity.

**Lists of developmental control and ASD brains available in Geschwind and Sestan labs for this project**

**Table 1. List of developmental control brains**

| NUMBER | ID# | Age | Sex | Ethnicity | PMI (hours) | Tissue level RNA-seq | psychENCODE dedicated brains |
|---|---|---|---|---|---|---|---|
| **Late mid-fetal period** | | | | | | | |
| **1** | HSB 242 | 17 pcw | F | TBD | <1 | N/A | Yes |
| **2** | HSB 267 | 17 pcw | M | TBD | <1 | N/A | Yes |
| **3** | HSB 239 | 17 pcw | F | TBD | <1 | N/A | Yes |
| **4** | HSB 268 | 17 pcw | M | TBD | <1 | N/A | Yes |
| **5** | HSB 265 | 21 pcw | F | TBD | 15 | N/A | Yes |
| **6** | HSB 274 | 21 pcw | M | TBD | 8 | N/A | Yes |
| **Infancy** | | | | | | | |
| **7** | HSB 220 | 0.001 y | F | H | 9 | N/A | Yes |
| **8** | HSB 121 | 0.3 y | M | C | 10 | **Available** | No |
| **9** | HSB 132 | 0.3 y | M | C | 22 | **Available** | No |
| **10** | HSB 139 | 0.3 y | M | AA | 20 | **Available** | No |
| **11** | HSB 131 | 0.5 y | F | C | 26 | **Available** | No |
| **12** | HSB 122 | 1 y | F | C | 18 | **Available** | No |
| **Childhood** | | | | | | | |
| **13** | HSB 143 | 2 y | F | C | 12 | **Available** | No |
| **14** | HSB 173 | 3 y | F | C | 8 | **Available** | No |
| **15** | HSB 172 | 3 y | M | H | 16 | **Available** | No |
| **16** | HSB 118 | 4 y | M | AA | 20 | **Available** | No |
| **17** | HSB 141 | 8 y | M | AA | 30 | **Available** | No |
| **18** | HSB 174 | 8 y | M | AA | 16 | **Available** | No |
| **Adolescence** | | | | | | | |
| **19** | HSB 175 | 11 y | F | AA | 22 | **Available** | No |
| **20** | HSB 124 | 13 y | F | AA | 20 | **Available** | No |
| **21** | HSB 119 | 15 y | M | AA | 14 | **Available** | No |
| **22** | HSB 105 | 18 y | M | C | 8 | **Available** | No |
| **23** | HSB 127 | 19 y | F | C | 10 | **Available** | No |
| **24** | To be prospectively collected by the Sestan lab | | | | | | |
| **Adulthood** | | | | | | | |
| **25** | HSB 187 | 37 y | F | AA | 22 | N/A | Yes |
| **26** | HSB 285 | 37 y | F | C | 6 | N/A | Yes |
| **27** | HSB 317 | 47 y | M | C | 5 | N/A | Yes |
| **28** | HSB 269 | 49 y | F | C | 10 | N/A | Yes |
| **29** | HSB 277 | 53 y | M | C | 12 | N/A | Yes |
| **30** | HSB 244 | 59 y | M | C | 22 | N/A | Yes |

M - Male, F - Female, AA - African American, C - Caucasian, H - Hispanic
CTL – control
TBD - to be determined from genome data
N/A - not available

## Table 2: List of ASD and matching control brains

| # | SOURCE | Disorder | ID# | Age | Sex | Tissue Level RNASeq | | SOURCE | Disorder | ID# | Age | Sex | Tissue Level RNAseq |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Period 10 (Early Childhood)** | | | | | | | | | | | | | |
| 1 | Geschwind Lab | ASD / AUTISM | AN03345 | 2 | M | Available | 1 | Sestan lab | CTL | HSB 173 | 3 | F | Available |
| 2 | Geschwind Lab | ASD / AUTISM | UMB5308 | 4 | M | Available | 2 | NICHD | CTL | 4670 | 4 | M | |
| 3 | Sestan Lab | ASD / AUTISM | 4671 | 4 | F | | 3 | Sestan lab | CTL | HSB 118 | 4 | M | Available |
| 4 | Geschwind Lab | ASD / AUTISM | AN08873 | 5 | M | Available | 4 | NICHD | CTL | 4327 | 5 | F | |
| 5 | Geschwind Lab | ASD / AUTISM | AN13872 | 5 | F | Available | 5 | NICHD | CTL | 3 | 5 | M | |
| 6 | Sestan Lab | ASD / AUTISM | 1349 | 5 | M | Available | | | | | | | |
| **Period 11 (Middle and late childhood)** | | | | | | | | | | | | | |
| 7 | Sestan Lab | ASD / AUTISM | 5144 | 7 | M | Available | 6 | NICHD | CTL | 4898 | 7 | M | |
| 8 | Sestan Lab | ASD / AUTISM | 4849 | 7 | M | Available | 7 | NICHD | CTL | 629 | 7 | M | |
| 9 | Sestan Lab | ASD / AUTISM | 1174 | 7 | F | | 8 | Geschwind Lab | CTL | UMB4337 | 8 | M | Available |
| 10 | Geschwind Lab | ASD / AUTISM | AN19511 | 8 | M | Available | 9 | Sestan lab | CTL | HSB 141 | 8 | M | Available |
| 11 | Geschwind Lab | ASD / AUTISM | UMB4334 | 8 | M | Available | 10 | Sestan lab | CTL | HSB 174 | 8 | M | Available |
| 12 | Sestan Lab | ASD / AUTISM | 4231 | 8 | M | Available | 11 | NICHD | CTL | M3835M | 9 | F | |
| 13 | Sestan Lab | ASD / AUTISM | 4721 | 8 | M | Available | 12 | NICHD | CTL | 5173 | 10 | F | |
| 14 | Geschwind Lab | ASD / AUTISM | AN16641 | 9 | M | Available | 13 | NICHD | CTL | 39 | 10 | M | |
| 15 | Geschwind Lab | ASD / dup15q | AN14762 | 9 | M | Available | 14 | NICHD | CTL | 616 | 11 | M | |
| 16 | Sestan Lab | ASD / AUTISM | 797 | 9 | M | | 15 | Sestan lab | CTL | HSB 175 | 11 | N/A | Available |
| 17 | Sestan Lab | ASD / AUTISM | 1182 | 9 | F | | 16 | NICHD | CTL | 5334 | 12 | M | |
| 18 | Geschwind Lab | ASD / dup15q | AN06365 | 10 | M | Available | | | | | | | |
| 19 | Sestan Lab | ASD / AUTISM | M2004M | 10 | M | | | | | | | | |
| 20 | Geschwind Lab | ASD / AUTISM | AN16115 | 11 | F | Available | | | | | | | |
| 21 | Geschwind Lab | ASD / AUTISM | AN17678 | 11 | M | Available | | | | | | | |
| 22 | Geschwind Lab | ASD / dup15q | AN09402 | 11 | M | Available | | | | | | | |
| 23 | Sestan Lab | ASD / AUTISM | 4334 | 11 | M | Available | | | | | | | |
| 24 | Sestan Lab | ASD / AUTISM | 5454 | 11 | M | | | | | | | | |
| 25 | Sestan Lab | ASD / AUTISM | 4305 | 12 | M | | | | | | | | |
| 26 | Sestan Lab | ASD / AUTISM | 5565 | 12 | M | | | | | | | | |
| **Period 12 (Adolescence)** | | | | | | | | | | | | | |
| 27 | Sestan Lab | ASD / AUTISM | 5710 | 13 | M | | 17 | Sestan lab | CTL | HSB 124 | 13 | F | Available |
| 28 | Geschwind Lab | ASD / AUTISM | NP27/11 | 14 | M | Available | 18 | NICHD | CTL | 5163 | 14 | M | |
| 29 | Sestan Lab | ASD / AUTISM | 4315 | 14 | M | | 19 | NICHD | CTL | 917 | 14 | M | |
| 30 | Sestan Lab | ASD / AUTISM | 4899 | 14 | M | Available | 20 | Geschwind Lab | CTL | UMB5163 | 15 | M | Available |
| 31 | Geschwind Lab | ASD / AUTISM | AN02987 | 15 | M | Available | 21 | Geschwind Lab | CTL | UMB5242 | 15 | M | Available |
| 32 | Geschwind Lab | ASD / AUTISM | AN04682 | 15 | M | Available | 22 | NICHD | CTL | 1843 | 15 | F | |
| 33 | Geschwind Lab | ASD / AUTISM | NP72/11 | 15 | M | Available | 23 | NICHD | CTL | 1065 | 15 | M | |
| 34 | Sestan Lab | ASD / AUTISM | 5531 | 15 | M | | 24 | Geschwind Lab | CTL | UMB5168 | 16 | F | Available |
| 35 | Geschwind Lab | ASD / AUTISM | UMB5278 | 16 | F | Available | 25 | Geschwind Lab | CTL | AN17425 | 16 | M | Available |
| 36 | Geschwind Lab | ASD / AUTISM | UMB5302 | 16 | M | Available | 26 | Geschwind Lab | CTL | AN00544 | 17 | M | Available |
| 37 | Geschwind Lab | ASD / dup15q | AN17138 | 16 | M | Available | 27 | Geschwind Lab | CTL | AN07444 | 17 | M | Available |
| 38 | Sestan Lab | ASD / AUTISM | 5403 | 16 | M | | 28 | Sestan lab | CTL | HSB 105 | 18 | M | Available |
| 39 | Geschwind Lab | ASD / AUTISM | AN01570 | 18 | F | Available | 29 | NICHD | CTL | 1571 | 18 | F | |
| 40 | Sestan Lab | ASD / AUTISM | 5419 | 19 | F | | 30 | Geschwind Lab | CTL | AN03217 | 19 | M | Available |
| 41 | Sestan Lab | ASD / ASPERGER'S | 5294 | 19 | M | | 31 | Sestan lab | CTL | HSB 127 | 19 | F | Available |

| # | SOURCE | Disorder | ID# | Age | Sex | Tissue Level RNASeq | | SOURCE | Disorder | ID# | Age | Sex | Tissue Level RNAseq |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period 13 (Young adulthood) | | | | | | | | | | | | | |
| 42 | Geschwind Lab | ASD / AUTISM | AN00764 | 20 | M | Available | 32 | Geschwind Lab | CTL | UMB4590 | 20 | M | Available |
| 43 | Geschwind Lab | ASD / dup15q | AN03935 | 20 | M | Available | 33 | NICHD | CTL | 1475 | 20 | M | |
| 44 | Sestan Lab | ASD / AUTISM | 1638 | 20 | F | | 34 | Geschwind Lab | CTL | AN07176 | 21 | M | Available |
| 45 | Geschwind Lab | ASD / AUTISM | UMB4999 | 21 | M | Available | 35 | Sestan lab | CTL | HSB 130 | 21 | F | Available |
| 46 | Geschwind Lab | ASD / AUTISM | AN09730 | 22 | M | Available | 36 | Geschwind Lab | CTL | A206/89 | 22 | M | Available |
| 47 | Geschwind Lab | ASD / AUTISM | UMB5176 | 22 | M | Available | 37 | Geschwind Lab | CTL | AN10833 | 22 | M | Available |
| 48 | Geschwind Lab | ASD / AUTISM | NP26/11 | 22 | M | Available | 38 | Geschwind Lab | CTL | UMB5342 | 23 | M | Available |
| 49 | Sestan Lab | ASD / AUTISM | 5176 | 22 | M | Available | 39 | Sestan lab | CTL | HSB 136 | 23 | M | Available |
| 50 | Sestan Lab | ASD / AUTISM | 5610 | 22 | M | | 40 | Geschwind Lab | CTL | AN14757 | 24 | M | Available |
| 51 | Sestan Lab | ASD / AUTISM | 5574 | 23 | M | | 41 | Geschwind Lab | CTL | AN19760 | 28 | M | Available |
| 52 | Geschwind Lab | ASD / dup15q | AN05983 | 24 | M | Available | 42 | Sestan lab | CTL | HSB 125 | 28 | M | |
| 53 | Geschwind Lab | ASD / dup15q | AN14829 | 26 | F | Available | 43 | Geschwind Lab | CTL | AN12137 | 31 | M | Available |
| 54 | Geschwind Lab | ASD / AUTISM | AN00493 | 27 | M | Available | 44 | Geschwind Lab | CTL | AN15566 | 32 | F | Available |
| 55 | Sestan Lab | ASD / AUTISM | M3663M | 27 | M | | 45 | Geschwind Lab | CTL | UMB5079 | 33 | M | Available |
| 56 | Geschwind Lab | ASD / AUTISM | AN08166 | 29 | M | Available | 46 | Sestan lab | CTL | HSB 138 | 33 | M | |
| 57 | Geschwind Lab | ASD / AUTISM | AN12457 | 29 | F | Available | 47 | Geschwind Lab | CTL | AN08161 | 36 | F | Available |
| 58 | Geschwind Lab | ASD / AUTISM | AN08792 | 30 | M | Available | 48 | Geschwind Lab | CTL | AN10028 | 36 | M | Available |
| 59 | Geschwind Lab | ASD / AUTISM | AN11989 | 30 | M | Available | 49 | Geschwind Lab | CTL | A268/93 | 37 | M | Available |
| 60 | Geschwind Lab | ASD / AUTISM | NP121/11 | 33 | M | Available | 50 | Geschwind Lab | CTL | UMB1376 | 37 | M | Available |
| 61 | Geschwind Lab | ASD / AUTISM | UMB5297 | 33 | M | Available | 51 | Sestan lab | CTL | HSB 123 | 37 | M | Available |
| 62 | Sestan Lab | ASD / AUTISM | 5578 | 35 | M | | 52 | Geschwind Lab | CTL | AN08677 | 38 | M | Available |
| 63 | Geschwind Lab | ASD / AUTISM | UMB5027 | 38 | M | Available | | | | | | | |
| 64 | Geschwind Lab | ASD / AUTISM | AN01971 | 39 | M | Available | | | | | | | |
| 65 | Geschwind Lab | ASD / AUTISM | AN06420 | 39 | M | Available | | | | | | | |
| 66 | Geschwind Lab | ASD / dup15q | AN11931 | 39 | F | Available | | | | | | | |
| Period 14 (Middle adulthood) | | | | | | | | | | | | | |
| 67 | Sestan Lab | ASD / AUTISM | 1575 | 40 | F | | 53 | Geschwind Lab | CTL | A247/92 | 40 | M | Available |
| 68 | Sestan Lab | ASD / AUTISM | 5712 | 43 | M | | 54 | Sestan lab | CTL | HSB 135 | 40 | F | Available |
| 69 | Geschwind Lab | ASD / AUTISM | NP167/08 | 44 | F | Available | 55 | Geschwind Lab | CTL | A246/93 | 41 | M | Available |
| 70 | Geschwind Lab | ASD / AUTISM | NP56/10 | 44 | M | Available | 56 | Geschwind Lab | CTL | AN01410 | 41 | M | Available |
| 71 | Geschwind Lab | ASD / AUTISM | UMB5115 | 46 | M | Available | 57 | Geschwind Lab | CTL | AN10679 | 41 | F | Available |
| 72 | Geschwind Lab | ASD / AUTISM | AN03632 | 49 | F | Available | 58 | Geschwind Lab | CTL | AN00142 | 44 | M | Available |
| 73 | Geschwind Lab | ASD / dup15q | AN17777 | 49 | F | Available | 59 | Geschwind Lab | CTL | AN04479 | 44 | M | Available |
| 74 | Geschwind Lab | ASD / AUTISM | NP90/09 | 50 | M | Available | 60 | Sestan lab | CTL | HSB 181 | 44 | M | |
| 75 | Geschwind Lab | ASD / AUTISM | AN17254 | 51 | M | Available | 61 | NICHD | CTL | 1936 | 46 | M | |
| 76 | Geschwind Lab | ASD / AUTISM | AN08043 | 52 | F | Available | 62 | Geschwind Lab | CTL | UMB4842 | 47 | M | Available |
| 77 | Geschwind Lab | ASD / AUTISM | UMB5340 | 53 | M | Available | 63 | Geschwind Lab | CTL | AN19442 | 50 | M | Available |
| 78 | Geschwind Lab | ASD / AUTISM | AN17515 | 54 | M | Available | 64 | Geschwind Lab | CTL | A012/12 | 51 | F | Available |
| 79 | Geschwind Lab | ASD / AUTISM | AN01093 | 56 | M | Available | 65 | Geschwind Lab | CTL | AN12240 | 51 | M | Available |
| | | | | | | | 66 | Geschwind Lab | CTL | AN15088 | 52 | F | Available |
| | | | | | | | 67 | Geschwind Lab | CTL | UMB1578 | 53 | M | Available |
| | | | | | | | 68 | Geschwind Lab | CTL | A358/08 | 55 | F | Available |
| | | | | | | | 69 | Geschwind Lab | CTL | AN01125 | 56 | M | Available |
| | | | | | | | 70 | Geschwind Lab | CTL | AN13295 | 56 | M | Available |
| | | | | | | | 71 | Geschwind Lab | CTL | AN11864 | 57 | M | Available |
| Period 15 (late adulthood) | | | | | | | | | | | | | |
| 80 | Geschwind Lab | ASD / AUTISM | AN09714 | 60 | M | Available | 72 | Geschwind Lab | CTL | AN10723 | 60 | M | Available |
| 81 | Geschwind Lab | ASD / AUTISM | UMB5303 | 67 | M | Available | 73 | NICHD | CTL | 5452 | 67 | M | |
| 82 | Geschwind Lab | ASD / AUTISM | AN06875 | 68 | M | Available | 74 | Sestan lab | CTL | HSB 229 | 70 | M | |
| 83 | Geschwind Lab | ASD / AUTISM | AN06133 | 81 | M | Available | | | | | | | |

## VERTEBRATE ANIMALS

Not applicable