Comprehensive functional genomic resource and integrative model for the adult brain

Daifeng Wang[1*], Shuang Liu[2,3*], Jonathan Warrell[2,3*], Hyejung Won[4,5,6*], Xu Shi[2,3*], Fabio Navarro[2,3*], Declan Clarke[2,3*], Mengting Gu[3*], Prashant Emani[2,3*], Min Xu[2,3], Yucheng T. Yang[2,3], Jonathan J. Park[2,3], Suhn Kyong Rhie[10], Kasidet Manakongtreecheep[2,3], Holly Zhou[2,3], Aparna Nathan[2,3], Jing Zhang[2,3], Mette Peters[11], Eugenio Mattei[12], Dominic Fitzgerald[13], Tonya Brunetti[13], Jill Moore[12], PsychENCODE Consortium[‡], Nenad Sestan[14], Andrew E. Jaffe[15], Kevin White[13], Zhiping Weng[12], Daniel H. Geschwind[4-7†], James Knowles[8†], Mark Gerstein[2,3,9†]

Affiliations:

[1]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA

[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

[3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

[4]Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.

[5]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.

[6]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.

[7]Department of Psychiatry, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.

[8]SUNY Downstate Medical Center College of Medicine, Brooklyn, NY 11203, USA

[9]Department of Computer Science, Yale University, New Haven, CT 06520, USA

[10]Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90007, USA

[11]Sage Bionetworks, Seattle, WA 98109, USA

[12]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA

[13]Institute for Genomics and Systems Biology, Department of Human Genetics, University of Chicago, Illinois 60637, USA

[14]Department of Neuroscience and Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT 06520, USA

[15]Lieber Institute for Brain Development, Johns Hopkins Medical Campus; Departments of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health Baltimore, MD, 21205, USA

* These authors contributed equally to this work

‡ The consortium authors are listed at the end of the paper.

† Co-corresponding authors

Abstract

Our molecular-level understanding of how genomic variants relate to brain disorders is limited. Addressing this challenge, the PsychENCODE consortium has generated ~5,500 genotype, transcriptome, chromatin, and single-cell datasets from 1,866 individuals and, by uniformly processing and analyzing them together with publically available data, has developed a comprehensive resource for the adult brain (available via Adult.PsychENCODE.org). In particular, we deconvolved the gene expression of bulk tissue using single-cell data, finding that differences in the proportions of cell types explain >85% of the cross-population variation observed. Moreover, using chromatin and Hi-C data from reference prefrontal-cortex samples, we found ~79,000 brain-active enhancers and linked them to genes and transcription factors in an extended regulatory network. We identified ~2.5M eQTLs (comprising ~238K linkage-disequilibrium-independent SNPs) and many additional QTLs associated with chromatin, splicing and cell-type-proportion changes. We, also, leveraged our QTLs, Hi-C data and regulatory network to connect more genes to GWAS variants for psychiatric disorders than possible before (e.g., 304 for schizophrenia). Finally, we developed a deep-learning model embedding the regulatory network in a framework connecting genotype to observed traits. Our model achieves a ~6X improvement in disease prediction over an additive model, highlights key genes for disorders, and allows imputation of missing transcriptome information from genotype data alone.

Introduction

Disorders of the brain affect nearly one fifth of the world's population (*1*). Decades of research has led to little progress in our fundamental understanding of the molecular causes of psychiatric disorders. This contrasts with cardiac disease, for which lifestyle and pharmacological modification of environmental risk factors has had profound effects on morbidity, or cancer, which is now understood to be a direct disorder of the genome (*2-5*). Although genome-wide association studies (GWAS) have identified many genomic variants associated with psychiatric disease risk, for the vast majority we have little understanding of the molecular mechanisms affecting the brain (*6*).

To this end, a number of studies have begun to elucidate the molecular steps on the path from genomic alteration to risk. For instance, the Psychiatric Genomics Consortium (PGC) has recently identified 142 GWAS loci associated with schizophrenia (*7*). Many of these lie in non-coding regions (*7*), suggesting roles in gene regulation. Other consortia have annotated non-coding regions using expression quantitative-trait loci (eQTLs) from the Genotype-Tissue Expression (GTEx) project and enhancers from the ENCODE and Epigenomics Roadmap projects. However, none of these projects have specifically tailored their efforts toward the brain. The initial work focusing on identifying brain-specific genomic elements has provided greater insight into brain-specific functional genomics (*8, 9*), but could be enhanced with larger sample sizes from both healthy and diseased samples. Moreover, many new assays for functional elements have been recently developed, such as Hi-C and single-cell sequencing, which have yet to be fully integrated with brain genomics data, at scale (*10-13*).

Hence, the PsychENCODE Consortium has generated a large-scale dataset for providing insights into the adult human brain and psychiatric disorders, including data derived through genotyping, bulk and single-cell RNA-seq, ChIP-seq, ATAC-seq, and Hi-C using brains from 1866 individuals (*14*). All raw and uniformly processed data at both tissue and single-cell level have been placed into a central, publically available resource for brain functional genomics, that also integrates relevant re-processed data from other related projects, including ENCODE, CommonMind (CMC), GTEx, Epigenomics Roadmap, with nearly ~12,000 data samples in total. By leveraging this resource, we were able to identify functional elements and QTLs specific to the adult brain, including novel psychiatric GWAS and gene linkages. Moreover, we combined these elements to build an integrated deep-learning model. This tool can utilize the richly structured data of the resource to identify interactions between genotype and molecular phenotypes at multiple layers, as well as predict high-level traits.

Resource construction

We designed the resource Adult.PsychENCODE.org to provide coherent structure to a large amount of data on brain functional genomics (*1*). Broadly, it organizes data hierarchically, with a large base of raw data files (many of which have restricted access, such as individual genotyping and raw next-generation sequencing of transcriptomics and epigenomics), a middle layer of uniformly processed and easily shareable results (such as open chromatin regions and gene expression quantifications), and a compact cap that consists of an integrative model based on imputed regulatory networks and QTLs. As shown in Fig. 1, to build the base layer we included all the adult data from PsychENCODE (~5,500 datasets derived from 1,866 individual brains) and merged these with relevant data from ENCODE, CMC, GTEx, Roadmap, and recent single-cell studies (~5,000 additional datasets) (*11, 13*). These data cover a representation of phenotypes and psychiatric disorders including Schizophrenia (SCZ), Bipolar (BPD), Autism Spectrum Disorder (ASD). Furthermore, the PsychENCODE project developed a specific "reference brain" project on adult prefrontal cortex (PFC) utilizing many matched assays on the same set of brain tissues, which we used (below) to develop an anchoring annotation (*15*).

Transcriptome analysis: bulk and single-cell

To identify the genomic elements exhibiting transcriptional activities specific to the brain, we used the ENCODE pipeline to uniformly process RNA-seq data from PsychENCODE, GTEx and Roadmap. Using these data, we identified a wide variety of interpretable brain functional elements, such as non-coding regions of transcription, and sets of differentially expressed and co-expressed genes - e.g., 12,080 genes were transcribed in the brains of 95% of the individuals surveyed and over 16,000 protein-coding and 9,000 non-coding genes were detected in total (*15, 16*).

Brain tissues are composed of a variety of cell types, including neuronal and non-neuronal cells. Previous studies have suggested that gene expression changes at the tissue level may be associated with changing proportions of basic cell types (*17-21*). However, studies have not systematically revealed how differing cell types can quantitatively contribute to population-level expression variation. Here, we address this question for expression over our cohort of 1,866 individuals.

We used two complementary strategies. First, we used the standard pipeline to uniformly process single-cell RNA-seq data in PsychENCODE, in conjunction with a number of other single-cell studies on the brain (*11, 13*), in order to assemble a list of brain cell types for the project. This includes previously identified neuronal types, major non-neuronal types, and a number of additional cell types involved in development (*15*). The results constitute a matrix, C of expression signatures, mostly concordant with what has been published (Fig. S2.4 and Conclusion). A number of genes had expression levels varying more substantially across these cell types than they did across individuals in a population (e.g., dopamine receptor DRD3, Fig. 2A). This implies that the changes in bulk expression can readily result from cell fraction variations.

To explore this further, we used a second strategy: an unsupervised analysis to identify the primary components of bulk expression variation as they relate to cell types. We decomposed the bulk gene-expression matrix, B from our resource using non-negative matrix factorization (NMF), $B \approx VH$, and then determined whether the top components capturing the majority of covariance (NMF-TCs, columns of V) were consistent with the single-cell signatures (Fig. 2B and C) (*15*). We found that a number of NMF-TCs highly correlated with neuronal, non-neuronal, and development-related cell types, demonstrating that an unsupervised analysis derived solely from bulk data roughly matches the single-cell signatures, partially corroborating them.

We then tried to understand how variation in proportions of cell types contributes to variation in bulk expression. In particular, we de-convolved the expression matrix of tissue, B using the single-cell signatures, C to estimate the cell-fractions W, solving the equation $B \approx CW$ (*15*) (Fig. 2B). As validation, our estimated fractions of NEU+/- cells matched the experimentally determined fractions from the reference brain samples (Median error = 0.04, Fig. S2.9). We also compared our results with previous deconvolution methods (*15*). Overall, we found that single-cell expression signatures could explain much of the population-level variation (Fig. 2D, i.e., across tissue samples from different individuals $1-||B-CW||^2/||B||^2>85\%$) (*15*).

Finally, we found that cell-fraction changes were associated with different observed phenotypes and disorders (Fig. 2E, S2.6 and S2.7). For example, particular excitatory and inhibitory neurons exhibited different fractions between male and female samples (i.e., Ex3 and In8). The fraction of Ex3 was also reduced in ASD ($p=0.0077$), where non-neuronal cells (e.g., oligodendrocytes) were represented in greater abundance. Another interesting association was with age. In particular, the fractions of neuronal types Ex3 and Ex4 significantly increased with age; by contrast, some non-neuronal types (e.g. oligodendrocytes) decreased (Fig. S2.8). These changes are potentially associated with differentially expressed genes. For example, Somatostatin (SST) expression decreases with age whereas its promoter methylation increases; other genes (e.g. EGR1 and CP) exhibit different trends (Figs. 2F, S2.10 and S2.11) (*15*).

Enhancers

Using an approach consistent with ENCODE, we used chromatin modification signals to identify enhancers active in the brain (*15*). We based this on the reference brain (see above), supplemented by the DNase and ChIP-seq data of the same brain region from Roadmap Epigenomics. Overall, we annotated a reference set of 79,056 enhancers active in PFC, enriched in H3K27ac and depleted in H3K4me3 (Fig. 3A).

Assessing the variability of enhancers across individuals and tissues is more difficult than performing the analogous comparison for gene expression. Not only does the chromatin signal change across the population, but the boundaries of enhancers grow and shrink, sometimes disappearing altogether (Fig. 3A). To investigate chromatin variability across the population, we uniformly processed the H3K27ac data from PFC, temporal cortex (TC), and cerebellum (CB) on a cohort of 50 individuals (*15*). Aggregating ChIP-seq data across the cohort resulted in a total of 37,761 H3K27ac "peaks" (enriched regions) in PFC, 42,683 in TC, and 26,631 in CB -- each of them present in more than half of the population. Comparing aggregate sets for these three brain regions, the PFC was more similar to TC than CB (~90% vs 34% overlap in H3K27ac peaks), consistent with previous reports (*22*).

We also examined the overlap of the reference brain enhancers with H3K27ac in each of the individuals. As expected, not every active enhancer in the reference annotation was active in every individual in the cohort. In fact, on average ~70% ± 15% (~54,000) of the enhancers in the reference brain were active in another individual in the cohort (Fig. 3B). As expected, only a core set of reference enhancers was ubiquitously active in every person, with a larger fraction (~68%) being active in more than half of the population. To estimate the total number of enhancers in PFC, we calculated the cumulative number of active regions across the cohort (Fig. S3.2). This number increased dramatically for the first 20 individuals sampled, but saturated at the 30th. Thus, we hypothesize that pooling the identified PFC enhancers from 30 individuals is sufficient to cover nearly all potential enhancers in PFC, estimated at ~120,000.

Consistent comparison: transcriptome and epigenome

As we uniformly processed the transcriptomic and epigenomic data across PsychENCODE, ENCODE, GTEx, and Roadmap datasets, we could compare the brain to other organs in a consistent fashion and also to compare across transcriptome and epigenome. We tried several approaches, including PCA, t-SNE, and reference component analysis (RCA) for an appropriate comparison. Although popular, PCA de-emphasizes local structure and can be easily influenced by outliers; in contrast, t-SNE preserves local relationships but “shatters” global structure (*15*). RCA is a compromise: it projects gene expression in an individual sample against a reference panel, and then reduces the dimensionality of the projections.

For gene expression, our comparison revealed that the brain separates from the other tissues in the first component (Fig. 3E). Inter-tissue differences were larger than intra-tissue ones (Fig. S4.1-4). A different picture emerged for chromatin: comparison showed that the chromatin levels at all regulatory positions were, overall, less distinguishable between brain and other tissues (Fig. 3C) (*15*). At first glance, this is surprising as one expects great differences in epigenetics between tissues. Note, however, our analysis compares chromatin signals over all non-coding regulatory elements from ENCODE (including enhancers and promoters), which is consistent with our expression comparison across all protein-coding genes (Fig. 3F vs. 3C). The total number of regulatory elements is much larger than brain-active enhancers (~1.3M vs. ~79K), so there are proportionately fewer brain-active regulatory elements than protein-coding genes (6% vs. 60%).

Our analysis focused on inter-tissue differences in annotated regions (i.e., genes, promoters, and enhancers). However, in addition to the canonical expression differences in protein-coding genes, we also found differences in unannotated non-coding and intergenic regions. In particular, testes and lung have the largest amount of transcriptional diversity overall for protein-coding genes (i.e., the most genes transcribed, Fig. 3D); however, when we shift to unannotated regions, brain tissues, such as cortex and cerebellum, now have a greater extent of transcription than any other tissue.

QTL analysis

We used the PsychENCODE data to identify QTLs affecting gene expression and chromatin activity. In particular, we calculated expression, chromatin, splicing-isoform, and cell-fraction QTLs (eQTLs, cQTLs, isoQTLs and fQTLs, respectively). For eQTLs, we adopted a standard approach, adhering closely to the established GTEx pipeline. In PFC, we identified ~2.5M cis-eQTLs (~238K independent SNPs after linkage-disequilibrium (LD) pruning) and ~33K eGenes (including non-coding ones) with FDR<0.05 (Fig. 4). We found ~1.3M SNPs involved in these from 5,297,875 tested in a 1 Mb window around genes. This conservative estimate has a substantially larger number of eQTLs and eGenes than previous studies and reflects the large PsychENCODE sample size (*15*). The number of eGenes, in fact, is approaching the total number of genes expressed in brain. We evaluated the similarity of GTEx and CMC eQTLs to our eQTL set using the π_1 statistic (*23*), finding a high replication rate (Fig. 4A). We also applied the same QTL pipeline to splicing, identifying ~160K isoQTLs (*15*).

For cQTLs, the situation is more complicated: no established methods exist for calculating these on a large scale, although there have been a variety of previous efforts (*24, 25*). To identify cQTLs, we focused on our reference set of enhancers and then examined how H3K27ac chromatin activity varied in these across 292 individuals (Fig. 4B) (*15*). Overall, we identified ~2,000 cQTLs in addition to the 6,200 identified using individuals from the CMC cohort (*26*).

Next, we determined if any SNPs were associated with changes in the relative fractions of cell types across individuals (fQTLs). In total, we identified 1672 distinct SNPs constituting 4199 fQTLs (Fig. S5.3). Of these, the proportions of excitatory neuron Ex4 and Ex5 were associated with the most. After factoring out these cell-type differences, we identified 200,729 SNPs significantly associated with gene expression changes across individual tissues; these "residual trans-eQTLs" represent variant-expression associations largely unexplained by changing proportions of cell types.

To further dissect the associations between genomic elements and the QTLs, we intersected our QTL lists with each other and a set of genomic annotations (Fig. 4D). As expected, eQTLs tended to be enriched at promoter regions, and cQTLs, at enhancer and TF-binding regions; fQTLs were spread over many different elements. Also, an appreciable number of eQTLs were enriched on the promoter of a different gene than the one regulated, suggesting e-promotor activity (*27*). For the overlap among different QTLs, we expected that most cQTLs, isoQTLs and fQTLs would be a subset of the much larger number of eQTLs; somewhat surprisingly, an appreciable number of these did not overlap (Fig. 4C). We calculated π_1 statistics to evaluate the sharing among eQTLs with other QTLs. We found that the eQTL sharing with cQTLs was the highest while that with fQTLs was lowest (0.89 vs 0.11). Moreover, the shared cQTLs often suggested that the expression-modulating function of an eQTL derived from chromatin changes (for example for the MTOR gene, Fig. 4C). Finally, there were 119 SNPs that functioned as QTLs in more than 3 different capacities (e.g. as eQTLs, cQTLs and isoQTLs), which we dubbed multi-QTLs.

Regulatory networks

We next integrated the genomic elements described above at the regulatory-network level. We created a network revealing how the genotype and regulators relate to target gene expression. We first processed a Hi-C dataset for adult brain in the same reference samples used for enhancer identification, providing a physical basis for interactions between enhancers and promoters (Fig. 5A) (*10, 15*). In total, we identified 2,735 topologically associating domains (TADs) and ~90K enhancer-promoter interactions (Fig. S6.1). Our adult Hi-C dataset substantially differed from an earlier fetal-brain Hi-C dataset (e.g. only ~31% of the interactions were detected in the fetal dataset) (*10*), highlighting the importance of the developmental stage for chromatin (Fig. S6.2 and S6.3).

As expected, ~75% of enhancer-promoter interactions occurred within the same TAD, and genes with more associated enhancers tended to have higher expression (Fig. 5B and S6.1). We next integrated the Hi-C data with the eQTLs and isoQTLs. Surprisingly, QTLs involving SNPs distal to the eGene but linked by Hi-C interactions showed significantly stronger associations than QTLs involving SNPs on the exons and promoters of the eGene (Fig. 5C and S6.4).

In addition to Hi-C and QTLs, we tried to predict further regulatory relationships based on directly relating the activity of transcription factors (TFs) to target genes (Fig. 5A). In particular, for each potential target of a TF, we required that (i) it has a "good binding site" (matching the TF's motif) in open chromatin regions near a gene (either in promoters or brain-active enhancers) and that (ii) it has a high coefficient in a regularized, elastic-network regression relating TF activity to target expression (*15*). Overall, we found the subset of interactions meeting these criteria could predict the expression of 8,930 genes with MSE< 0.05 (mean-square error, Fig. S6.5). For example, we could predict the expression of the ASD risk gene CHD8 with MSE=0.034 (*15*). Moreover, the subset of these interactions involving TFs binding to enhancers, necessarily instantiated a third set of putative enhancer-to-gene links.

Collectively, we generated a full regulatory network, linking enhancers, TFs, and target genes. It contained ~43k proximal linkages (TF-to-target gene via promoters), and ~37k distal linkages (enhancer-target-gene) that are supported by at least two of the three evidence sources (Hi-C, QTLs, or activity relationships)(*15*).

Linking GWAS variants to genes

We used our above regulatory network to connect non-coding GWAS loci to potential genes. We exploited all three possible evidence sources including Hi-C, QTLs, and activity relationships. For the newly identified 142 schizophrenia GWAS loci (*28*), we identified a set of 1,097 putative schizophrenia-associated genes, covering 119 loci (hereby referred as "SCZ-genes," Fig. 5E). 304 of these constitute a “high-confidence” set supported by more than two evidence sources (e.g., QTL and Hi-C, Figs. 5D-F and S7.1), exemplified by CACNA1C, which is regulated by multiple neuronal TFs via enhancers. The SCZ-genes represent a substantial increase from the previously reported 22 genes across 19 loci based on a smaller QTL set (*8, 28*) and also a much larger number than can be linked by simple genomic proximity (176, Fig. 5D). The majority of SCZ-genes were not in linkage disequilibrium with index SNPs (734 genes [~66%] with r^2<0.6, Fig. S7.1), consistent with previous observations that regulatory relationships often do not follow linear genome organization (*10*).

We then looked at the characteristics of the SCZ-genes. As expected, they shared many characteristics with known schizophrenia-associated genes. In particular, they were enriched for genes intolerant to loss-of-function mutations (*28*), translational regulators, cholinergic receptors, calcium channels, synaptic genes, and genes that are known to be differentially expressed in schizophrenia (Fig. S7.1). Next, we integrated SCZ-genes with single-cell profiles and found that they are highly expressed in neurons with the highest expression in excitatory neurons (Fig. 5G).

Finally, in a more general context, we found aggregate associations between our eQTLs and many brain-disorder GWAS variants, not just those for schizophrenia. In particular, compared to the GWAS-SNPs for non-brain related disorders, we found more significant enrichment for cis-eQTL SNPs and GWAS SNPs for many brain disorders (Fig. 4E). We find a similar, and, in fact, stronger enrichment for our brain-active enhancers (Fig. 4E).

Integrative deep-learning model

The full interaction between genotype and phenotype involves many levels, beyond those encapsulated in the regulatory network. We addressed this by embedding our regulatory network into a larger multilevel model. For this purpose, we developed an interpretable deep-learning framework, a Deep Structured Phenotype Network (DSPN) (*15*). This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from the gene regulatory network. As shown in Fig. 6A, traditional classification methods such as logistic regression predict phenotype directly from genotype, without inferring intermediates such as the transcriptome. In contrast, the DSPN (Fig. 6B) is constructed via a series of intermediate models that add layers of structure; these include intermediate molecular phenotypes (i.e., gene expression and chromatin state) and defined groupings of these (cell-type marker genes and co-expression modules), multiple higher layers for inferred groupings (hidden nodes), and a top layer for observed phenotypes (psychiatric disorders and other traits). Finally, we used special types of connectivity, including sparsity and lateral, intra-level relationships, to integrate our knowledge of QTLs, regulatory networks, and co-expression modules from sections above. By using a generative architecture, we ensure that the model is able to impute intermediate phenotypes, as well as provide forward predictions from genotypes to observed phenotypes.

Using the full model with the genome and transcriptome data provided, we demonstrated that the extra layers of structure in the DSPN allowed us to achieve substantially better prediction of diseases and traits than traditional additive models; further, the transcriptome carries additional information, which the DSPN is able to extract (Fig. 6D). For instance, a logistic predictor was able to gain a 2.4X improvement when using the transcriptome vs. the genome alone (+9.3% for transcriptome vs. +3.8% for the genome, above a 50% random baseline). In comparison, the DSPN was able to gain a larger 6X improvement (+22.9% vs. +3.8%), which may reflect its ability to incorporate non-linear interactions between intermediate phenotypes. Moreover, the DSPN also allows us to perform joint inference and imputation of intermediate phenotypes (i.e., transcriptome and epigenome, Fig. S8.1) and observed traits from just genotype alone, achieving a ~3.4X improvement over a logistic predictor in this context (Fig. 6D). These results demonstrate the usefulness of even a limited amount of functional genomic information for unraveling gene-disease relationships and show that the structure learned from such data can be used to make more accurate predictions of observed traits even when absent.

We transformed our results to the liability scale for comparison with narrow-sense heritability estimates (Fig. 6D) (*15*). Prior studies have estimated that common SNPs explain 25.6%, 20.5%, and 19% of the genetic variance for SCZ, BPD and ASD, respectively (*29*). These may be taken as upper bounds for additive predictive models, given unlimited common-variant data; by contrast, non-linear predictors can potentially exceed these limits. Our best liability scores (based on just the genotype at QTL-associated variants) are substantially below these bounds, implying that additional data will be beneficial. In contrast, the variance explained by the full DSPN model exceeds that explained by common SNPs (32.8%, 37.4%, and 14.4%, respectively for the three conditions), possibly reflecting the influence of rare variants and epistatic interactions; the degree to which this variance may be captured through improved imputation however is limited by the proportion of total variance in the imputed variables which is genetically determined (Fig. S8.2).

A key aspect of the DSPN is its interpretability. In particular, we examined the specific connections learned by the DSPN between intermediate and high-level phenotypes. We included known co-expression modules in the model (DSPN-mod) and examined which of these the DSPN prioritized, as well as new sets of genes associated with latent nodes that were uncovered at each hidden layer (Fig. S8.3 and Table S8.1) (*15*). We provide a full summary of the enrichment analysis for the prioritized modules and highlight some of the associations found using the schizophrenia model (Fig. 6C and S8.4). Overall, we show the modules prioritized by the DSPN were enriched for known SCZ and BPD GWAS variants (Fig. S8.5). In particular, among the highest schizophrenia-prioritized modules and higher-order groupings, we found enrichments for (i) glutamatergic-synapse pathway genes, (ii) calcium-signaling pathways and astrocyte-marker genes, and (iii) complement cascade pathway genes including C4A, C4B, and CLU -- confirming and extending previous analyses (*30*). Furthermore, for groupings prioritized for aging, we found enrichment in Ex4 cell-type genes and the specific gene NRGN (in a module associated with synaptic and longevity functions), both consistent with differential expression analysis (Fig. S8.4 and S2.10).

Conclusion

Here, we uniformly integrated PsychENCODE datasets with other datasets, developing a comprehensive resource for functional genomics of the adult brain. Overall, our study identified a set of eQTLs several fold greater than previous studies, achieving close to saturation for protein-coding genes. Our data are consistent with the stage and tissue specific nature of gene regulation, indicating that it will be valuable to profile different regions and developmental stages at similar scale. It also indicates that increasing individual sample size and quality of chromatin data, such as identifying enhancers via STARR-seq, will help with cQTLs. More fundamentally, one-dimensional fluctuations in chromatin signal reflect changes in three-dimensional changes in architecture and new metrics beyond cQTLs may need to be developed to measure chromatin variation better. In addition, some other epigenetic marks might exhibit distinguishable patterns in the brain, e.g. the methylation landscape. Likewise, inter-tissue expression comparisons might be boosted by including microRNAs. Nevertheless, using current approaches, we were able to identify over 300 high confidence SCZ risk genes, implicated by 142 published genome wide significant loci and supported by at least two independent methods. This is more than an order of magnitude higher number of SCZ-genes than identified in previous studies, highlighting the power of our sample size and integrative approach.

Another area for future development is single-cell analysis. In this study, we found that varying proportions of basic cell types (with different expression signatures) could explain a large fraction of expression variation across the human population. This assumes that expression signatures, at least for biomarker genes, are fairly constant over same cell types. Larger-scale single cell studies will allow us to examine this assumption in greater detail, perhaps quantifying and bounding environment-associated transcriptional variability. In addition, current single-cell techniques suffer from low capture efficiency; thus, it remains challenging to reliably quantify low-abundance transcripts (*12, 31*). This is particularly the case for specific cell sub-structures such as axons and dendrites (*12*).

Further, we envision how our DSPN deep-learning approach can be readily extendable to modeling genotype-phenotype relationships involving other kinds of intermediate phenotypes (e.g., from brain imaging); we can naturally embed new types of QTLs and phenotype-phenotype interactions. Comparison of the variance explained in terms of liability when particular intermediate phenotypes are imputed versus known provides natural bounds on the variance in observed traits mediated by these phenotypes. Finally, although our focus has been on common SNPs, the DSPN may be capturing the effects of rare variants through their influence on intermediate phenotypes; the interpretable structure of the model may help identify such variants by their association with prioritized phenotypes and higher-order groupings.

In summary, our integrative analyses here and with respect to the disease and developmental transcriptome (*16, 32*) demonstrate that functional annotation of gene regulatory elements is useful for unraveling molecular mechanisms in the brain.

Acknowledgment

We would like to acknowledge the National Institute of Mental Health (NIMH) for funding. Also, we acknowledge program staff, in particular Thomas Lehner, Lora Bingaman, David Panchision, Alexander Arguello and Geetha Senthil, for providing institutional support and guidance for this project.

Figures

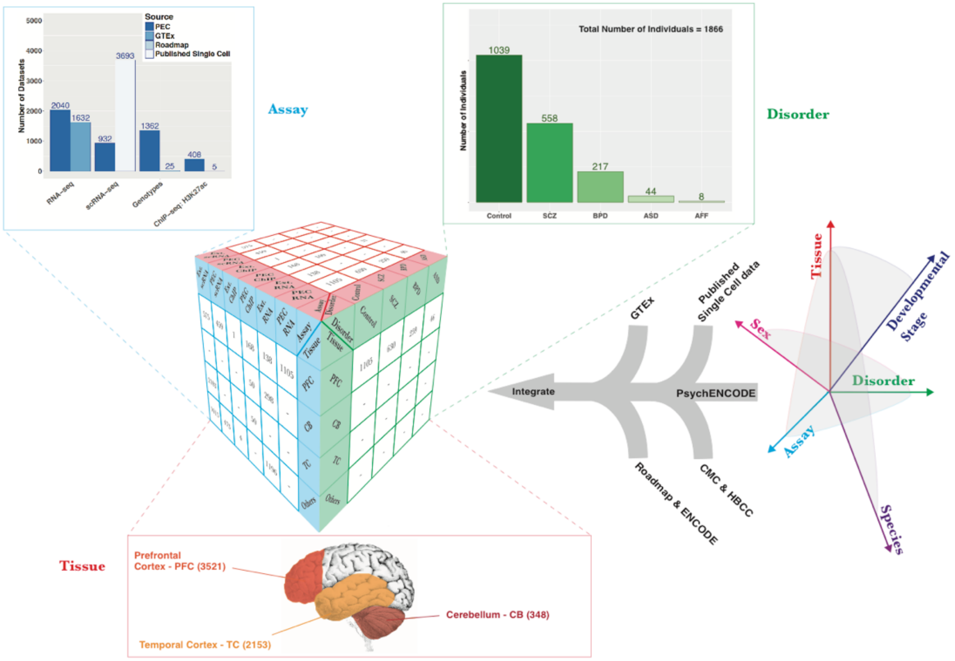


Figure 1. Comprehensive data resource for functional genomics in the adult brain. The functional genomics data generated by the PsychENCODE consortium (PEC) constitute a

multidimensional exploration across tissue, developmental stage, disorder, species, assay, and sex. From this larger corpus of PEC samples, we focused on adult datasets, integrated with those from consortia such as GTEx, the Roadmap Epigenomics Consortium, ENCODE, CMC and Human Brain Collection Core studies, and previously published single-cell transcriptomic data. The central data cube represents the results of this integration for the three dimensions of disorder, assay, and tissue, where only the numbers of datasets used in the current analysis are depicted. Projections of the data onto each of these three parameters are shown in graph form for assay and disorder, and in schematic form for the primary brain regions of interest. **Assay:** The bars represent datasets across a subset of the assay types, including RNA-seq (N = 2040 PEC + 1632 uniformly processed GTEx samples), genotypes (N = 1362 PEC + 25 GTEx = 1387 individuals matched to RNA-seq samples for eQTL analysis), scRNA-seq (N = 932 PEC + 3693 external datasets), and H3K27ac ChIP-seq (= 408 PEC + 5 uniformly processed Roadmap samples). **Disorder:** The number of individuals under the control category include the 113 from GTEx and 926 from PEC, while individuals from PEC provide data on the remaining disorders of schizophrenia (SCZ, N = 558), bipolar disorder (BPD, N = 217), ASD (N = 44), and affective disorder (AFF, N = 8), resulting in a total of 1,866. **Tissue:** In this schematic, we focus on the datasets derived from three primary brain regions evaluated in our integrative study: the prefrontal cortex (PFC, N = 3521), the temporal cortex (TC, N = 2153), and the cerebellum (CB, N = 348). See supplement (*15*) and Adult.psychencode.org for more details.

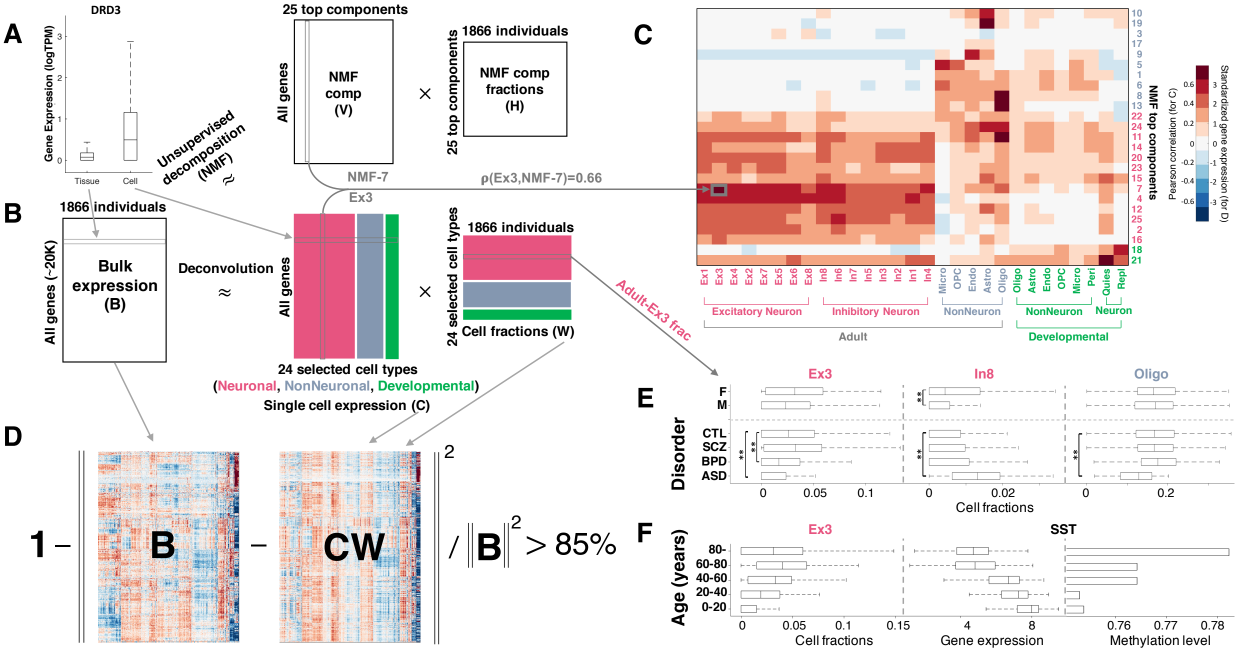


Figure 2. Deconvolution analysis of bulk and single-cell transcriptomics reveals cell fraction changes across the population. **(A)** Genes had significantly higher expression variability across single cells, sampled from different types of brain cells, than equivalent tissue samples, taken from a population of individuals. Left: dopamine gene, DRD3. **(B)** Top: the bulk tissue gene expression matrix (B, genes by individuals) can be decomposed by NMF into the product of two matrices: an NMF component matrix (V, genes by top NMF components; i.e.,

NMF-TCs) and a component fraction matrix (H, top NMF components by individuals); i.e., B≈VH. Bottom: the bulk tissue gene expression matrix B can be also deconvolved by the single-cell gene expression matrix (C, genes by cell types) to estimate the cell fractions across individuals (the matrix, W); i.e., B≈CW. Three major cell types were neuronal cells (blue), non-neuronal cells (red), and developmental (dev) cells (green), as highlighted by columns groups in C (also row groups in W). **(C)** The heatmap shows the Pearson correlation coefficients of gene expression between the NMF-TCs and single-cell signatures (for N=457 biomarker genes, see (*15*)). For example, NMF-7 is highly correlated with the Ex3 cell type (r=0.66). **(D)** The estimated cell fractions can explain >85% of bulk tissue expression variation across the population; i.e., $1-||B-CW||^2/||B||^2>0.85$. **(E)** The cell fractions changed across genders (just control samples) and brain disorders. In particular, the neuronal cell types (e.g. In8) had a significantly higher fraction in female than male samples (p<1.2e-4). Disorder types that showing significant changes compared to control samples after accounting for age distributions are labeled (**). For example, Ex3 neuronal cells and oligodendrocytes had lower fractions in ASD than other cell types. **(F)** Across age, changing cell fractions (for Ex3), gene expression (for SST) and promoter methylation level (median level, for SST) are shown. Note, the excitatory neuronal cell type Ex3 had a significant increase with age (trend analysis p<6.3e-10).

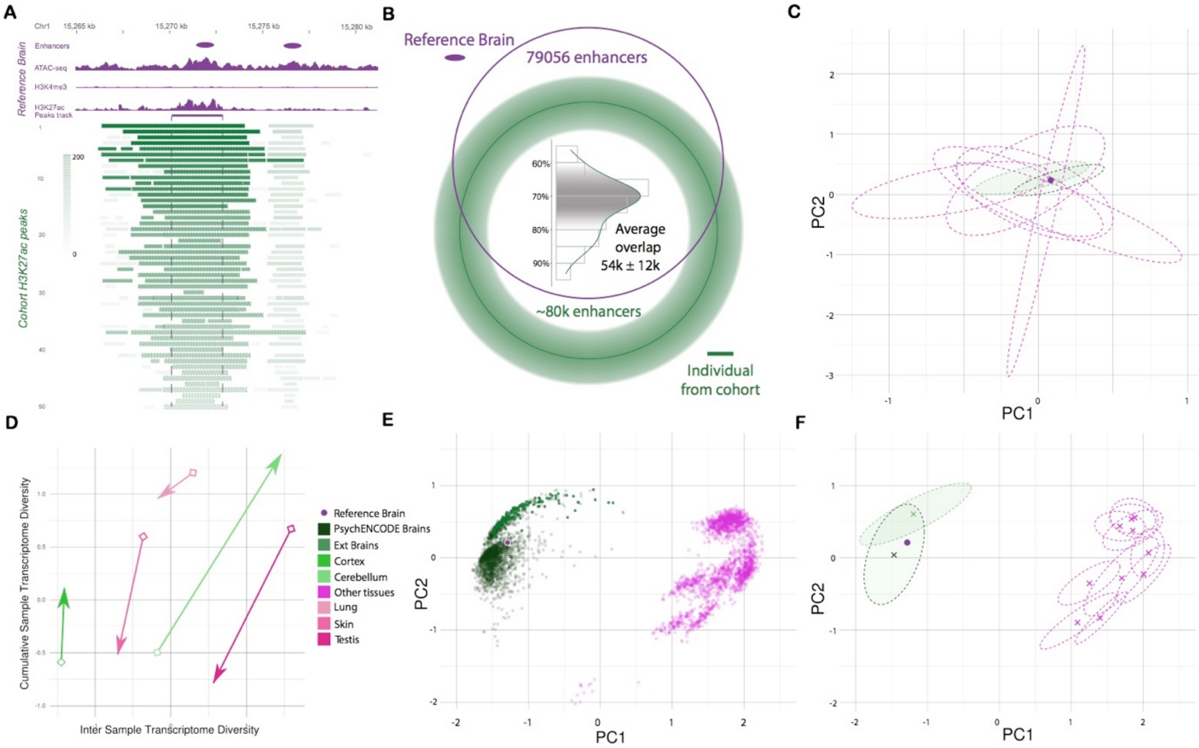
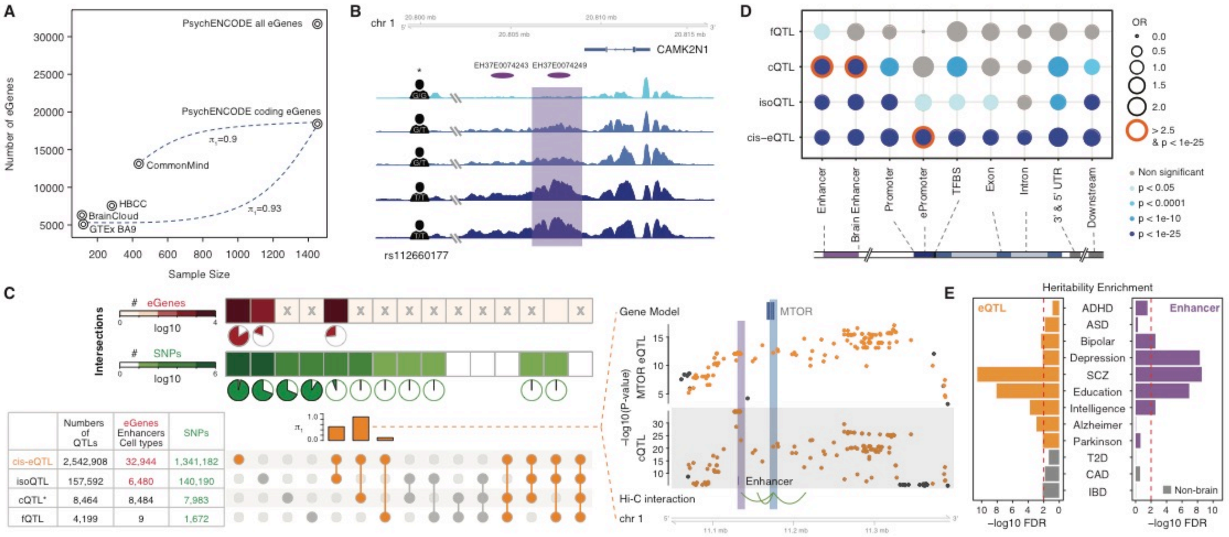


Figure 3. Comparative analysis for transcriptomics and epigenomics between brain and other tissues. **(A)** Chromatin features of the reference brain (purple dot) were used to identify active enhancers, located in the open chromatin regions (as manifest by ATAC-seq peaks), with

strong H3K27ac/H3K4me1 signal and low H3K4me3 signal. Enhancer activity varied among individuals, as indicated by the varying H3K27ac peaks at the enhancer region in the population. Each row corresponds to an individual in the cohort (green), with the gradient showing the normalized signal value for each peak **(B)** The overlap of individual H3K27ac peaks with reference brain enhancers in the population is shown as the Venn diagram. The histogram shows the varying percentage of H3K27ac peaks across individuals. **(C)** The tissue clusters of RCA coefficients (PC1 vs. PC2) for chromatin data of any potential regulatory elements are shown. Clusters of PsychENCODE samples (dark green ellipses), Roadmap Epigenomics brain samples (light green ellipses), and other non-brain tissues (magenta ellipses) are plotted. The reference brain is shown as the purple dot (same in E and F). *Panels E and F are drawn similarly to D, but now for transcription rather than epigenetics*. **(E)** The coefficients (PC1 vs. PC2) of RCA analysis for gene expression data of PsychENCODE samples are shown in dark green. The brain samples from GTEx are shown in light green, and other tissue samples are shown in magenta. **(F)** The center (cross) and ranges of different tissue clusters (dashed ellipses) are shown on an RCA scatterplot of (**E**). Finally, **(D)** The transcriptional diversity for coding (circle) and non-coding (triangle) regions among the tissue samples (inter-sample on x-axis) is shown compared to the diversity on cumulative tissue samples (y-axis) for select tissue types including cerebellum, cortex, lung, skin, and testes, using PolyA RNA-seq data.



	Numbers of QTLs	eGenes Enhancers Cell types	SNPs
cis-eQTL	2,542,908	32,944	1,341,182
isoQTL	157,592	6,480	140,190
cQTL*	8,464	8,484	7,983
fQTL	4,199	9	1,672

Figure 4. Summary of QTLs in adult brain PFC. **(A)** Numbers of genes with at least one eQTL (eGenes) are shown compared to the sample size in different studies. The number of eGenes increased as the sample size increased. The eGenes of PsychENCODE is close to saturation for protein coding genes. The estimated replication π_1 values of GTEx and CMC eQTLs versus PsychENCODE are 0.93 and 0.90 respectively (*23*). **(B)** Example of H3K27ac signal of individual brains in a representative genomic region showing largely congruent identification of regions of open chromatin. Region in the dashed frame represents a chromatin QTL; the signal magnitudes of individuals with a G/G or G/T genotype were lower than the ones with a T/T genotype. **(C)** Numbers of identified QTLs, associated elements (eGenes, enhancers, and cell types) and QTL SNPs are shown in the left table. *For cQTLs we only show the number of top SNPs for each enhancer. Overlap of eQTL, isoQTL, fQTL, and cQTL SNPs and overlap of eQTL and isoQTL eGenes are shown. Overlap numbers are shown in heatmaps while

overlap percentage are shown with pies. Sharing of the QTLs vs. eQTLs are shown using π_1 values in the orange bar plot indicating the highest sharing is between cQTLs vs. eQTLs. An example on the right for the MTOR gene shows the overlapping (based on co-localization analysis) of eQTL SNPs for expression of the gene and cQTL SNPs for the H3K27ac signal on an enhancer ~50kb upstream of the gene. Hi-C interactions demonstrate that the enhancer interacts with the promoter of MTOR, indicating that the cQTL SNPs potentially mediate the expression modulation manifest by the eQTL SNPs. **(D)** Enrichment of genomic regions annotations of QTLs is shown. **(E)** Brain disorder GWAS show stronger heritability enrichment in brain regulatory variants (eQTLs) and elements (enhancers) than non-brain disorder GWAS. ADHD, attention-deficit/hyperactivity disorder; T2D, type 2 diabetes; CAD, coronary artery disease; IBD, inflammatory bowel disease.

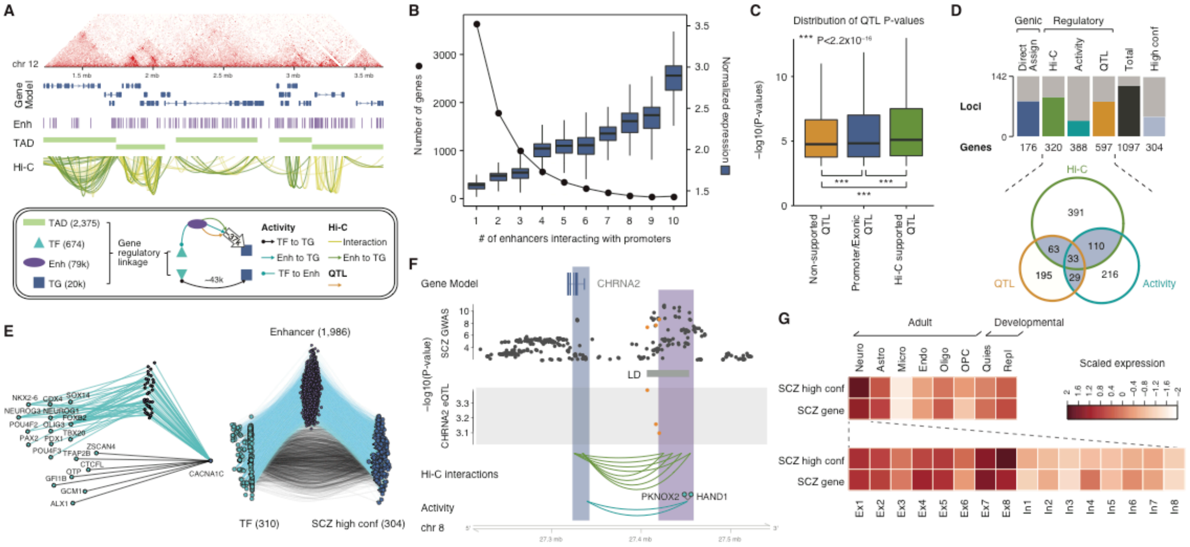
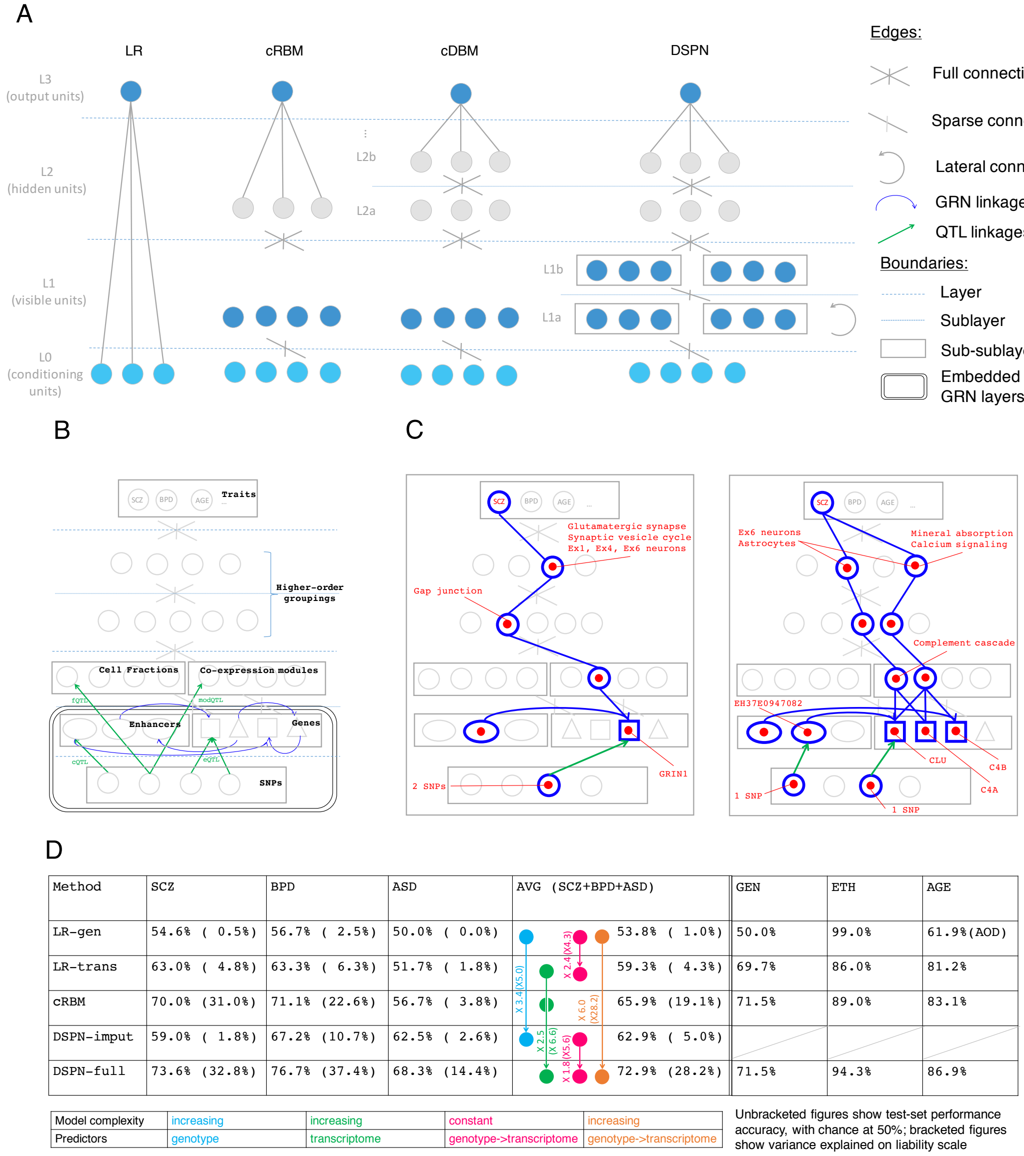


Figure 5. Data integration and modeling predicts a gene regulatory network, linking additional GWAS genes for psychiatric disorders. **(A)** A full Hi-C data from adult brain reveals the large-scale structure of the genome, ranging from contact maps (top), TADs, and promoter-based interactions. We leveraged gene regulatory linkages involving TADs, TFs, enhancers, and target genes to build a full gene regulatory network consisting of ~150,000 Hi-C interactions, ~2.5 million eQTL-eGene linkages, ~211k TF-to-target and ~448k enhancer-to-target-promoter linkages based on activity relationships. **(B)** We compared the number of genes (left y-axis, dotted line) and the normalized gene expression levels (right y-axis, boxes) with the number of enhancers that interact with the gene promoters. **(C)** QTLs that were supported by Hi-C evidence showed more significant P-values than those that were not. **(D)** The number of schizophrenia GWAS loci and their putative target genes (SCZ-genes) annotated by each assignment strategy. SCZ-genes with more than 2 evidence sources were defined as SCZ high-confidence (high conf.) genes. The overlap between SCZ-genes defined by QTL associations (QTL), chromatin interactions (Hi-C), and activity relationships (Activity) is depicted in a Venn diagram at the bottom. **(E)** A gene regulatory network of TFs (cyan), enhancers (purple), and 304 highly confident SCZ high-confidence genes (blue) as targets, based on TF activity linkages. A subnetwork including multiple neuronal TFs targeting the SCZ gene CACNA1C via enhancers is highlighted on the left. **(F)** Evidence depicting that GWAS SNPs that overlap with

CHRNA2 eQTLs also have chromatin interactions and activity correlations with the same gene. **(G)** SCZ-genes show higher expression levels in neuronal cell types (particularly excitatory neurons) than others cell types.



Method	SCZ	BPD	ASD	AVG (SCZ+BPD+ASD)	GEN	ETH	AGE
LR-gen	54.6% (0.5%)	56.7% (2.5%)	50.0% (0.0%)	53.8% (1.0%)	50.0%	99.0%	61.9%(AOD)
LR-trans	63.0% (4.8%)	63.3% (6.3%)	51.7% (1.8%)	59.3% (4.3%)	69.7%	86.0%	81.2%
cRBM	70.0% (31.0%)	71.1% (22.6%)	56.7% (3.8%)	65.9% (19.1%)	71.5%	89.0%	83.1%
DSPN-imput	59.0% (1.8%)	67.2% (10.7%)	62.5% (2.6%)	62.9% (5.0%)			
DSPN-full	73.6% (32.8%)	76.7% (37.4%)	68.3% (14.4%)	72.9% (28.2%)	71.5%	94.3%	86.9%

Model complexity	increasing	increasing	constant	increasing
Predictors	genotype	transcriptome	genotype->transcriptome	genotype->transcriptome

Unbracketed figures show test-set performance accuracy, with chance at 50%; bracketed figures show variance explained on liability scale

Figure 6. Deep-learning model predicts genotype-phenotype relations and reveals intermediate molecular mechanisms. **(A)** The schematic outlines the model structures for Logistic Regression (LR), conditional Restricted Boltzmann Machine (cRBM), conditional Deep Boltzmann Machine (cDBM), and Deep Structured Phenotype Network (DSPN) models. Nodes are partitioned into four possible layers (L0-L3) and colored according to their status as (i) conditioning nodes visible during training and testing (light blue); (ii) nodes visible during training and visible or imputed during testing (dark blue); or (iii) hidden nodes (grey). **(B)** The DSPN structure is shown in further detail, with the biological interpretation of layers L0, L1, and L3 highlighted. The gene regulatory network (GRN) structure learned previously (Fig. 5) is

embedded in layers L0 and L1, with different types of regulatory linkages and functional elements shown. **(C)** Shown are examples of associations found: model traces are shown for three co-expression modules and associated higher-order groupings prioritized by the DSPN schizophrenia model, along with functional annotations enriched at each level. Genes, enhancers, and SNPs associated with each module are shown. **(D)** The performance of different models is summarized, comparing performance (i) across models of different complexity, and (ii) using transcriptome vs. genome predictors, corresponding to with/without imputation for the DSPN (colors highlight relevant models for each comparison). Performance accuracy on a balanced sample is shown first, with variance explained on the liability scale shown in brackets. LR-gen and LR-trans are logistic models using the genotype and transcriptome as predictors respectively; DSPN-imput and DSPN-full are the DSPN model with imputed intermediate phenotypes (genotype predictors only) and fully observed intermediate phenotypes (transcriptome predictors) respectively. Differential performance of models is shown in terms of improvement above chance, for instance comparing LR-gen and DSPN-imput accuracy improves from 53.8% to 62.9%, which can be expressed as a 3.4X improvement above chance (+12.9% vs. +3.8%, blue). Corresponding improvements in liability variance scores are shown in brackets. Disorders are abbreviated as in the main text, and GEN=Gender, ETH=Ethnicity, AOD=Age of death.

Reference

1. R. C. Kessler *et al.*, Design and field procedures in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A). *Int J Methods Psychiatr Res* **18**, 69-83 (2009).
2. P. W. Wilson *et al.*, Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-1847 (1998).
3. N. Cancer Genome Atlas Research *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120 (2013).
4. D. M. Lloyd-Jones *et al.*, Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. *Circulation* **113**, 791-798 (2006).
5. M. R. Stratton, P. J. Campbell, P. A. Futreal, The cancer genome. *Nature* **458**, 719-724 (2009).
6. G. C. C. C. Psychiatric *et al.*, Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry* **166**, 540-556 (2009).
7. D. H. Geschwind, J. Flint, Genetics and genomics of psychiatric disease. *Science* **349**, 1489-1494 (2015).
8. M. Fromer *et al.*, Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).
9. C. Colantuoni *et al.*, Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**, 519-523 (2011).
10. H. Won *et al.*, Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527 (2016).
11. B. B. Lake *et al.*, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590 (2016).
12. A. E. Saliba, A. J. Westermann, S. A. Gorski, J. Vogel, Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42**, 8845-8860 (2014).

13. S. Darmanis *et al.*, A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290 (2015).
14. E. C. Psych *et al.*, The PsychENCODE project. *Nat Neurosci* **18**, 1707-1712 (2015).
15. Materials and methods are available as supplementary materials.
16. M. J. Gandal, e. al., Dysregulation of cortical splicing, isoform and noncoding gene regulatory networks in ASD, schizophrenia, and bipolar disorder. *submitted*.
17. I. Voineagu *et al.*, Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384 (2011).
18. M. J. Gandal *et al.*, Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693-697 (2018).
19. M. C. Oldham *et al.*, Functional organization of the transcriptome in human brain. *Nat Neurosci* **11**, 1271-1282 (2008).
20. T. E. Bakken *et al.*, A comprehensive transcriptional map of primate brain development. *Nature* **535**, 367-375 (2016).
21. A. E. Jaffe *et al.*, Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat Neurosci* **18**, 154-161 (2015).
22. W. Sun *et al.*, Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* **167**, 1385-1397 e1311 (2016).
23. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445 (2003).
24. R. C. del Rosario *et al.*, Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nat Methods* **12**, 458-464 (2015).
25. F. Grubert *et al.*, Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-1065 (2015).
26. J. Bryois *et al.*, Evaluation Of Chromatin Accessibility In Prefrontal Cortex Of Schizophrenia Cases And Controls. *bioRxiv*, (2017).
27. L. T. M. Dao *et al.*, Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet* **49**, 1073-1081 (2017).
28. A. F. Pardinas *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* **50**, 381-389 (2018).
29. V. Anttila *et al.*, Analysis of shared heritability in common disorders of the brain. *bioRxiv*, (2017).
30. A. Sekar *et al.*, Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177-183 (2016).
31. S. Liu, C. Trapnell, Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5**, (2016).
32. M. Li, e. al., Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risk. *submitted*.

‡The PsychENCODE Consortium:

Alexej Abyzov, Mayo Clinic Rochester; Ashley Adams, The University of Chicago; Schahram Akbarian, Mount Sinai; Majd Alsayed, The University of Chicago; Anahita Amiri, Yale University; Joon Yong An, University of California, San Francisco; Alexander Arguello, NIMH; Christoper Armoskus, University of Southern California; Allison Ashley-Koch, Duke University; Lora Bingaman, NIMH; Miguel Brown, The University of Chicago; Mimi Brown, The University of Chicago; Julien Bryois, Karolinska Institutet; Adrian Camarena, University of Southern California; Becky Carlyle, Yale University; Alexander Charney, Mount Sinai; Chao Chen, Central South University; Lijun Cheng, The University of Chicago; Jinmyung Choi, Yale University; Declan Clarke, Yale University; Leonardo Collado-Torres, Lieber Institute; Gianfilippo Coppola, Yale University; Greg Crawford/Sullivan, Duke University; Diane DelValle, Mount Sinai; Stella

Dracheva, Mount Sinai; Tara Dutka, NIMH; Prashant Emani, Yale University; Oleg Evgrafov, SUNY Downstate Medical Center; Peggy Farnham, University of Southern California; Dominic Fitzgerald, The University of Chicago; Nancy Francoeur, Mount Sinai; Mike Gandal, University of California, Los Angeles; Tianliuyun Gao, Yale University; Mark Gerstein, Yale University; Dan Geschwind, University of California, Los Angeles; Gina Giase, University of Illinois at Chicago; Kiran Girdhar, Mount Sinai; Fernando Goes, Johns Hopkins University; Kay Grennan, SUNY Upstate Medical University; Mengting Gu, Yale University; Gamze Gursoy, Yale University; Evi Hadjimichael, Mount Sinai; Liza Herendeen, The University of Chicago; Gabriel Hoffman, Mount Sinai; Thomas Hyde, Lieber Institute; Nikolay Ivanov, Lieber Institute; Andrew Jaffe, Lieber Institute; Yan Jiang, Mount Sinai; Yi Jiang, Central South University; Amira Kefi, University of Illinois at Chicago; Yunjung Kim, University of North Carolina - Chapel Hill; James Knowles, SUNY Downstate Medical Center; Alexey Kozlenkov, Mount Sinai; Thomas Lehner, NIMH; Mingfeng Li, Yale University; Zhen Li, Yale University; Shuang Liu, Yale University; Chunyu Liu, SUNY Upstate Medical University; Lara Mangravite, Sage Bionetworks; Eugenio Mattei, University of Massachusetts; Angus Nairn, Yale University; Fabio Navarro, Yale University; Mingming Niu, St. Jude Children's Hospital; Temi Okubadejo, University of Chicago; Xinghua Pan, Yale University; David Panchision, NIMH; Junmin Peng, St. Jude Children's Hospital; Mette Peters, Sage Bionetworks; Dalila Pinto, Mount Sinai; Sirisha Pochareddy, Yale University; Damon Polioudakis, University of California, Los Angeles; Amanda Price, Lieber Institute; Michael Purcaro, University of Massachusetts; Tim Reddy, Duke University; Suhn Rhie, University of Southern California; Panagiotis Roussos, Mount Sinai; Tanmoy Roychowdhury, Mayo Clinic Rochester; Stephan Sanders, University of California, San Francisco; Gabriel Santpere, Yale University; Soraya Scuderi, Yale University; Geetha Senthil, NIMH; Nenad Sestan, Yale University; Xu Shi, Yale University; Annie Shieh, SUNY Upstate Medical University; Mario Skarica, Yale University; Andre Sousa, Yale University; Valeria Spitsyna, University of Southern California; Matthew State, University of California, San Francisco; Patrick Sullivan, University of North Carolina - Chapel Hill; Vivek Swarup, University of California, Los Angeles; Jin Szatkiewicz, University of North Carolina - Chapel Hill; Anna Szekely, Yale University; Junko Tsuji, University of Massachusetts; Flora Vaccarino, Yale University; Ramu Vadukapuram, SUNY Upstate Medical University; Harm van Bakel, Mount Sinai; Daifeng Wang, Yale University; Xusheng Wang, St. Jude Children's Hospital; Jonathan Warrell, Yale University; Sherman Weissman, Yale University; Zhiping Weng, University of Massachusetts; Donna Werling, University of California, San Francisco; Kevin White, University of Chicago; Jeremy Willsey , University of California, San Francisco; Heather Witt, University of Southern California; Hyejung Won, University of California, Los Angeles; Shannon Wood, University of Southern California; Susan Wright, NIMH; Feinan Wu, Yale University; Yan Xia, SUNY Upstate Medical University; Min Xu, Yale University; Yucheng Yang, Yale University; Peter Zandi, Johns Hopkins University; Ying Zhu, Yale University.

Supplement to
Comprehensive functional genomic resource and integrative model for the adult brain

Quick Guide to Finding Information in the Supplement

To most clearly link and cross-reference between the main text and this supplement, we use common section headings for both. Thus, *supplementary content to a given main text section within the supplementary section is named in a parallel fashion*:

Supplementary methods, figures and tables are contained within their respective sections. Items are numbered to provide ease of access to supplementary content. For instance, “Fig. S2.1b” refers to panel b within the first figure within section S2. Also, note that many associated data files are on the website Adult.psychENCODE.org (S9).

Introduction on PsychENCODE data & more on the supplement

This document provides an organized reference to support datasets, pipelines, and analyses associated with this study. It is presented in a parallel fashion to the main text. It is also connected to the main text through the major results presented in the form of main text figures – captions associated with main text figures point to relevant subsections within this supplement. In cases where the related supplementary section is not readily apparent, we note "see supp. section xyz” to refer to a specific section.

Large datasets produced by the psychENCODE consortium include over 2,000 human brain samples for healthy controls and individuals afflicted by neuropsychiatric diseases. These include full genotyping, RNA-seq, ChIP-seq, and single-cell data. It also includes processed data such as expression QTLs and chromatin QTLs trait loci, enhancers that are active in different brain regions, in addition to differentially expressed genes, transcripts, and novel non-coding RNAs. These are also provided at the resolution of brain sub-regions, thereby providing valuable resources for investigating potential underlying factors for an array of psychiatric diseases.

However, the very richness of this data introduces considerable challenges with respect to data organization. Our analyses rely on multiple methodologies, the details of which are difficult to include within the main text of this paper.

The data resources may be organized into a pyramid-like structure, with large raw data files at the base, and more processed summary data organized at higher levels. The raw data files include datasets from PyschENCODE, ENCODE, CommonMind, GTEx, Epigenomics Roadmap, and others. These comprise RNA-seq expression quantification data, ChIP-seq signal track qualifications and peak identifications using ENCODE standard pipelines, in addition to private data such as imputed genotypes. Further up the pyramid, more readily human-interpretable data and descriptors populate the top. These more processed datasets include patient metadata and phenotypes (such as disease status), fully processed epigenomic signals and peaks, active enhancers, QTLs, differentially expressed genes and transcripts, and regulatory networks.

With the aim of presenting data and results (including software packages) in an organized way, we have written about this study in roughly a hierarchical fashion. The main text lies at the top of this hierarchy and synthesizes everything in a broad manner. It refers to more detailed descriptions of our methods and datasets, as provided in this supplement. Raw data files, which lie at the bottom of the hierarchy (and which are hosted as online resources) form the bedrock from which our results are built.

S1. Supp. content to main text section "Resource construction"

The PsychENCODE data covers a number of phenotypes on mental health. These include normal controls (n=1104), as well as schizophrenia (n=558), bipolar (n=217), autism spectrum disorder, (n=44), and affective disorder (n=8) (Fig. 1). There are 1246 males and 685 females. We integrated standard pipelines to uniformly process raw sequencing and genotyping data (Fig. S1.1). Details are provided in following Sections S2.1, S2.2, S3.1, S5.1, and S6.1-6.2.

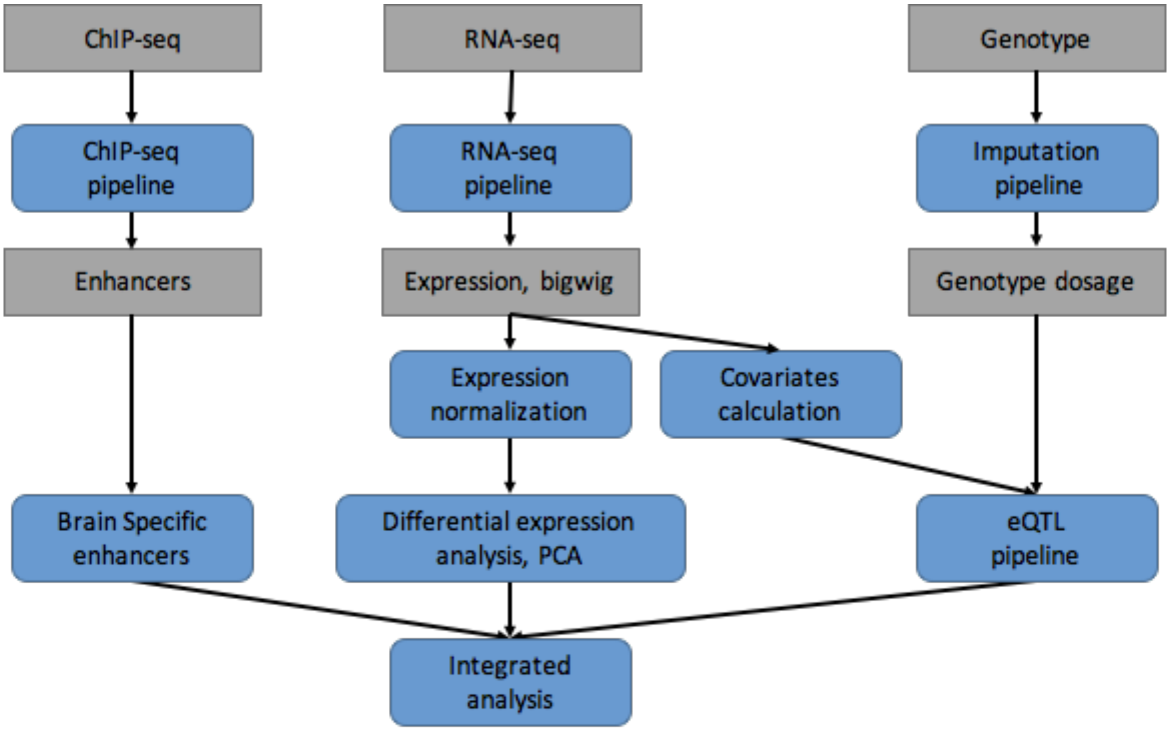


Fig. S1.1 - Integrated analysis pipeline of PsychENCODE. We used the standard pipelines from ENCODE and other large consortia to uniformly processed the raw sequencing data from PsychENCODE, including RNA-seq, ChIP-seq and Genotype, and identified functional genomic elements such as brain enhancers, expressed genes and eQTLs. We also processed other data types such as Hi-C and single cell and provided details on data processing in the following sections. As shown by this flowchart, we then performed the integrative modeling and analysis for functional genomic elements in adult brain.

S2. Supp. content to main text section "Transcriptome analysis”

S2.1 Data processing

Note that the data files for this section are described in detail in Section S9 (Resource website).

S2.1.1 GTEx brain and other tissues

We used several types of data from the GTEx version 7 dataset (GTEx Consortium, 2017). GTEx version 7 contains RNA-seq and matching genotype data for 10 brain regions: anterior cingulate cortex, caudate nucleus, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens, and putamen. We used the raw RNA-seq data to quantify the proportion of the transcribed non-coding genome. For eQTL calculations and weighted gene co-expression network analysis (WGCNA) analysis, we used individual trusted platform module (TPM) data, and renormalized it using probabilistic estimation of expression residuals (PEER) factors calculated in combination with PsychENCODE data. Further, for the eQTL calculations, we re-imputed the genotype data from the raw genotype calls using the pipeline described below to match the processing of the PsychENCODE data.

We used data from GTEx7 (GTEx Consortium, 2017) to compare the brain transcriptome to that of other tissues. GTEx7 contains RNA-seq data from 34 other tissues. As above, we used the raw RNA-seq data to quantify the proportion of transcribed non-coding regions. For WGCNA analysis, we used the individual TPM data, pre-normalized by the PEER factors calculated in GTEx7 to identify modules in individual tissues, and the median TPM data by tissue to identify modules across tissues.

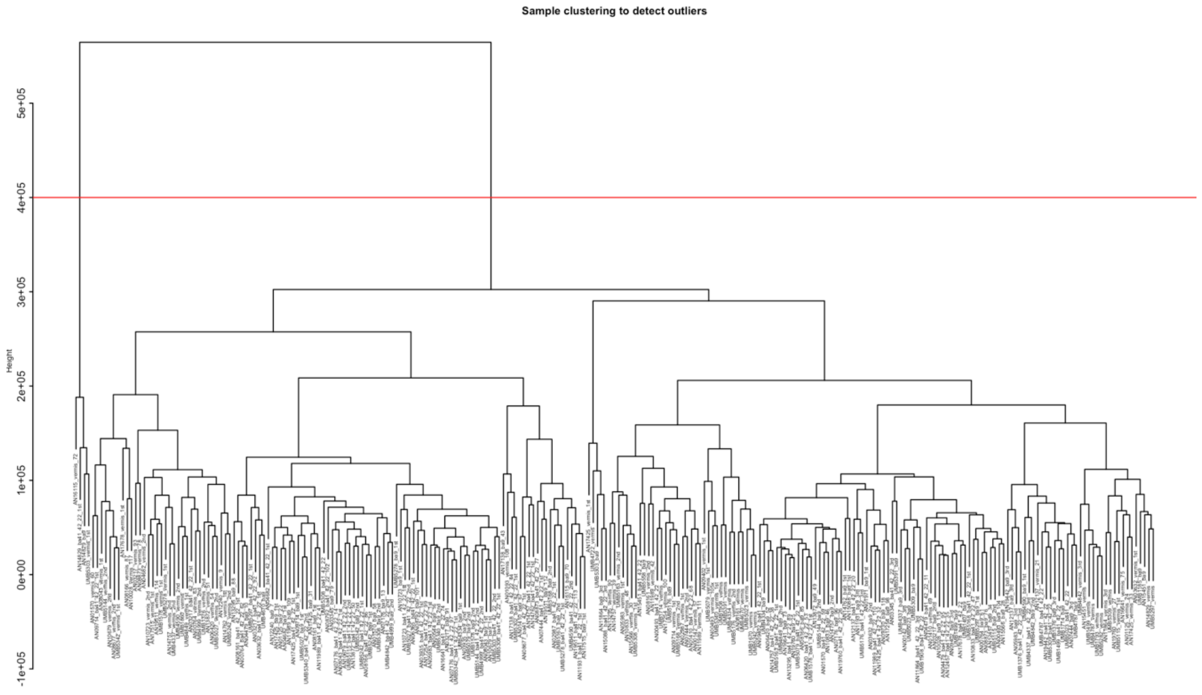


Fig. S2.1 Dendrogram of clustering analysis for identifying outliers of gene expression. An example of removing 4 outlier samples from a UCLA-ASD study according to hierarchical clustering of the gene expression data.

Table S2.1 Summary of dataset. This table provides the number of samples incorporated into the integrative analyses in this manuscript, categorized by study, the disease status of the individual from which the sample is acquired (CTL = Control, SCZ = Schizophrenia, BPD = Bipolar Disorder, ASD = Autism Spectrum Disorder, AFF = Affective Disorder), the source tissue(s), and the downstream analyses conducted as a part of this manuscript.

Study	**Disease**	**Brain Tissue(s)**	**Assay**	**Analyses done**	**No. of Samples**
Roadmap	CTL	Dorsolateral Prefrontal Cortex	ChIP-seq: H3K27ac	Chromatin RCA	1
	CTL	Caudate nucleus, Cingulate gyrus, Hippocampus, Cortex	ChIP-seq: H3K27ac	Chromatin RCA	4
	CTL	Non-brain tissues: Adipose Tissue = 2, Adrenal Gland = 8, Adipose Tissue = 2, Blood = 12, Blood Vessel = 9, Bodily Fluid = 6, Bone Element = 9, Brain Cell = 9, Breast = 1, Connective Tissue = 12, Embryo = 22, Epithelium = 1, Esophagus = 5, Extraembryonic Component = 1, Gonad = 2, Heart = 9, Intestine = 6, Kidney = 3, Large Intestine = 15, Limb = 11, Liver = 9, Lung = 12, Lymph Node = 9, Mammary Gland = 9, Mouth = 3, Musculature of Body = 11, Pancreas = 11, Penis = 11, Placenta = 1, Prostate Gland = 25, Skin of Body = 15, Small Intestine = 3, Spinal Cord = 1, Spleen = 4, Stomach = 7, Thymus = 2, Thyroid Gland = 7, Urinary Bladder = 1, Uterus = 4, Vagina = 3, Vein = 3	ChIP-seq: H3K27ac	Chromatin RCA	294
ENCODE	CTL	Frontal Cortex	DNase-seq	TF imputation	2
GTEx	CTL	Frontal Cortex (BA9)	RNA-seq	QTL analyses, Gene Expression RCA	138
	CTL	Cerebellum	RNA-seq	Gene Expression RCA	298
	CTL	Amygdala = 99, Anterior Cingulate Cortex = 114, Caudate (basal ganglia) = 157, Cortex = 148, Hippocampus = 122, Hypothalamus = 121, Nucleus Accumbens (basal ganglia) = 144, Putamen (basal ganglia) = 118, Spinal cord (cervical c-1) = 87, Substantia Nigra = 86	RNA-seq	Gene Expression RCA	1196
	CTL	Frontal Cortex (BA9)	Genotypes	QTL analyses	25

	CTL	All non-brain tissues (GTEx V7)	RNA-seq	Weighted Gene Co-expression Analysis (WGCNA)	11688
	CTL	Non-brain tissues (GTEx V6p): Adipose - Visceral (Omentum) = 110, Esophagus - Gastroesophageal Junction = 166, Esophagus - Mucosa = 328, Esophagus - Muscularis = 282, Liver = 128, Lung = 350, Nerve - Tibial = 333, Pancreas = 194, Spleen = 120, Uterus = 39	RNA-seq	Gene Expression RCA	2050
Published Methylation data: Jaffe et al., 2016	CTL	Dorsolateral Prefrontal Cortex (BA46/9)	DNA Methylation Microarray studies	Methylation Analysis	255
PEC: BrainSpan	CTL	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	6
Published Single-cell: Lake et al., 2016	CTL	Dorsolateral Prefrontal Cortex (BA10)	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	575
	CTL	Temporal Cortex (BA21, BA22, BA41)	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	1771
	CTL	Intermediate Frontal Cortex (BA8)	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	490
	CTL	Primary Visual Cortex X1 (BA17)	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	391
Published Single-cell: Darmanis et al., 2015	CTL	Temporal Cortex	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	332
	CTL	Developmental Cortex	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	134
PEC: scRNA-seq	CTL	Dorsolateral Prefrontal Cortex	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	459
	CTL	Dorsal Pallium	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	473

PEC: Reference Brain	CTL	Dorsolateral Prefrontal Cortex	ChIP-seq: H3K27ac	Enhancer Definition	1
	CTL	Dorsolateral Prefrontal Cortex	HiC	Enhancer Definition	1
	CTL	Dorsolateral Prefrontal Cortex	ATAC-seq	Enhancer Definition	1
PEC: CommonMind	CTL	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	295
	SCZ	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	263
	BPD	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	47
	AFF	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	8
	CTL	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	285
	SCZ	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	263
	BPD	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	47
	AFF	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	8
PEC: CommonMind-HBCC	CTL	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	220
	SCZ	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	97
	BPD	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	70
	CTL	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	191
	SCZ	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	85
	BPD	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	25
PEC: BrainGVEX	CTL	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	259
	SCZ	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	95
	BPD	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	73

	CTL	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	47
	SCZ	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	45
	BPD	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	45
PEC: LIBD_szControl	CTL	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	320
	SCZ	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	175
	CTL	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	96
	SCZ	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	104
PEC: BipSeq	BPD	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	69
	BPD	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	55
PEC: UCLA-ASD	CTL	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	46
	ASD	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	43
	CTL	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	35
	ASD	Dorsolateral Prefrontal Cortex	Genotypes	eQTL	31
	CTL	Dorsolateral Prefrontal Cortex	ChIP-seq: H3K27ac	cQTL, Enhancer Definition	50
	ASD	Dorsolateral Prefrontal Cortex	ChIP-seq: H3K27ac	cQTL	31
	CTL	Cerebellar Cortex	ChIP-seq: H3K27ac	Enhancer Definition	50
	CTL	Temporal Cortex	ChIP-seq: H3K27ac	Enhancer Definition	50
PEC: Yale-ASD	CTL	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	23
	ASD	Dorsolateral Prefrontal Cortex	RNA-seq	eQTL	3
PEC: EpiDiff	CTL	NeuN+/- from Dorsolateral Prefrontal Cortex	ChIP-seq: H3K27ac	cQTL	117
	SCZ	NeuN+/- from Dorsolateral Prefrontal Cortex	ChIP-seq: H3K27ac	cQTL	109

S2.1.2 RNA-seq processing (Adapted from the Synapse Website)

The PsychENCODE RNA-seq pipeline (Fig. S2.2) is mostly based on that of ENCODE, which is compatible with stranded and unstranded mRNAs from (poly-A(+)), rRNA-depleted total RNA, or poly-A(-) RNA libraries. The inputs are RNA-seq reads (from paired-end stranded or single-end unstranded libraries), a reference genome and a gene annotation file (by default, GENCODE). We used GRCh37 (hg19) as a reference genome and Gencode v19 for gene annotation. Coding and non-coding transcripts were used to quantify gene expression. For each sample, the pipeline outputs included: A bam file with reads mapped to the genome, a bam file with reads mapped to the transcriptome, bigwig files with normalized RNA-seq signal track for unique and multi-mapping reads (split between +strand and -strand if the library was stranded), gene quantifications, and transcript quantifications.

The mapping of the reads was done using STAR (2.4.2a) and the quantification of genes and transcripts was done with RSEM (1.2.29). Although there is general agreement between the mappings and the gene quantifications produced by different RNA-seq pipelines, quantifications of individual transcript isoforms, being much more complex, can differ substantially depending on the processing pipeline employed, and are of unknown accuracy. Therefore, mapping and gene quantifications can be used confidently, whereas transcript quantifications should be used with care. Quality control metrics were calculated using RNA-SeQC (v1.1.8), featureCounts (v1.5.1), PicardTools (v1.128), and Samtools (v1.3.1). Pipeline source code can be found at doi:10.7303/syn12026837.1 at Synapse. All PsychENCODE sample FASTQ files were run through a unified RNA-seq processing pipeline (Fig. S1.1) run at the University of Chicago on an OpenStack cloud system. GTEx samples were processed at Yale University.

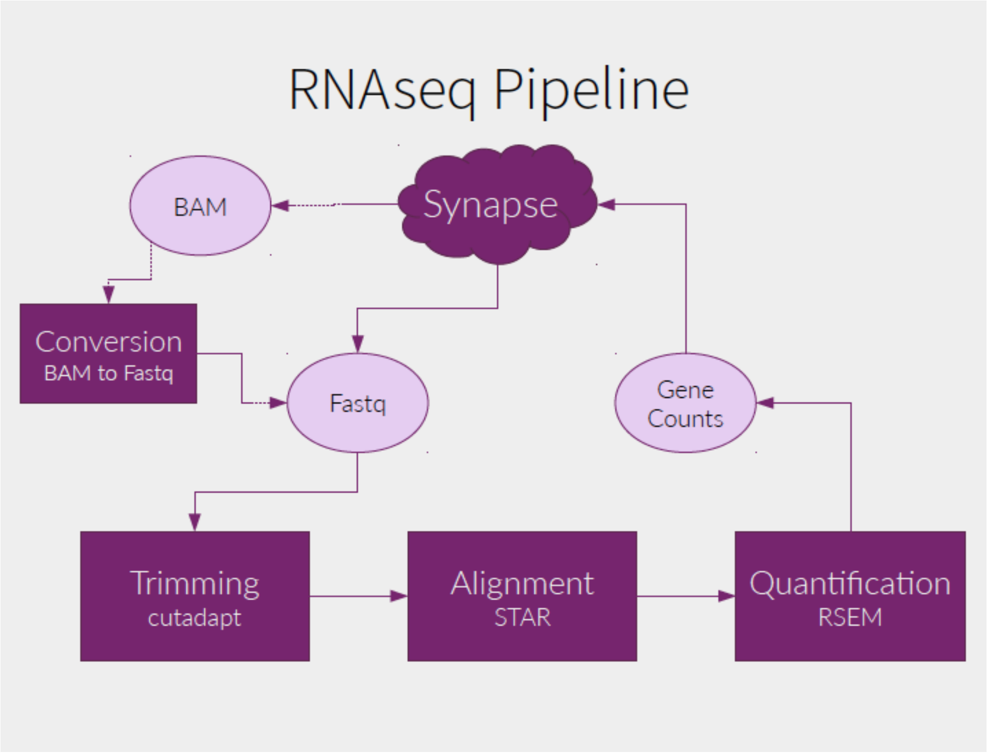


Fig. S2.2 PsychENCODE RNA-seq pipeline. The flowchart of the uniform RNA-seq pipeline is shown. This pipeline was modified based on the long-RNA-seq-pipeline used by the ENCODE Consortium.

S2.2 Single-cell RNA-seq analysis

S2.2.1 Datasets of single-cell transcriptomics

We integrated and used the same pipeline, including ENCODE RNA-seq analysis, to uniformly process single-cell RNA-seq data for ~900 cells from PsychENCODE with 11 novel cell types in embryonic and developmental tissues. The expression of ~3,000 neuronal cells with 8 excitatory and 8 inhibitory types (Lake et al., 2016), and ~400 cells including two developmental types, one adult neuronal type and 5 adult non-neuronal types, astrocytes, endothelial, microglia, oligodendrocytes, and oligodendrocyte progenitor cells (OPCs; Darmanis et al., 2015) were downloaded from corresponding publications. The details of cell types are shown in Table S2.2.

The basic cell types have been shared and used by other PsychENCODE capstone projects focusing on non-coding regulation and development. For PsychENCODE single-cell data, we first applied quality control on ~900 cells using the R 'scater' package (McCarthy et al., 2017) to filter the cells with low library size and high mitochondrial RNA concentration. Furthermore, the cells with a total library size less than 0.2 million were also filtered for future analysis. In total, we built a gene expression profile of ~800 high-quality cells quantified in TPM. We merged the PsychENCODE, Lake et al., and Quake et al. data by matching the gene names. As the single-cell data suffers from high dropout rates, we used MAGIC (van Dijk et al., 2017) to impute the missing values in the expression matrix. We compared these single cells based on (biomarker) gene expression similarity using tSNE, and found that cells of the same type generally could be clustered together (Fig. S2.3). In particular, 99.4% PsychENCODE cells clustered together with known developmental cell types from a previous report (Darmanis et al., 2015).

We also found that the gene expression changes across individual tissue samples could be largely explained by single-cell gene expression, and the changes of single-cell fractions were associated with the individual phenotypes. Therefore, we deconvolved the tissue-level gene expression data of all 1,866 individuals' tissue samples using single-cell gene expression data of 457 biomarker genes to find the fraction of different cell types that corresponded, and compared cell fractions across different phenotypes.

S2.2.2 Quantification of gene expression

The gene expression in both bulk and single-cell RNA-seq data were quantified in TPM and further transformed into log scale by log2(TPM+1). Later, we subjected the transformed gene expression to decomposition and devolution analysis (see below).

S2.3 Decomposition of brain tissue gene expression data

To check if the brain tissue expression was due to the combinations of single-cell types in Section 2.4 (i.e., the cell fractions), we decomposed the brain tissue gene expression data using an unsupervised approach to find the principal components of the tissue data, and compared them with single-cell expression data. Specifically, given the brain tissue gene expression matrix X (N by M) for a phenotype/disorder where M is the number of tissue samples and N is the number of select genes (e.g., the cell biomarker genes), we used non-negative matrix factorization (NMF) to decompose X into the product of two matrices, H and V so that $||X\text{-}V*H||^2$ was minimized and all elements of H is non-negative. H is a K by M matrix with the (i,j) element describing the contribution coefficient of the jth NMF "top-component" (NMF-TC) to the ith tissue sample, K is the number of select NMF-TCs (e.g., equal to the number of select cell types as above), and V is an N by K matrix with the (i,j) element being the expression level of the jth select gene on the ith NMF-TC.

We then correlated NMF-TCs with the select gene expression data of different single-cell types, and obtained a correlation map between NMF-TCs and single cells (Fig. 2B). For example, No. 10 and 19 NMF-TCs of the non-neuronal group highly correlated with astrocytes, No. 21 NMF-TC correlated with developmental cells, and No. 4, 7, 12, and 25 NMF-TCs of the neuronal group correlated with excitatory neuronal cell types. This suggests that a large portion of the tissue gene expression changes was a linear combination of these cell types' gene expression. Thus, we wanted to further identify the cell fractions showing how individual single cells contribute the tissue's gene expression, using a deconvolution. In addition, previous studies have identified cell type-specific expression patterns from co-expression analysis (Oldham et al., 2008). We found here that some of our NMF-TCs correlated with the eigengenes of gene co-expression modules (Gandal, M.J. et al., submitted), especially for the cell type modules, supporting again that they connect the cell type information from the bulk tissue data.

S2.4 Deconvoluting brain tissue gene expression data using single-cell data to estimate cell fractions

We used an unsupervised approach (NMF) to decompose tissue expression and found that NMF-PCs recovered the expression patterns of both neuronal and non-neuronal cells. This suggests that it is highly likely that a linear combination of single cells contributes to the brain tissue expression. Thus, to more accurately identify the single-cell fractions that determine the tissue expression, especially for various phenotypes/disorders, we further applied a supervised approach that used the single-cell expression data to deconvolve brain tissue expression data to find the fractions of different cell types of individual tissues.

In particular, we defined the brain tissue gene expression matrix B (N by M) for a phenotype/disorder, where M is the number of tissue samples and N is the number of select genes (e.g., the cell biomarker genes), and the single-cell gene expression matrix C (N by K), where K is the number of select cell types. We used the non-negative least square method to find a non-negative K by M matrix, with W to minimize $||B-C*W||^2$. The (i,j) element of W represents the linear combination coefficient of the ith single-cell type to the jth tissue expression, which is proportional to the jth single-cell fraction. In the deconvolution analysis, the gene expression quantified in TPM was transformed into log scale by log2(TPM+1).

We further evaluated the goodness-of-fit for the deconvolution model by calculating the coefficient of determination (also known as R^2), which accounts for the percentage of variance in the individual gene expression of tissue samples that has been explained by varying the cell proportions of cell types. Specifically, the variance in the gene expression of tissue samples was $||B||^2$ and the variance that had not been explained by the model was $||B-C*W||^2$. The R^2 could be calculated as $1-||B-C*W||^2/||B||^2$, which was further normalized to an adjusted R^2 by incorporating the degree of freedom. In addition, we deconvolved the tissue expression data and compared the cell fraction changes for various phenotypes and psychiatric disorders (Figs. S2.6 and S2.7). Fig. S2.8 shows the cell fractions across different ages. We found that Ex3 and Ex4 had a significant increasing trend across age (trend analysis $p<6.3e-10$ and 1.5e-6), but some non-neuronal types such as oligodendrocytes were found to decrease ($p<2.1e-14$). Furthermore, these age-related cell changes were potentially associated with differentially expressed genes across age groups; for example, a gene involved in early growth response was down-regulated in older age groups, whereas ceruloplasmin was down-regulated among middle-aged groups (Fig. 2F). In addition, we observed reduced microglia fractions for bipolar disorder and increased astrocyte fractions for SCZ.

We have validated our estimated cell fractions on a subset of samples from the EPIMAP study with experimentally measured NeuN+ fractions. Fig. S2.9 shows the NeuN+ fractions measured in experiments and estimated in our deconvolution analysis on 14 samples with RIN > 7.3. Our estimation was very close to the experimental NeuN+ fractions.

We further compared the performance of deconvolution with one popular deconvolution tool CIBERSORT (Newman et al., 2015). We performed CIBERSORT to deconvolve the tissue expression data with single-cell data of selected 24 types and further calculated the variance as an adjusted R-square; this value (0.8132) was lower than that calculated by our deconvolution method (0.8779).

Data files associated with both the decomposition (NMF components and fractions) and deconvolution (cell fractions) analyses are available on the website (adult.psychencode.org).

Table S2.2 Summary of cell types. This table includes PsychENCODE developmental cell types and public adult cell types from Lake et al. 2016 and Darmanis et al. 2015.

Abbreviation	Adult/Developmental	Full name	Source
Ex	Adult	Excitatory Neuron	Lake et al. 2016
In	Adult	Inhibitory Neuron	Lake et al. 2016
OPC	Developmental	Oligodendrocyte progenitor cells	Li, M. et al. (*submitted*)
Trans	Developmental	Transient cell type (nascent neurons)	Li, M. et al. (*submitted*)
NEP	Developmental	Neuroepithelial cells	Li, M. et al. (*submitted*)
IPC	Developmental	Intermediate progenitor cells	Li, M. et al. (*submitted*)
Quiescent/Quies	Developmental	Quiescent newly born neurons	Darmanis et al., 2015
Replicating/Repli	Developmental	Replicating neuronal progenitors	Darmanis et al., 2015
IntN	Developmental	Inhibitory Neuron	Li, M. et al. (*submitted*)
ExtN	Developmental	Excitatory Neuron	Li, M. et al. (*submitted*)
Oligo	Developmental	Oligodendrocyte cells	Li, M. et al. (*submitted*)
Astrocytes/Astro	Developmental	Astrocytes	Li, M. et al. (*submitted*)
Pericytes/Peri	Developmental	Pericytes	Li, M. et al. (*submitted*)
Endothelial/Endo	Developmental	Endothelial cells	Li, M. et al. (*submitted*)
Microglia/Micro	Developmental	Microglia	Li, M. et al. (*submitted*)
Microglia/Micro	Adult	Microglia	Darmanis et al., 2015
OPC	Adult	Oligodendrocyte progenitor cells	Darmanis et al., 2015
Endothelial/Endo	Adult	Endothelial cells	Darmanis et al., 2015
Astrocytes/Astro	Adult	Astrocytes	Darmanis et al., 2015
Oligo	Adult	Oligodendrocyte	Darmanis et al., 2015
OtherNeuron	Adult	Mixed of excitatory and inhibitory neuronal cells	Darmanis et al., 2015

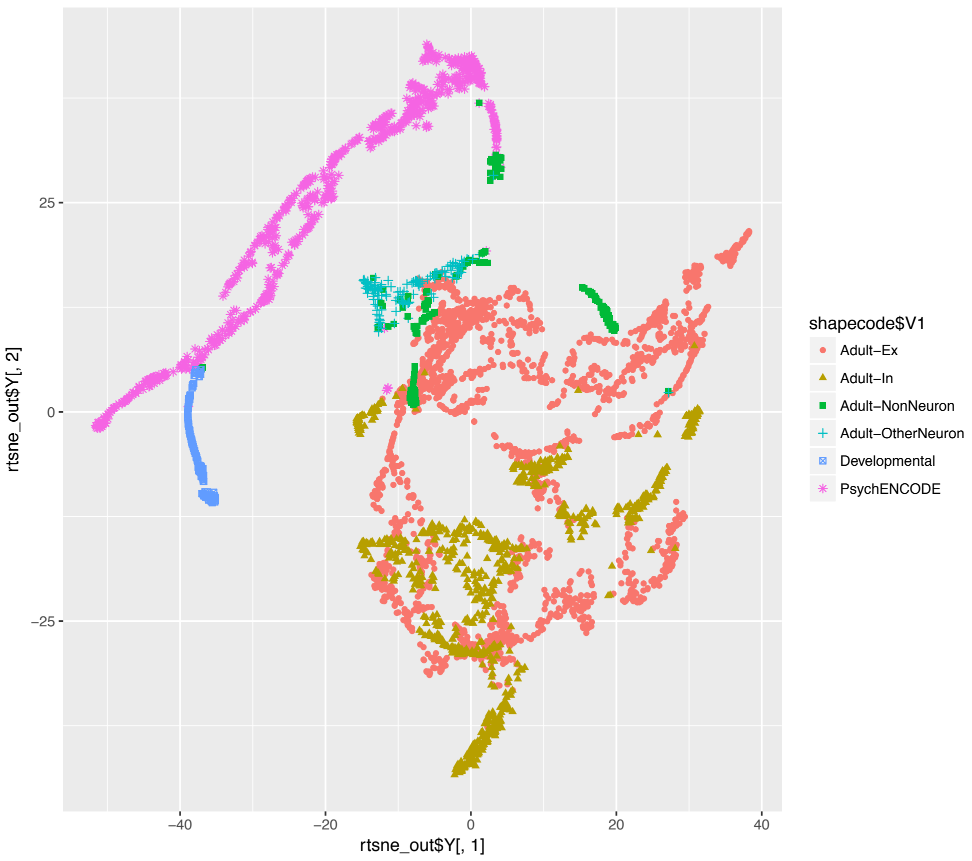


Fig. S2.3 t-SNE plot of the PsychENCODE and public single-cell data. Most of the PsychENCODE data were found to be clustered together with public developmental data in Darmanis et al., 2015.

S2.5 Differentially expressed genes for brain phenotypes

We used the limma R package for linear modeling to find genes that are differentially expressed for neuropsychiatric disorders, sex, and brain regions. Normalized gene expression data was partitioned into the control and schizophrenia samples or male and female samples using a merged matrix. We then constructed a design matrix representing these partitions, which we used to fit a linear model and estimate fold changes/standard errors. We then applied empirical Bayes smoothing to the standard errors. The output was represented in a table form or as a heatmap using the heatmap.2 R package. This pipeline was used for brain region analysis using gene expression data from GTEx, where either brain regions (amygdala, anterior cingulate cortex, caudate, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens, putamen, spinal cord, and substantia nigra) or all brain samples were compared with select control tissues (liver, colon, lung, esophagus, pancreas, spleen, and stomach) for region-specific or brain-specific differential gene expression, respectively. In addition, the differentially expressed and spliced genes and transcripts for psychiatric disorders were identified by a submitted report (Gandal, M.J. et al., submitted). Associated data files with the differentially expressed (DEX) and spliced genes and transcripts from the both the current manuscript and the submitted report are available on the website (adult.psychencode.org).

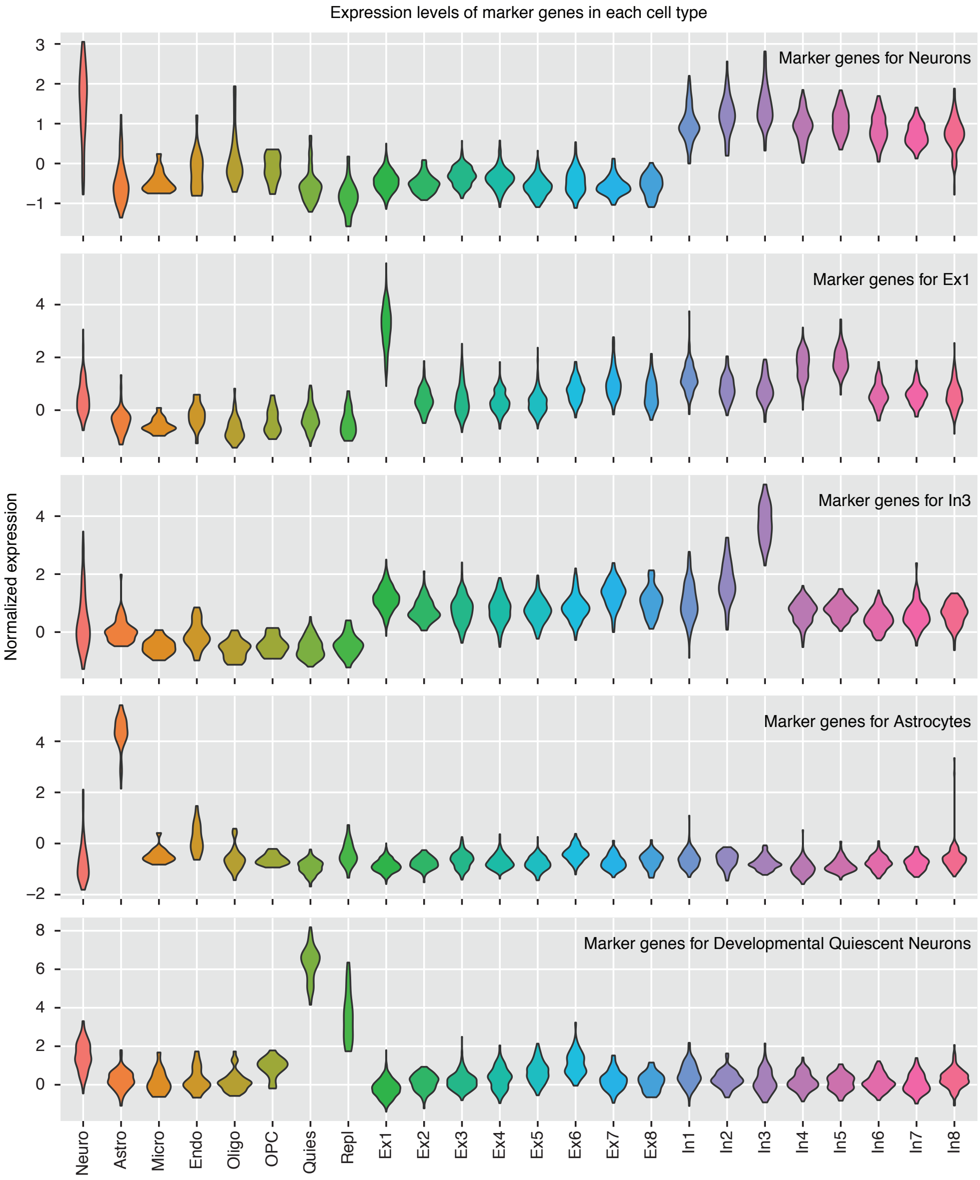


Fig. S2.4 Biomarkers show higher expression in the cell type from which they were defined compared to other cell types. Expression signatures of biomarkers are conserved in the newly constructed expression matrix, which integrates multiple sources of single-cell expression data.

S2.6 Gene co-expression network analysis

We used WGCNA to identify modules of co-expressed genes, both within and between tissues (Zhang et al., 2005). Briefly, each gene was associated with a vector of normalized expression values across either individuals or tissues (using median expression). A weighted network was constructed where the weight between any two genes had a similar score, calculated by normalizing the Pearson correlation of their expression vectors to lie between 0 and 1, and raising this to the power β.

We followed Zhang et al. in setting β such that connectivity of the network was as close to scale-free as possible (using the R^2 statistic described in Zhang et. al.). The genes were then hierarchically clustered using a *topological overlap score*, which compares how similar the patterns of connection are from each node to all other nodes. Disjoint modules were extracted using the Dynamic Tree Cut algorithm (Langfelder et al., 2008). We further extracted submodules in addition to the disjoint modules extracted by WGCNA, by adding the subtrees formed on each merge where both left and right subtrees were larger than a minimal size (which we set at 30 genes). To find brain specific modules/submodules using clusters calculated on median expression variation across tissues, we further calculated the *module eigengenes* (as described in Zhang et al.), and calculated the correlation of each eigengene with a binary vector, which was 1 for brain regions and 0 otherwise. We called a module 'brain specific' if this correlation was significant at the 0.001 level (under a permutation test of the tissue labels).

Our co-expression analysis indeed found several modules with eigengenes showing very different expression levels between brain and non-brain samples (Fig. S2.5), which suggests that brain-specific regulatory mechanisms drive these brain co-expression modules (Gandal, M.J. et al., submitted). Associated data files with the gene and isoform co-expression modules from the both the current manuscript and the submitted report are available on the website (adult.psychencode.org).

S2.7 Gene expression and DNA methylation over aging

To find the effect of age on gene expression, we selected genes that showed significant correlation with age. Samples were segregated by age bins of 20 years, for a total of five bins (0-20, 20-40, 40-60, 60-80, and 80-100). Gene expression was estimated using uniform processing with the PsychENCODE RNA-seq pipeline (See S2.1.3). Fig. S2.10 displays 90 protein-coding and non-coding genes that correlate with age. In particular, EGR1 (early growth response - ENSG00000120738.7) and CP (ceruloplasmin - ENSG00000047457.9) are displayed. Similarly, we processed array methylation data to investigate the effect of aging in promoter and enhancer methylation. Published data from (Jaffe et al., 2016) were used. We used the normalized (scaled) proportion of methylated CpGs across individuals' age bins near gene TSS (Fig. S2.11).

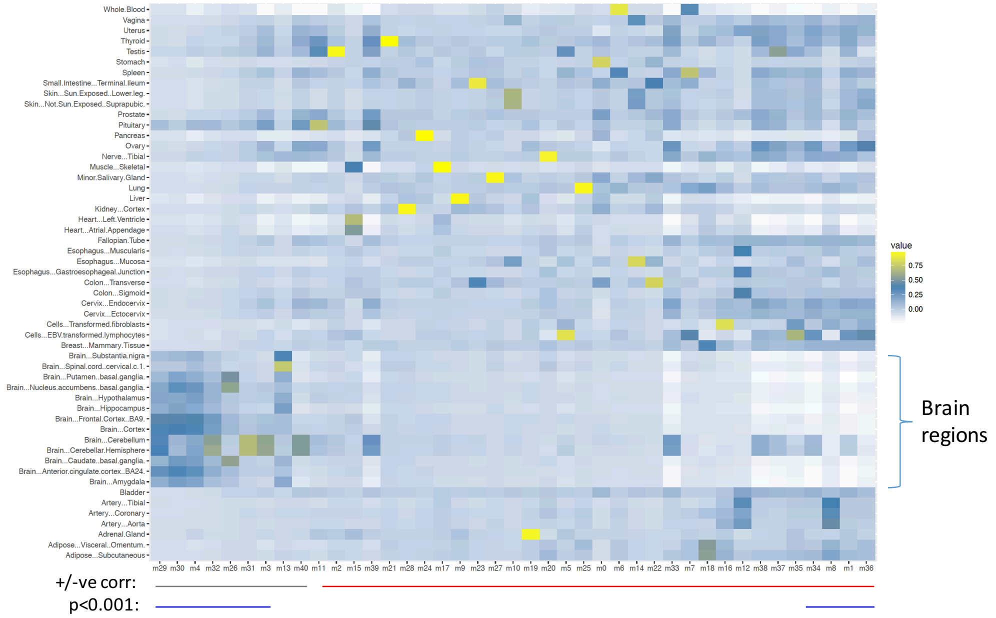


Fig. S2.5 Brain-specific co-expression modules and submodules. Module eigengenes are plotted as columns, which are ordered by the degree to which their expression is specific to the brain (see text). Lines beneath the plot show positive (green) and negative (red) correlations, with correlations that are significant at the p<0.001 level (either positive or negative) highlighted in blue.

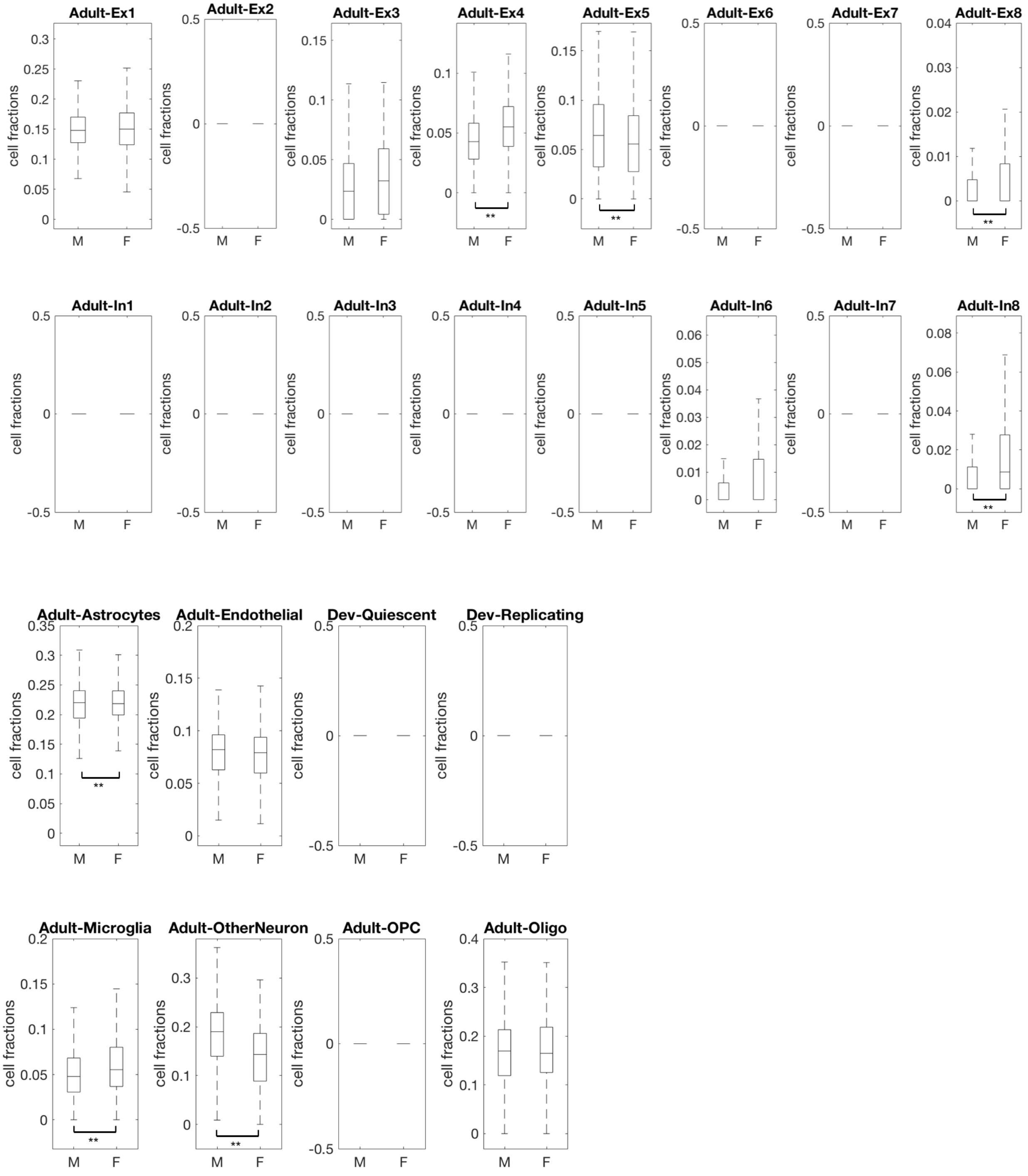


Fig. S2.6 Estimated cell fractions of 24 selected cell types in control samples. The cell types with significant changes (FDR < 0.05) between genders after balancing age distribution are labeled with double asterisks (**)..

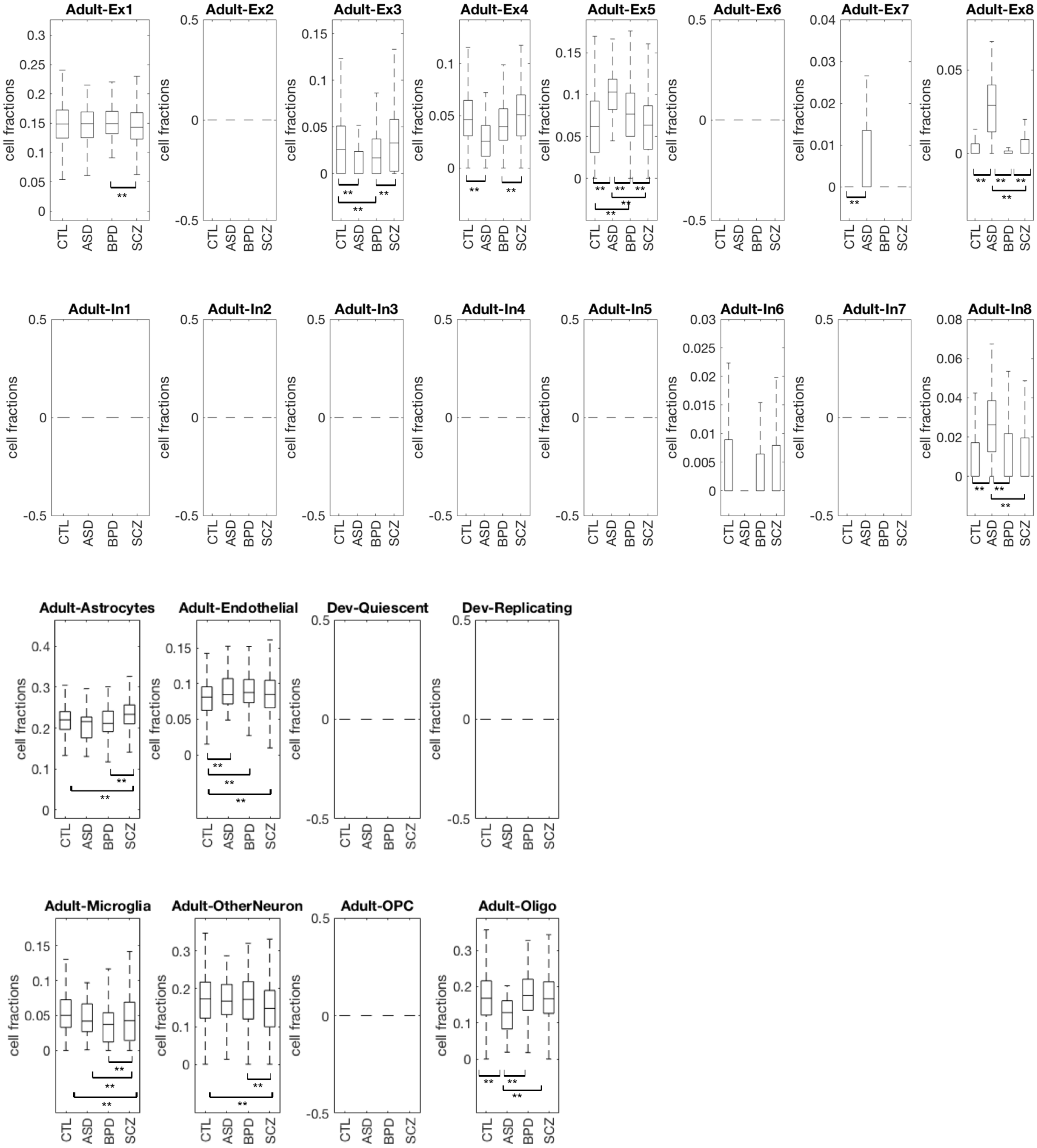


Fig. S2.7 Estimated cell fractions of 24 selected cell types in samples with different disorders. For each cell type, the boxes with double asterisks (**) indicate the disorder types that show significant differences (FDR < 0.05) after balancing the age distribution.

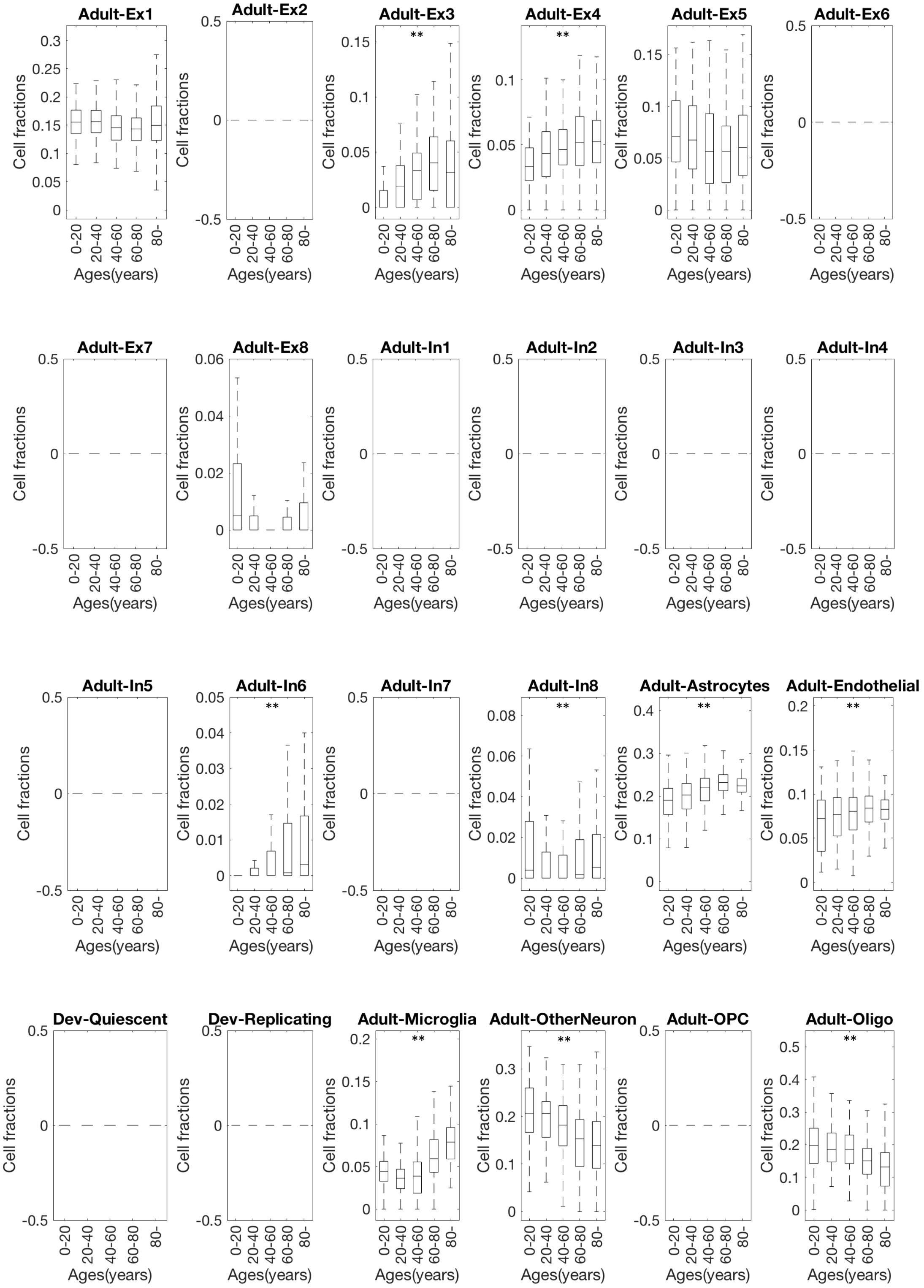


Fig. S2.8 Estimated cell fractions of 24 selected cell types in control samples with different ages. The cell types showing significant increasing/decreasing trends across ages (trend analysis p-value < 1e-2) are labeled with double asterisks (**).

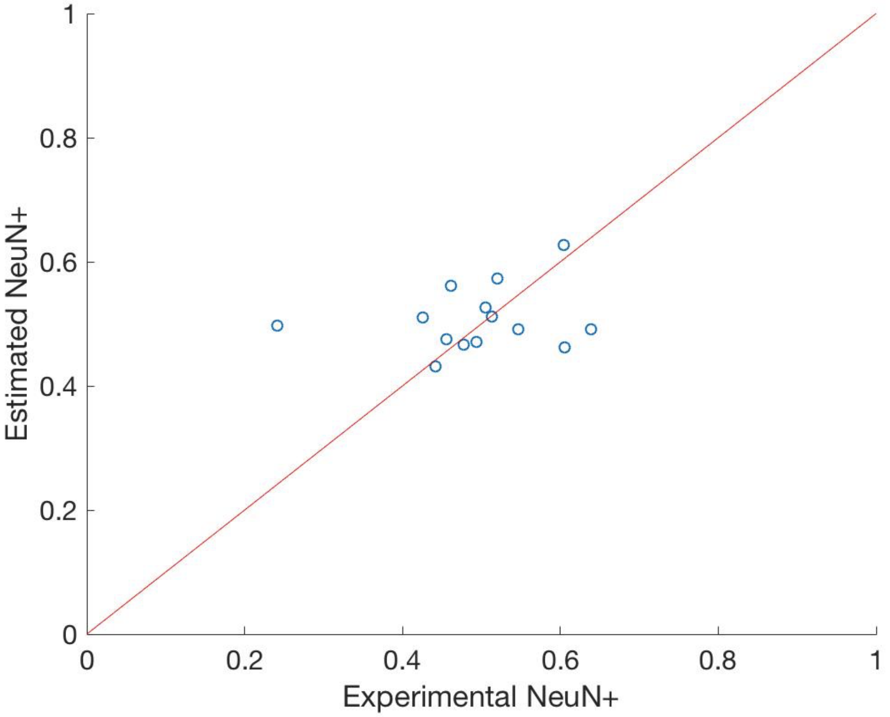


Fig. S2.9 Validation of estimated cell fractions from deconvolution. The X-axis shows the NeuN+ cell fractions measured in experiments and the y-axis shows the NeuN+ cell fractions estimated from deconvolution. The median error is 0.04.

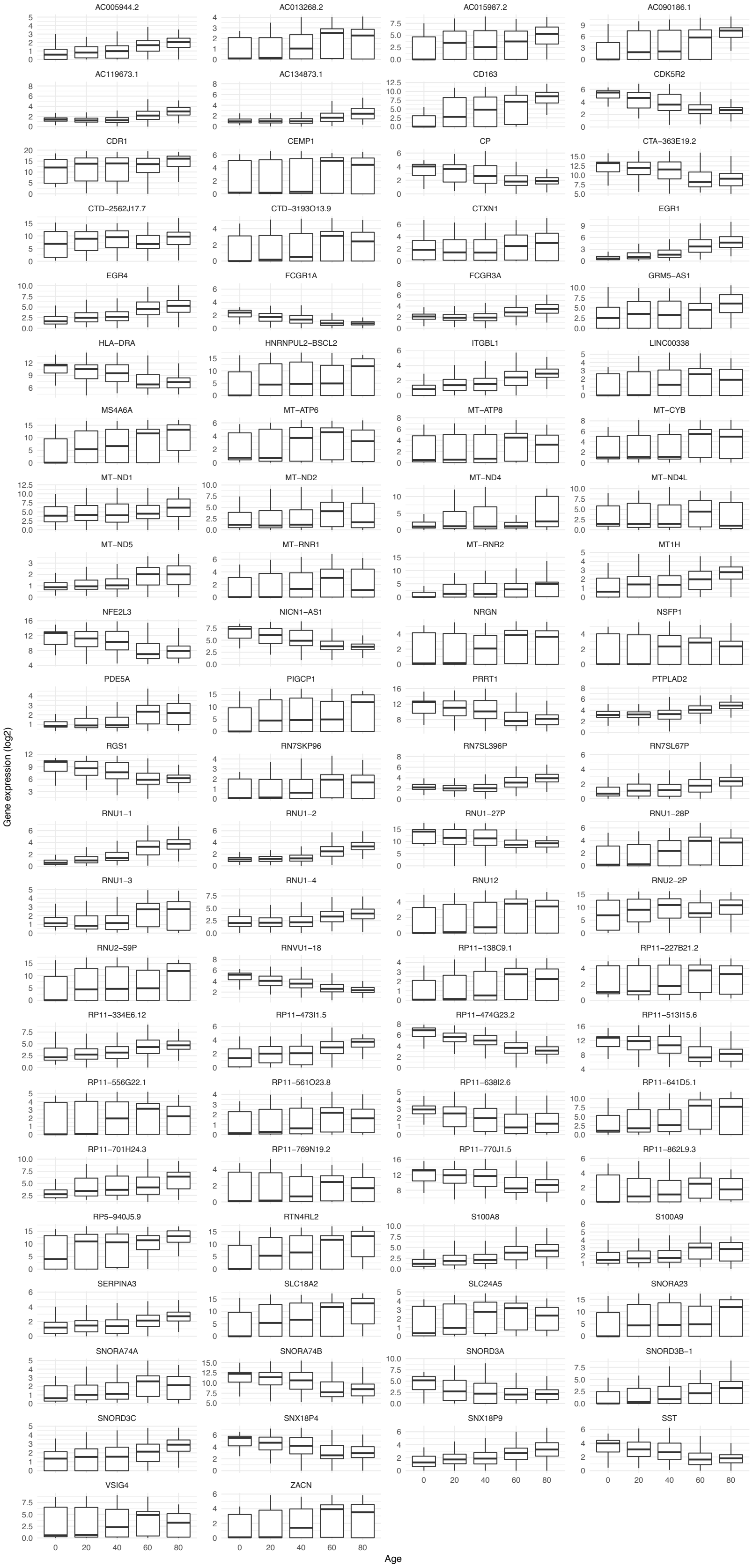


Fig. S2.10 Gene expression variation in the human brain across ages. The X axis shows 5 bins of age and the Y axis shows the log2(rpkm) for genes positively or negatively correlated with age. Each panel refers to a gene, where the identification was made by ENSEMBL ID.

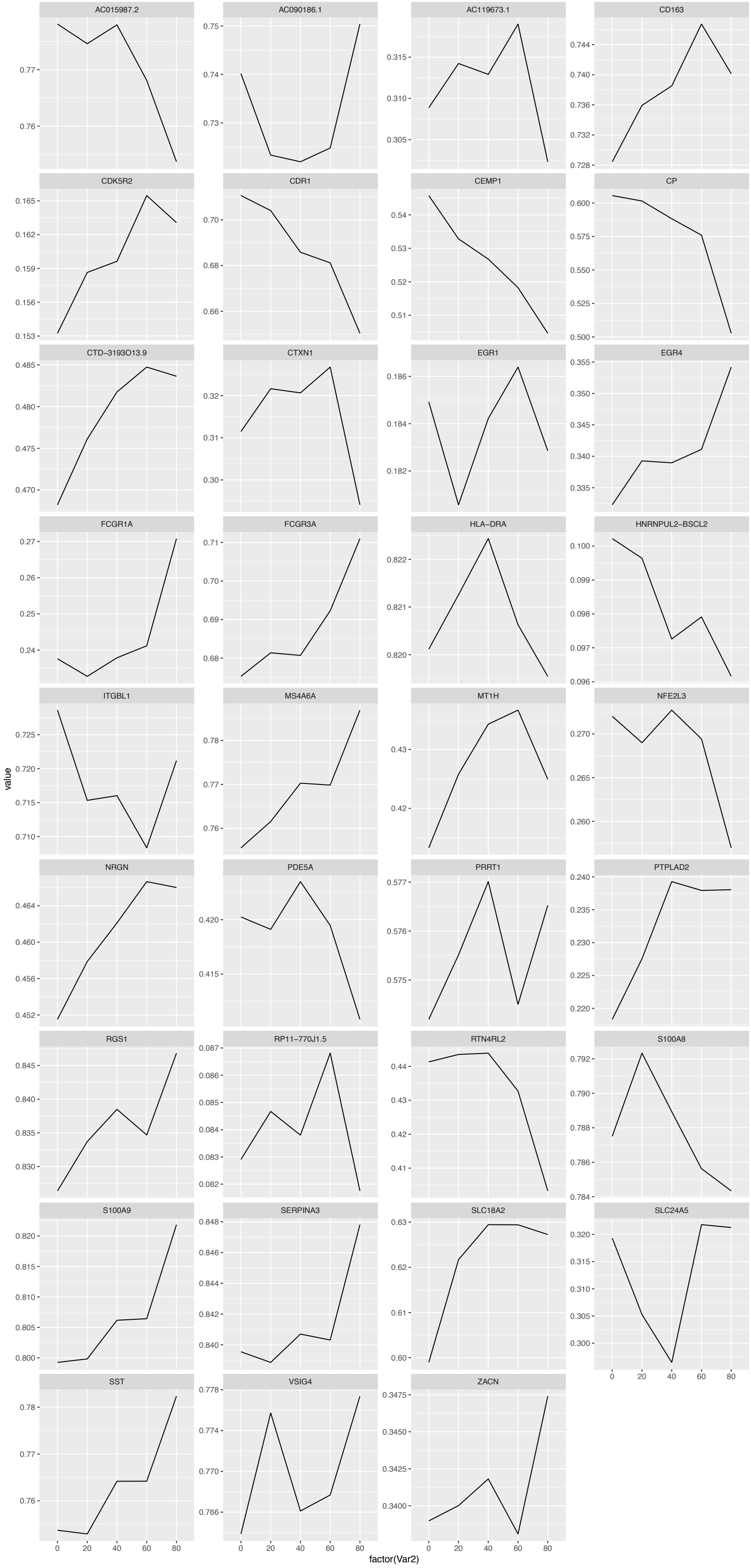


Fig. S2.11 Promoter and enhancer region methylation of genes correlated to aging. Genes with methylation data available were assessed for their methylation status. The X axis shows 5 bins of age and the Y axis shows the normalized distribution of methylation near the gene TSS. Each panel refers to a gene, where the identification was made by gene name.

S3. Supp. content to main text section "Enhancers"

S3.1 PsychENCODE ChIP-seq pipeline and processing

We used the modified parallel version of the ENCODE ChIP-seq pipeline (Fig. S3.1). This was improved over the ENCODE pipeline using the workflow system Snakemake for more efficient computation (https://github.com/weng-lab/psychip_snakemake). The original ENCODE pipeline can be found at https://goo.gl/KqHjKH. The PsychENCODE ChIP-seq data were processed at the University of Massachusetts and Yale University.

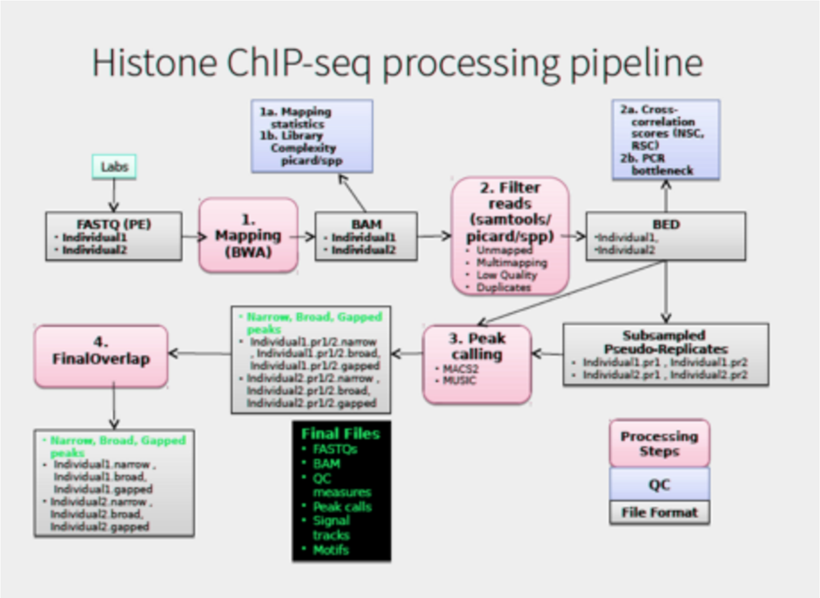

Histone ChIP-seq processing pipeline
Labs
1a. Mapping statistics
1b. Library Complexity picard/spp
2a. Cross-correlation scores (NSC, RSC)
2b. PCR bottleneck
FASTQ (PE)
• Individual1
• Individual2
1. Mapping (BWA)
BAM
• Individual1
• Individual2
2. Filter reads (samtools/ picard/spp)
• Unmapped
• Multimapping
• Low Quality
• Duplicates
BED
•Individual1,
•Individual2
4. FinalOverlap
• Narrow, Broad, Gapped peaks
• Individual1.pr1/2.narrow , Individual1.pr1/2.broad, Individual1.pr1/2.gapped
• Individual2.pr1/2.narrow , Individual2.pr1/2.broad, Individual2.pr1/2.gapped
3. Peak calling
• MACS2
• MUSIC
Subsampled Pseudo-Replicates
• Individual1.pr1 , Individual1.pr2
• Individual2.pr1 , Individual2.pr2
• Narrow, Broad, Gapped peaks
• Individual1.narrow , Individual1.broad, Individual1.gapped
• Individual2.narrow , Individual2.broad, Individual2.gapped
Final Files
• FASTQs
• BAM
• QC measures
• Peak calls
• Signal tracks
• Motifs
Processing Steps
QC
File Format

Fig. S3.1 PsychENCODE ChIP-seq processing pipeline. This pipeline flowchart was adapted and modified from the ENCODE ChIP-seq pipeline (https://goo.gl/KqHjKH). FASTQ files were aligned using BWA and the reads were filtered to get only unique mapped reads for peak calling using MACS2. Pseudo-replicates were generated before peak calling for each individual to find robust peaks. NSC, RSC, and PCR bottlenecks were generated for QC.

S3.2 Epigenomics Roadmap, ENCODE ChIP-seq for identifying regulatory regions

We incorporated ChIP-seq datasets from the Roadmap Epigenomics Consortium and the ENCODE project in our analysis. To integrate them consistently with the PsychENCODE dataset, ChIP-seq experiments were uniformly processed using the ENCODE standard pipeline (See below, Section S2.3), including alignment, quality control, and peak-calling. Each released experiment consists of the raw sequencing data (in FASTQ) and the processed output, including alignment, signal, and peak files.

With the set of uniformly processed ChIP-seq experiments, a comprehensive statistical model was used to generate a registry of candidate regulatory elements (cREs) for major cell lines and tissues (Moore et al., under review). The cREs are based on a combined set of high-quality DHSs. For a particular cell or tissue, z-scores for DNase, H3K4me3, H3K27ac, and CTCF were calculated for these high-quality DHSs. Using the maximum z-score across all cell types and the distance to the nearest TSS, the cREs were classified into promoter-like elements, enhancer-like elements, and regions bound by CTCF only. As described in later sections (S2.6 and S2.10), we used the epigenetics signals of these cREs to annotate enhancers, calculate cQTLs, and perform comparative chromatin signal RCA analysis.

S3.3 Activated brain enhancers

To annotate a set of active enhancers, we uniformly processed the H3K27ac, H3K4me1, and H3K4me3 ChIP-seq data from the reference brain using the standard ENCODE pipeline. We also processed the ATAC-seq data generated on the same reference. Supplemented by the DNase-seq and ChIP-seq data of the prefrontal cortex (PFC) from the ENCODE and Roadmap Epigenomics projects, we identified 79,056 active enhancers. An active enhancer was considered to be in open chromatin regions (ATAC-seq signal or DNase signal Z-score > 1.64), with H3K27ac and H3K4me1 signals (Z-score > 1.64), which are characteristic markers for enhancers. To exclude promoters, we excluded regions with enriched H3K4me3 signals. These identified enhancer regions largely overlapped with ChromHMM enhancer annotations of the PFC (>90%).

We uniformly processed 150 H3K27ac ChIP-seq data from healthy individuals, 50 each from PFC, PC, and CB regions. For each sample, we called H3K27ac peaks using the standard ENCODE ChIP-seq pipeline. The H3K27ac peaks were pooled across the cohort, generating a total of 37,761 H3K27ac pooled peaks in PFC, 42,683 in TC, and 26,631 in CB. Each pooled peak was present in more than half of the samples in its corresponding brain region. Note that although the numbers of aggregate peaks were smaller than the number of reference enhancers, they actually covered a larger fraction of the genome, as the average width of H3K27ac peaks was larger than that of reference enhancers.

To investigate the enhancer activity across the population, we intersected the set of active enhancers identified in the reference sample with the H3K27ac PFC ChIP-seq peaks in each individual from the cohort. Any H3K27ac peaks intersecting with the reference enhancers were considered to be active enhancers in the corresponding individual. Among the 50 healthy samples, a median of 53,976 (~70%) enhancers from the reference brain were found to be active in the cohort. We also examined the cumulative number of reference enhancers that could be found in the cohort with individuals sorted by the number of overlapping enhancers, as shown in Fig. S3.2. The cumulative number grew fast at the beginning, and saturated at the 20th person of the sorted cohort. Thus, we hypothesize that pooling together the active enhancers of 20 people should recover most of the potential regulatory elements in brain prefrontal cortex. The PsychENCODE enhancer list is available on the website (adult.psychencode.org).

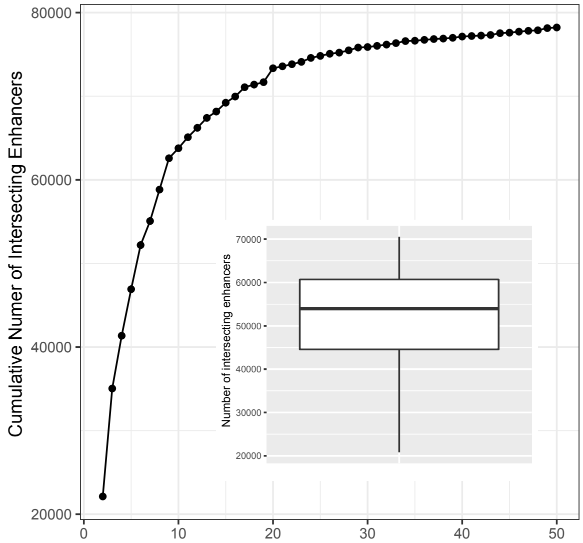


Fig. S3.2 Active reference brain enhancers in the population. The dotted line shows the cumulative number of identified reference sample enhancers in the cohort, which saturates at the 20th individual from the sorted cohort. The boxplot shows the number of identified reference enhancers found active in each individual, with the lower and the upper boundaries of the box showing the first and the third quantiles.

S4. Supp. content to main text section "Consistent comparison"

S4.1 Spectral analytic approaches (PCA, tSNE, RCA) to compare transcriptomic and epigenomic data across brain and other tissues

One key aspect of our analysis is that we, as consistently as possible, processed the transcriptomic and epigenomic data from PEC, GTEx (GTEx Consortium, 2017), and the Epigenetic Roadmap (Kundaje et al., 2015). This approach allowed us to compare the brain to other organs in a consistent fashion to assess if the human brain has unique gene expression and chromatin activities. This comparison could not be achieved without such a large-scale uniform data processing. We attempted several methods for an appropriate comparison; in particular, we used methods to reduce the dimensionality of genes or enhancers to compare the underlying structure of brain and other tissues. PCA and t-SNE are two popular techniques, but PCA tends to capture global structures, ignoring most of the local structure, but be overly influenced by data outliers (Johnstone et al., 2009). In contrast, t-SNE tends to separate samples from the same tissue so that the cluster distances on the t-SNE space are not proportional to real gene expression dissimilarities, and thus does not give a sense of overall effects (Maaten et al., 2008). As an alternative, we found another very useful technique to be reference component analysis (RCA), which projects the gene expression in an individual sample against a reference panel, and then essentially reduces dimensionality of individual projections (Li et al., 2017). Moreover, as shown in Fig. 3E, all the brain tissue samples from the different projects tend to group together, which is a consequence of our uniform processing.

In order to perform an RCA analysis, we first built a reference gene expression panel based on GTEx, which consisted of the average expression of genes across a panel of tissues. To select the genes in this panel, we searched for expression outliers (i.e., genes for which at least one sample had a delta log10(rpkm) higher than 1). This yielded 4,162 coding and non-coding genes in the reference panel. The average expression level for these genes was extracted from the GTEx v6 average expression file. We next used the gene expression from uniformly processed PsychENCODE and GTEx samples and selected only the 4,162 genes in the reference panel. We then calculated the correlation between each sample x reference tissue pair and built a correlation matrix.

Finally, to extract structures from the dataset, we performed PCA on the correlation matrix. Median sample was defined as the median PC1 and PC2. In order to account for sample variance within tissues, we fit the PC1 and PC2 to a multivariable Gaussian distribution and plotted the ellipse defined by median PC1, PC2, with width and height equal to one standard deviation in PC1 and PC2 space, respectively. We calculated the distances between tissues and samples. Overall, the distance of the brain centroid to other tissues was approximately one order of magnitude higher than the distance between brain samples. Distance was calculated using Euclidean distance on RCA space (Fig. S4.1, Table S4.1, and S4.2).

In order to assess which genes were responsible for differences in RCA PC1, we simulated RNA-seq samples with a step function equal to discrete changes in gene expression. For each step, we selected the gene representing the biggest change in the PC1 dimension. We simulated 5,253 steps (Fig. S4.2; the path is represented by the dark like moving from the brain to other tissues). In total 1,226 gene were selected multiple times as the biggest change in the PC1 dimension. Selecting top-ranked genes and performing Reactome term enrichment analysis with Panther resulted in enrichment for brain pathways.

Similar to the transcriptome RCA analysis, we built a reference panel using H3K27ac signals overlapping CREs as previously defined. For reference tissues, we used uniformly processed Epigenome RoadMap samples and calculated the average H3K27ac signals across CREs. We further filtered outlier CREs to select informative CREs. Similar to the transcriptome analysis, we selected CREs with average signal across the CRE higher than 0.1 from 40 tissues. That filter yielded 5,506 reference CREs. We calculated the correlation between each sample and the reference tissue pair, built a correlation matrix, and performed PCA analysis at the correlation space. Median and ellipses were calculated as described above. To remove batch effects from H3K27Ac, we used well-established methods. First, we computed the PCA in the RCA space and selected the first principal component; indeed, most of the variance in the first component was derived from experimental differences. In order to consistently compare the transcriptome and epigenome, we selected tissues on roadmap that were also represented in the transcriptome RCA analysis. Namely, we used roadmap_brain, esophagus, liver, lung, pancreas, spleen, and uterus. We also performed a PCA analysis for these samples (Fig. S4.3).

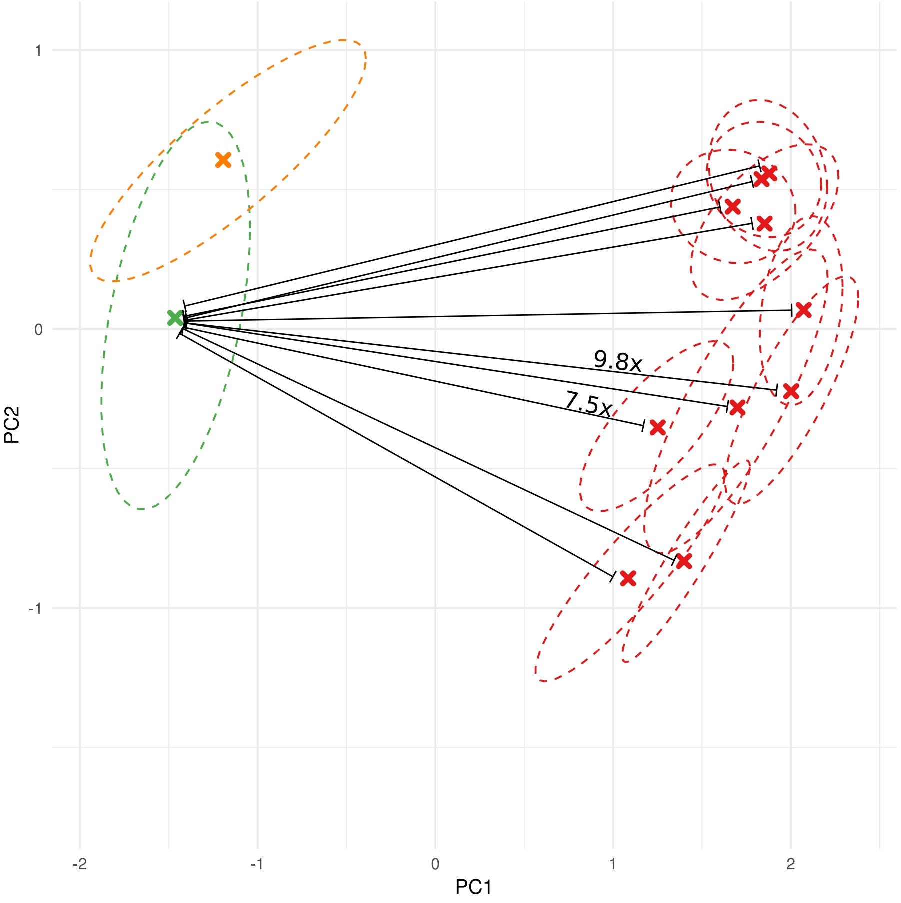


Fig. S4.1 Distances between brain to other tissues centroids. Transcriptome RCA plot as in Fig. 3. Brain samples' distributions are displayed in green and orange. Other tissues are shown in red. Euclidean distance was calculated between all centroids (Table S4.1) and normalized by the median brain distance (Table S4.2).

	PEC	Adipose	Esophagus	Liver	Lung	Nerve	Pancreas	Spleen	Uterus
PEC	0.00								
Adipose	4.32	0.00							
Esophagus	4.01	0.56	0.00						
Liver	3.32	1.69	1.85	0.00					
Lung	4.25	0.37	0.84	1.39	0.00				
Nerve	3.86	0.67	0.15	1.78	0.90	0.00			
Pancreas	3.35	1.13	1.17	0.69	0.93	1.10	0.00		
Spleen	3.66	1.38	1.63	0.39	1.05	1.59	0.61	0.00	
Uterus	4.07	0.46	0.10	1.82	0.76	0.23	1.16	1.58	0.00

Table S4.1 RCA centroid distances. Euclidean distance was calculated between all tissue centroids in RCA space.

	PEC	Adipose	Esophagus	Liver	Lung	Nerve	Pancreas	Spleen	Uterus
PEC	0.00								
Adipose	9.80	0.00							
Esophagus	9.11	1.27	0.00						
Liver	7.54	3.83	4.19	0.00					
Lung	9.64	0.83	1.91	3.16	0.00				
Nerve	8.76	1.51	0.35	4.05	2.05	0.00			
Pancreas	7.61	2.56	2.66	1.57	2.12	2.49	0.00		
Spleen	8.31	3.12	3.69	0.89	2.38	3.61	1.39	0.00	
Uterus	9.24	1.05	0.22	4.13	1.72	0.53	2.63	3.59	0.00

Table S4.2 RCA centroid distances normalized by median interbrain distance. Euclidean distance was calculated between all tissue centroids in RCA space and normalized by median brain distance.

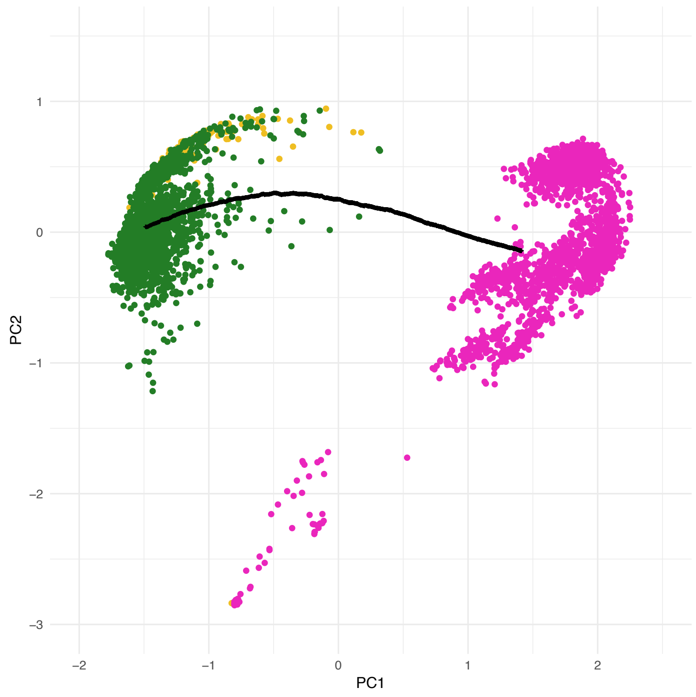


Fig. S4.2 Assessment of the most impactful genes in the PC1 dimension. All analyzed RNA-seq samples are displayed. Green and yellow samples are brain samples and pink samples were extracted from other tissues. Dark line represents hypothetical samples with gene expression changes.

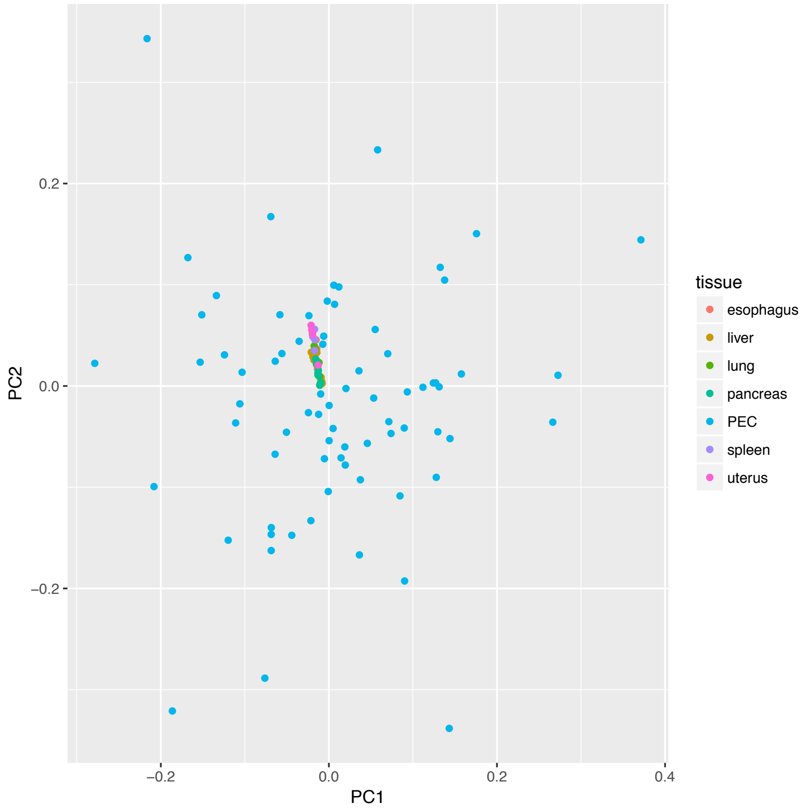


Fig. S4.3 PCA plot for regulatory data. H3K27Ac signals were used to calculate the PCA after batch correction. Brain samples are scattered in both PC1 and PC2, whereas the roadmap samples are clustered together.

Reactome pathways	#	#	expected	Fold Enrichment	+/-	P value
Serotonin Neurotransmitter Release Cycle	17	4	.16	25.01	+	4.14E-02
Neurotransmitter Release Cycle	50	6	.47	12.75	+	1.68E-02
Neuronal System	337	14	3.17	4.41	+	8.43E-03
Dopamine Neurotransmitter Release Cycle	22	5	.21	24.15	+	4.52E-03

Table S4.3 Reactome pathway enrichment for most impactful genes in the RCA PC1 dimension. Pathway enrichment for the top genes selected in the Fig. S4.2 analysis.

S4.2 Non-coding RNAs and TARs

We used uniformly processed RNA-seq signal data from healthy individuals from GTEx 6p and PsychENCODE to quantify the expression activity of annotated and non-annotated regions of the human genome. In order to create signal files, we used alignment files (bam files) as input to RSEM to create both uniquely aligned and multiple aligned signal tracks. Signal values were normalized within samples using the total number of reads mapped to the genome and by generating RPM values. We divided the genome into bins of 100 base pairs and calculated the average expression (RPM) in windows. We finally selected regions in the genome with an RPM higher than 0.1 to filter transcriptionally active regions. The union of all bins in the human genome above the threshold was used to build a resource of active regions of the human brain. To estimate the proportion of coding and non-coding (i.e., non-coding and unannotated) regions, we overlapped active regions to the GENCODE v19 annotation. For each annotation class, we estimated the cumulative proportion of coding and non-coding regions (Fig. S4.4).

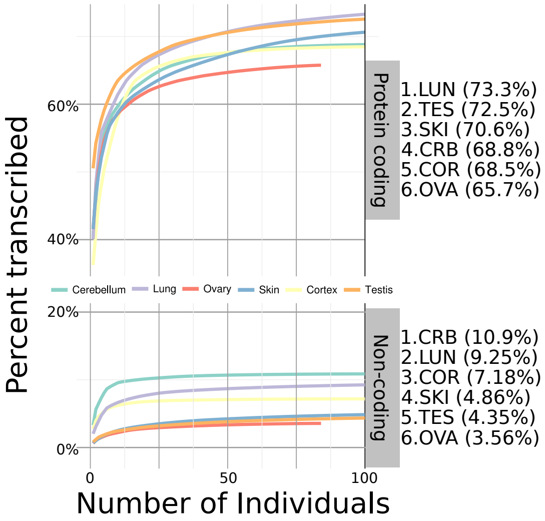


Fig. S4.4 Cumulative distribution of transcribed regions in the human brain and other tissues. The Y axis shows the cumulative transcribed proportion of annotated and unannotated regions (coding or non-coding). The X axis shows the number of transcriptomes (or individuals) analyzed. Labels on the right-hand side of the figure display the maximum cumulative proportion found.

We fit the curves on Fig. S4.4 to cumulative exponential curves to estimate a per tissue upper bound of the proportion of coding and non-coding transcribed regions. We observed that most tissues were transcriptomically saturated at approximately 100 individuals. Moreover, although a large (65-75%) of the coding transcriptome was active, only (3-10%) of the non-coding transcriptome was active. By contrast, the absolute number of nucleotides active in non-coding regions (which include non-annotated regions) was much larger than in coding regions. In Fig. 3, we estimated inter-tissue variability by calculating the cumulative transcriptome diversity as stated above; inter-sample diversity was defined as the average diversity across samples in a tissue-based fashion. Values displayed in Fig. 3 were normalized by average diversity in coding and non-coding regions, respectively. The inter-sample variability was estimated by calculating the mean difference. Absolute values for coding and non-coding transcriptome diversity were also estimated.

S5. Supp. content to main text section "QTL analysis"

S5.1 Genotype data processing

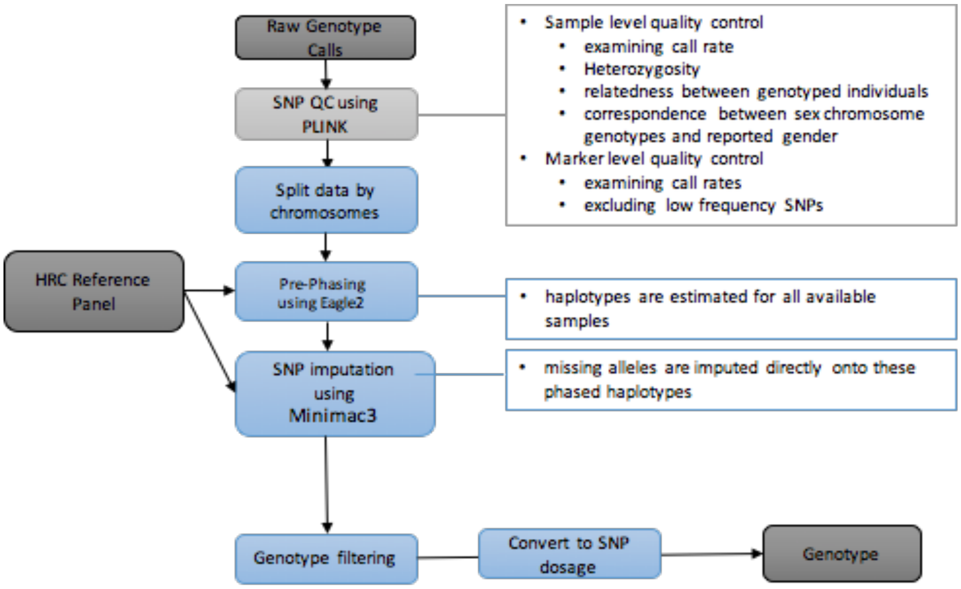

Raw Genotype Calls
SNP QC using PLINK
Split data by chromosomes
HRC Reference Panel
Pre-Phasing using Eagle2
SNP imputation using Minimac3
Genotype filtering
Convert to SNP dosage
Genotype
• Sample level quality control
• examining call rate
• Heterozygosity
• relatedness between genotyped individuals
• correspondence between sex chromosome genotypes and reported gender
• Marker level quality control
• examining call rates
• excluding low frequency SNPs
• haplotypes are estimated for all available samples
• missing alleles are imputed directly onto these phased haplotypes

Fig. S5.1 PsychENCODE genotype data processing pipeline. The raw genotype data were called and converted to PLINK files. We ran an initial quality sample level and marker level using PLINK. The quality controlled genotype data were then prepared by prephasing using Eagle2. The prephased data were imputed using Minimac3 and HRC. After imputation, we filtered genotype using R2>0.3 to get high-quality imputation data.

S5.1.1 Genotyping arrays, data generation, and quality control

Genotyping was done on several different genotyping platforms listed in Supplemental Table S5.1 and Section S9. Initial QC was performed using PLINK (Purcell et al., 2007) to remove markers with: zero alternate alleles, genotyping call rate < 0.95, Hardy-Weinberg p-value < 1 x 10-6, and individuals with genotyping call rate < 0.95. We also corrected for the strand flipping problem using snpflip (https://github.com/biocore-ntnu/snpflip).

S5.1.2 Imputation of genotypes

Genotypes of all studies were imputed using a uniform genotype QC and imputation pipeline in order to streamline quality control and genotype imputation of genome-wide single nucleotide polymorphism (SNP) data. This imputation pipeline consisted of four primary, independent modules: (1) pre-imputation data processing and quality control; (2) PCA of raw genotype data; (3) genotype imputation of untyped variants; and (4) post-imputation statistical analysis. Briefly, in the pre-imputation step, input genotype data (PLINK binary format) was reformatted for downstream analysis, and initial summaries of classic technical parameters (e.g., minor allele frequency, per individual and per site missing rates, case/control missingness, Hardy-Weinberg equilibrium) were produced.

Datasets	**#samples**	**DataPlatform**
BipSeq	179	Illumina_1M and Illumina_h650
LIBD_szControl	493	Illumina_1M, Illumina_Omni5, Illumina_h650
CMC-HBCC	696 (896 totoal)	Illumina_1M, Illumina_Omni5, Illumina_h650
BrainSpan	41	HumanOmni2.5
CommonMind	620	IlluminaInfiniumHuman Omni Express Exome 8 v 1.1b chip
GTEx	450 (97 DFC)	Illumina OMNI 5M or 2.5M
BrainGVEX	138+280	Affymetrix6.0, PsychChips
UCLA-ASD	97	Omni-2.5 and Omni-2.5-Exome
iPSC	3	WGS
EpiGABA	9	Illumina_HumanOmni1-Quadv1.0

Table S5.1. Summary of genotype data generated in PsychENCODE and used in our paper. Most of these studies used different genotyping platforms. There were overlap of the individuals in BipSeq, LIBD_szControl and CMC_HBCC studies and the number of total individuals of these three studies are 896.

The second module consisted of genotype PCA using peddy (Pedersen et al., 2017) to identify ancestry structure (Fig. S5.2). In the third, prior to imputation, SNP positions, identifiers, and alleles were aligned to the relevant reference genome assembly using LiftOver, and genotype data was divided into chromosomes and overlapping segments for parallel haplotype pre-phasing and imputation using eagle2 and Minimac3 on the Michigan Imputation Server (Das et al., 2016). We used the recently released HRC Reference Panel for imputation. In the final module, we used the summary of R2 from Minimac3 to evaluate the imputation accuracy and only kept imputed SNPs with R2>0.3 for QTL analysis.

S5.2 eQTL and isoform QTL

We used a conservative approach for eQTL and isoQTL processing. We adhered closely to the GTEx pipeline, and we benchmarked our results with direct comparisons to available data files in the GTEx portal (gtexportal.org) and published GTEx results. We used the QTLtools software package for eQTL and isoform QTL (iso-QTL) identification. Following the normalization scheme used by GTEx, the gene expression matrix was normalized using quantile normalization, followed by inverse quantile normalization to map to a standard normal distribution. Probabilistic estimation of expression residuals (PEER) factors, genotype PCs, gender, and respective study were used as covariates in our calculations to identify cis-eQTL. For cis-eQTLs, we calculated the associations between gene expression and variants within a 1Mb window of each gene TSS. These calculations were performed using genotype and gene expression data from 1,387 individuals (associations between a total of 43,854 genes and 5,312,508 variants were evaluated for potential QTLs).

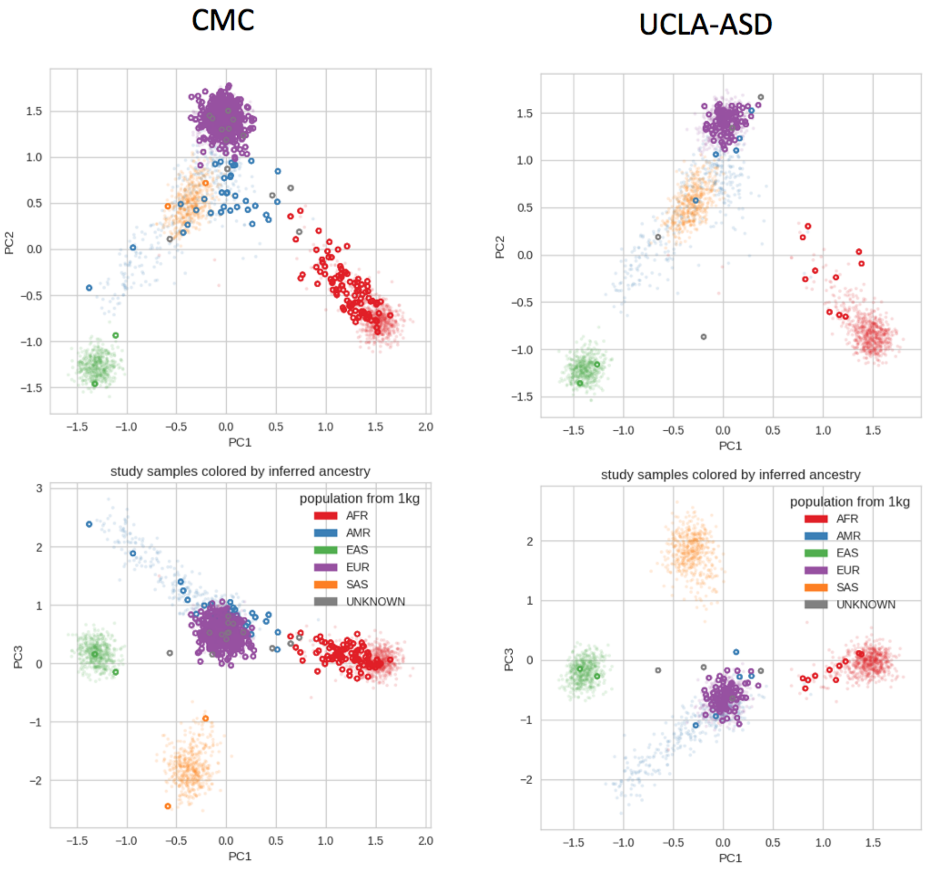


Fig. S5.2 Genotype PCs showing the population structure in CMC and UCLA-ASD studies. The first 3 genotype PCs could capture most of the population structures. The top panels show genotype PC1 vs. PC2. The bottom panels show genotype PC1 vs. PC3. A majority of the individuals in these two studies were from EUR populations.

We performed multiple testing correction on nominal P values by limiting FDR values less than 0.05. We identified 2,542,908 significant cis-eQTLs. Because of linkage disequilibrium (LD), many of the eQTL SNPs for the same gene were correlated. We pruned such SNPs for a given gene by restricting the genotype correlation coefficient (r^2) values to exceed 0.5. Enforcing this resulted in 373,686 eQTLs.

These conservative approaches for searching for eQTLs identified a substantially larger number of cis-eQTLs and eGenes than previous brain eQTL studies. This may reflect the greater statistical power offered by our large sample size. We also identified 157,592 iso-QTLs, using a similar pipeline to that in our search for eQTLs. For 1,147 individuals, we used isoform percentages of 43,820 transcripts using the same set of variants that we used in our search for eQTLs.

S5.3 cQTLs

To calculate cQTLs, we used the uniformly processed ChIP-seq data from PsychENCODE (3 different brain regions) and Roadmap ChIP-seq data for different tissues. cQTLs were calculated using candidate regulatory regions (cREs). We extended (in rare cases truncated) each cRE to 1kb (a typical enhancer's size). We calculated the average signal on each of the extended regions across PsychENCODE and roadmap samples. We identified 74 individuals from UCLA_ASD and 218 from Epidiff correlating this signal matrix with nearby variants within 1Mb window of the peak center. Then, we used the QTLtools for cQTL calculation using FDR<0.05 and identified the most significant SNP for each enhancer.

S5.4 Cell fraction & residual QTL

We used the QTLtools package (Delaneau et al., 2017) to calculate the cell fraction and residual QTLs based on the cell fractions and estimated residuals. QTLtools was run in nominal pass mode to identify fQTLs. We used gender and disease as covariates. To best deal with population structure as potential confounding factor, we restricted our analysis to European adult samples, which comprise a substantial subset of all available genotyped data (Fig. S5.3A).

We take the conservative approach of defining significant fQTLs to be those associated with Bonferroni-corrected p-values of no more than 0.05. By using this approach, we identified 9 different cell types with significant fQTLs (Ex3, Ex4, Ex5, In6, In8, Astrocytes, Microglia, and Endothelial cells). Specifically, these 9 cell types are those which exhibit fQTLs when using gender and disease status as input covariates. We find that different cell types exhibit considerable heterogeneity in terms of their abundance within the set of high-confidence fQTLs (Fig. S5.3B). The SNVs associated with these fQTLs coincide with 106 distinct SNVs associated with cis-eQTLs. A supplementary data file listing all fQTLs (along with associated data) is available online.

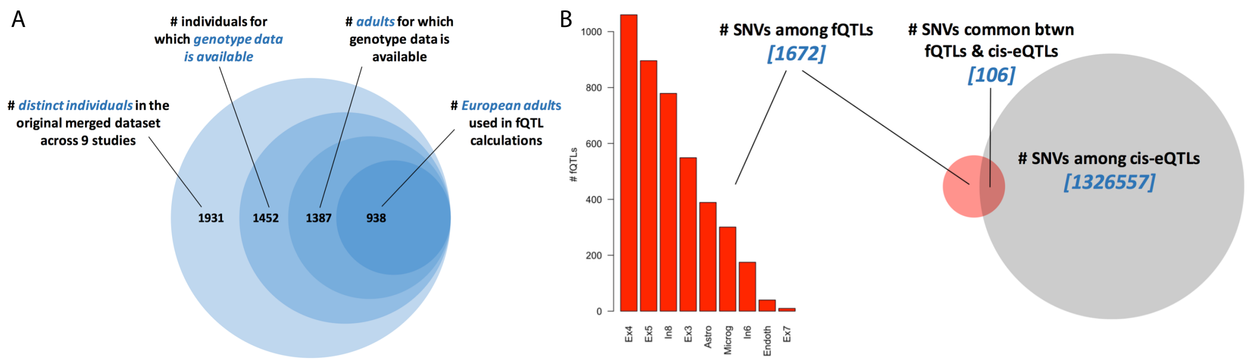


Fig. S5.3 Datasets, counts, and cis-eQTL overlaps associated with fQTLs. A. In calculating fQTLs, we restricted our analyses to a 938 European adult samples for which genotype data is available. **B.** The histogram on the left represents the counts for the number of fQTLs across 10 different cell types. These fQTLs encompass 1672 distinct SNVs, of which 106 (6.3%) also appear among the cis-eQLTs.

S5.5 QTL replication and sharing

We evaluated the replication of GTEx and CommonMind PFC eQTLs in our study using the π_1 statistic (Storey et al., 2003; Ng et al., 2017), which estimated the proportion of eQTLs that were significant based on the p-value distribution in our dataset. In this calculation, we used top SNPs from our eQTLs and found overlap with the eQTL SNPs in GTEx and CommonMind. Then, we used the p values of associations between these overlapped SNPs with protein-coding genes in the 1Kb window to calculate π_1. We determined π_1 values of 0.93 and 0.9 for GTEx and CommonMind, respectively, which indicated a good replication rate. We also used the π_1 statistic to investigate the sharing of SNPs between different types of QTLs in our study. In this case, we found shared SNPs between eQTL top SNPs and other QTL SNPs. Then, the π_1 statistic was calculated based on the p values of the associations of these shared SNPs with all genes in the 1Kb window. We found that the π_1 value of cQTL was 0.89, which is the highest among all QTL SNP sharing comparisons.

Lists of the identified QTLs are available on the website (adult.psychencode.org).

S6. Supp. content to main text section "Regulatory networks"

S6.1 Generation of Hi-C libraries

Hi-C libraries were generated as previously described (Won et al., 2016). Briefly, adult dorsolateral prefrontal cortices (DLPFC) from three individuals (sample information provided below) were acquired through a Reference Brain Project as a component of the psychENCODE project. Frozen pulverized tissue (100mg) was homogenized in 2mL of ice-cold lysis buffer (10mM Tris-HCl pH8.0, 10mM NaCl, 0.2% NP40, protease inhibitor). Ten million nuclei were collected, and chromatin was crosslinked in 1% formaldehyde (diluted in 1X PBS) for 10 min. Crosslinked chromatin was first digested by HindIII (NEB, R0104), and digested sites were labelled by biotin-14-dCTP (ThermoFisher, 19518-018). Proximity-based ligation was performed within nuclei to prevent random collision-based ligation (Rao et al., 2014). Biotin-marked DNA was then purified and sequenced by Illumina 50 bp paired-end sequencing.

S6.2 Hi-C data processing

Hi-C reads were mapped and filtered as previously described (Won et al., 2016) using hiclib (https://bitbucket.org/mirnylab/hiclib). Only cis reads (which refer to intra-chromosomal interactions) were used to construct contact matrices at 40kb and 10kb resolution for compartment and loop analyses, respectively. To obtain maximum resolution for loop detection (10kb), we pooled datasets from three individuals (see below for read depths for pooled samples). To compare interaction profiles in adult and fetal brain, we combined previously generated Hi-C datasets from two fetal cortical laminae to obtain comparable read depths (Won et al., 2016; see below for read depths for pooled samples).

Compartments were analyzed by calculating the leading principal component (PC1) values from Pearson's correlation matrix generated from contact matrices in 40kb resolution. Regions with PC1s positively and negatively correlated with the gene density were defined as compartment A and B, respectively. TADs were called based on contact matrices in 40kb resolution using Hi-C domain callers (http://chromosome.sdsc.edu/mouse/hi-c/download.html). Briefly, the directionality index was calculated by measuring the degree of interaction bias of a given 40kb bin to its upstream (2Mb) and downstream (2Mb) regions, which was subsequently processed by a hidden Markov model.

Table S6.1 Summary of Hi-C datasets

Samples	Sample information	cis filtered reads	total filtered reads
HBS189	Male 36yr (Ancestry unknown)	197,394,146	251,515,059
HBS106	Male 64yr (Ancestry unknown)	170,057,582	209,571,512
HBS181	Male 44yr (Caucasian)	243,396,052	299,801,452
Pooled samples Adult brain		610,847,780	760,888,023
Pooled samples Fetal brain	Won et al., 2016	855,987,816	1,834,759,860

S6.3 Detection of promoter-based interactions

Promoter-based interactions were identified as previously described (Won et al., 2016). Briefly, we constructed background interaction profiles from randomly selected length- and GC content-matched regions to promoters (defined as 2kb upstream of transcription start sites based on Gencode v19). Using these background interaction profiles, we fit interaction frequencies into Weibull distribution at each distance for each chromosome using the *fitdistrplus* package in R. Significance of interaction from each promoter was calculated as the probability of observing higher interaction frequencies under the fitted Weibull distribution, and interactions with FDR<0.01 (which corresponds to P-values~$1x10^{-4}$) were selected as significant promoter-based interactions. In total, we detected 149,098 promoter-based interactions. We overlapped promoter-based interactions with genomic coordinates of TADs, and found that the majority (~75%) of promoter-based interactions were located within the same TADs.

We used a binomial test as previously described (McLean et al., 2010) to evaluate the epigenetic state enrichment of regions that interact with promoters, using a 15 state chromatin state model in adult prefrontal cortices (PFC) from Roadmap Epigenomics (Kundaje et al, 2015). To assess whether promoter-interacting regions are enriched in enhancer states, we calculated the significance of the overlaps by binomial probability of P=Pbinom(k>=s, n=n, p=p), when p = fraction of genome in enhancer states, n = the number of promoter-interacting regions, s = the number of promoter-interacting regions that overlap with enhancer states.

To assess whether epigenetic states affect their target gene expression levels, we used transcriptomic profiles of PFC from neurotypical individuals (see section S2.1). Quantile normalized expression values were log transformed and centered to the mean expression level for each sample using a *scale(center=T, scale=F)+1* function in R. The centered expression values denote each gene's relative expression level in a given individual, and were used throughout the integrative analysis. We selected genes that interact with enhancers (EnhG=Genic enhancers, Enh=Enhancers), promoters (TssA=Active transcription start sites, TssAFlnk=Active transcription start site flanking regions), bivalent enhancers (EnhBiv), and repressive states (Het=Heterochromatin, ReprPC=Polycomb repressive sites) and average centered expression values for each group were calculated and plotted.

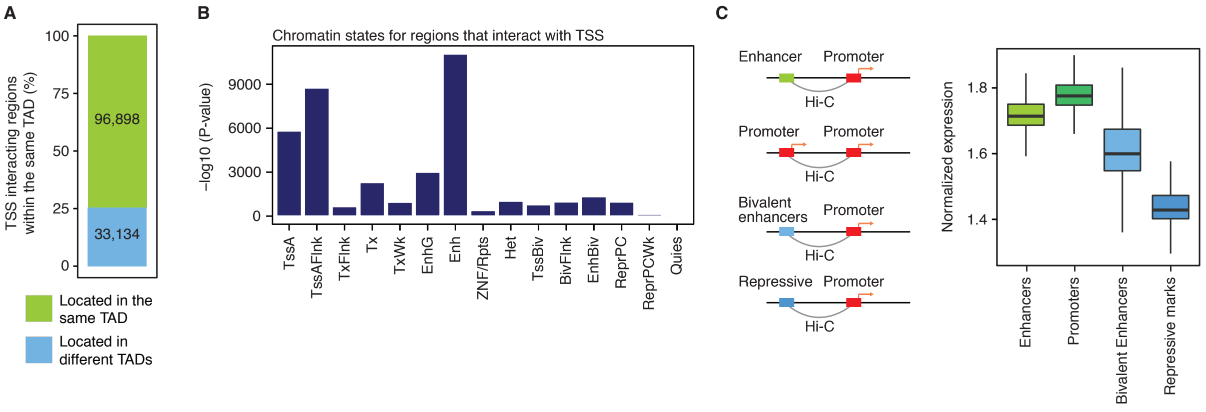


Fig. S6.1 Regulatory relationships in the adult cortex. A. The majority of promoter-based interactions reside within the same topologically associating domains (TADs). **B.** Regions that interact with transcription start sites (TSS) are enriched with other TSS and enhancers. **D.** Distribution of the number of putative enhancers assigned to each promoter. **E.** Genes that interact with enhancers or promoters are more highly expressed than genes that interact with bivalent enhancers or repressive marks.

S6.4 Integrative analysis

Compartment changes across brain development. Genomic regions were classified into (1) regions that undergo compartment A to B switching from fetal to adult brain, (2) regions that undergo compartment A to B switching from adult to fetal brain, (3) regions that do not switch their compartments across brain development (stable).

Genes were then grouped according to the compartment categories they locate in, and centered expression values for each group were calculated. As our RNA-seq data mainly focus on adult brain transcriptome, we processed expression values from Kang et al. to generate centered expression values (Kang et al., 2011). Prenatal and postnatal centered expression values were plotted for each group of genes. We also overlapped chromatin states in adult PFC and fetal brain defined by chromHMM with compartment categories. We then counted the total number of each chromatin state in a given compartment category, which was subsequently normalized by the size and the number of total chromatin states in that compartment category. We compared these normalized counts for each chromatin state between fetal and adult brains using the Fisher's exact test.

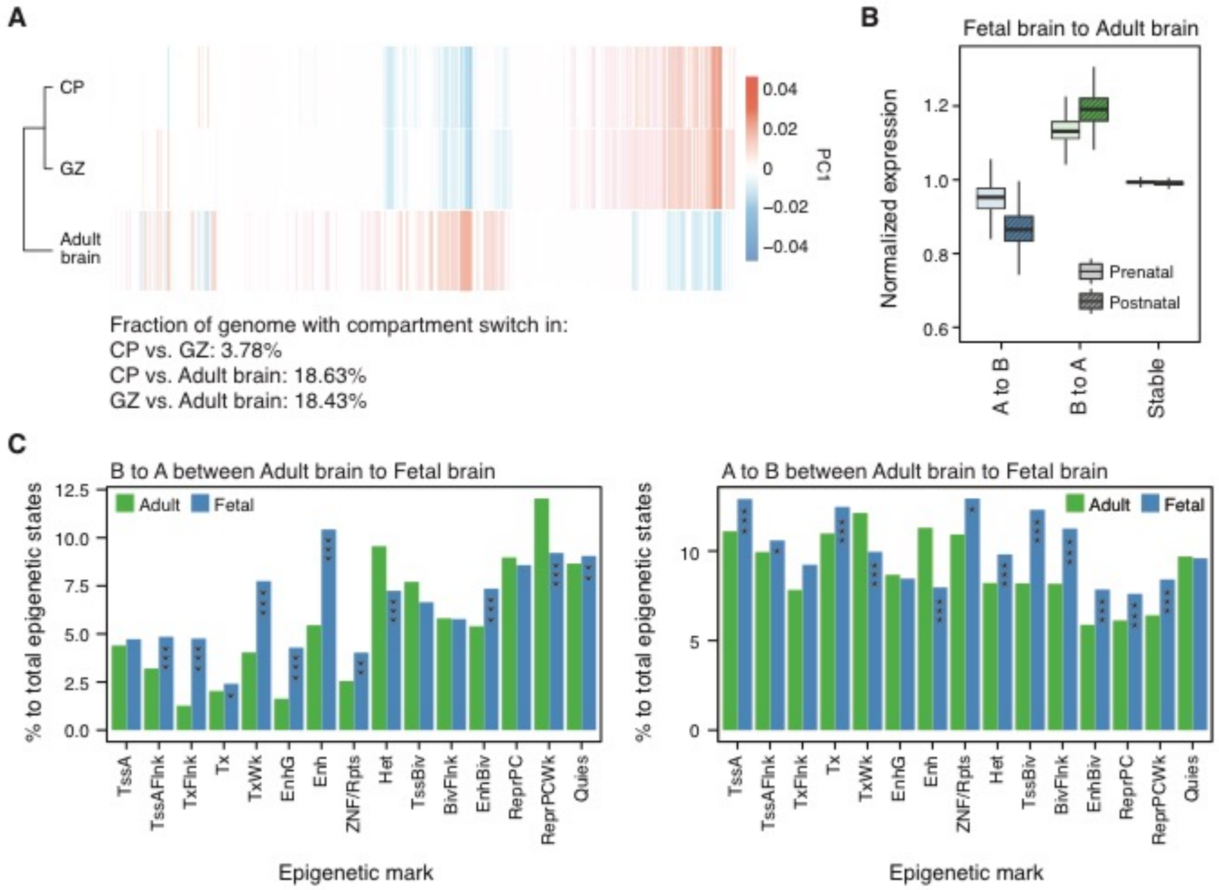


Fig. S6.2 Compartment switching across brain development is associated with expression and epigenetic changes. A. Heat map of the first principal component (PC1) values for regions that undergo compartment switching between fetal brain (CP and GZ) and adult brain. **B.** Brain expression levels for genes located in compartments that switch during development. **C.** Fraction of epigenetic states for regions that undergo compartment switching across brain development. For example, B to A shift in adult to fetal brain is accompanied by an increased proportion of active promoters (TssA, TssAFlnk), transcribed regions (Tx, TxWk), and enhancers (EnhG, Enh), and a decreased proportion of repressive elements (ReprPCWk) and heterochromatin (Het) in fetal brain compared with adult brain. *P < 0.05, **P < 0.01, ***P < 0.001. P values from Fisher's test.

Regulatory relationships across brain development. To compare the shared proportion of enhancer-promoter interactions in fetal vs. adult brain, we first collapsed putative enhancers (identified as promoter-based interactions) to each gene. We generated enhancer-gene links (e.g. chr10:100130000:ENSG00000230928) from fetal and adult brain and directly compared them. According to this analysis, 30.8% of enhancer-gene links detected from adult brain were also detected in fetal brain.

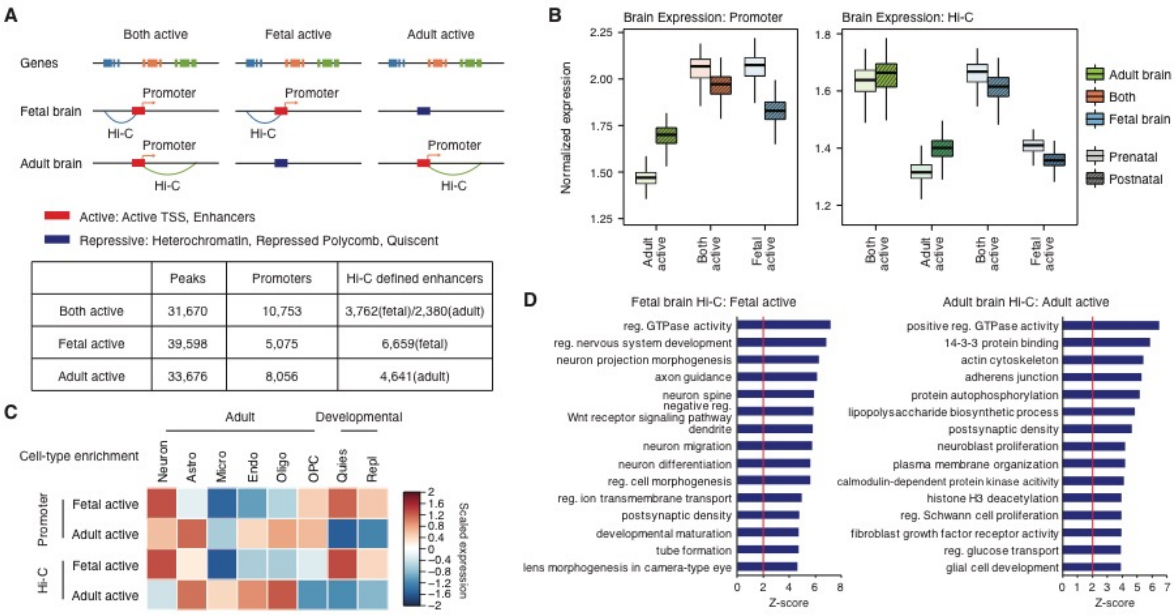
Using chromatin states defined by chromHMM (Kundaje et al., 2015) in fetal brain and adult PFC, we defined regulatory regions according to their developmental state changes: (1) Both active: elements that are active in both adult and fetal brain, (2) Fetal active: elements that are active in fetal brain and become repressive in adult brain, (3) Adult active: elements that are repressive in fetal brain then become active in adult brain. Active elements were defined as TssA, TssAFlnk, EnhG, Enh, while repressive elements were defined as Het, ReprPC, ReprPCWk (weak Polycomb repressive sites), and Quies (quiescent states). These elements are referred as developmental regulatory elements. Since developmental regulatory elements contain both promoters and enhancers, we then overlapped them with the promoter coordinates used to detect promoter-based interactions (see section S6.3). In total, we identified 6 types of developmental regulatory elements: both active promoters, both active enhancers, fetal active promoters, fetal active enhancers, both active promoters, both active enhancers.

We next assigned genes to developmental regulatory elements: elements that overlap with promoter coordinates were directly assigned to their genes based on linear genome, while the ones that do not overlap with promoter coordinates were thought as enhancers and assigned based on promoter-based interactions either from adult or fetal brain. Fetal active enhancers were assigned to their target genes based on fetal brain Hi-C, adult active enhancers were assigned based on adult brain Hi-C data, while both active enhancers were assigned based on both adult and fetal brain Hi-C data. In total, this analysis leads to 7 groups of genes that were linked to each element: both active promoters-linear assignment, fetal active promoters-linear assignment, adult active promoters-linear assignment, both active enhancers-fetal Hi-C, both active enhancers-adult Hi-C, fetal active enhancers-fetal Hi-C, adult active enhancers-adult Hi-C. Average centered expression values were calculated and plotted for each group, and gene ontology (GO) enrichment for each group was assessed using GoElite v77 (http://www.genmapp.org/go_elite/).

We also processed single cell expression values (in log2(TPM+1) forms, see Section S2.2.2) by centering to the mean expression level for each cell using a *scale(center=T, scale=F)* function in R. This results in centered expression values denoting each gene's relative expression level in a given cell, hereby referred as cell-level centered expression values. We then calculated average cell-level centered expression values for each group of genes mapped to distinct types of developmental regulatory elements.

Relationships between the enhancer number and gene expression. To measure the relationship between enhancer numbers and gene expression level, we integrated promoter-based interactions, brain active enhancers, and expression data. As enhancers and Hi-C interactions were defined in different resolution (Hi-C was defined at 10kb bin level, while enhancers were defined at much higher resolution), we clumped enhancers within 10kb bins so that they match with the Hi-C resolution. Intersecting brain active enhancers and promoter-based interactions led to 17,719 bin-level enhancer-promoter interactions. We grouped genes based on their number of interacting enhancers and their average centred expression values were calculated and plotted for each group. We also identified 90,015 enhancer-promoter interactions when we didn't clump enhancers into a bin-level.

Cis-regulatory relationship mediated by chromatin interactions. We overlapped eQTLs, isoQTLs, and cQTLs (hereby referred as QTLs) with Hi-C to measure the proportion of cis-regulatory relationship mediated by 3D interactions. As the type of chromatin interactions that mediate cis-regulatory relationship has not been well understood, we did not want to restrict our interaction search space into promoter-based interactions. Therefore, we first obtained chromatin interaction profiles of QTLs and then overlapped the profiles with (1) gene coordinates both at the exon and promoter levels (eQTL/isoQTL) or (2) coordinates of chromatin marks (cQTL).



	Peaks	Promoters	Hi-C defined enhancers
Both active	31,670	10,753	3,762(fetal)/2,380(adult)
Fetal active	39,598	5,075	6,659(fetal)
Adult active	33,676	8,056	4,641(adult)

Fig. S6.3 Dynamics of chromatin landscape across brain development. A. A schematic showing how brain regulatory elements were mapped to their putative target genes based on chromatin interaction profiles. Brain regulatory elements were first grouped into three categories: regulatory elements that are active in both developmental epochs (both active), regulatory elements in fetal brain (fetal active), and regulatory elements in adult brain (adult active). Brain regulatory elements that reside within promoters were directly assigned to their target genes (promoter-based assignment), while intergenic/intronic regulatory elements were assigned based on chromatin interactions either in fetal or adult brain (Hi-C based assignment). The number of brain regulatory elements (peaks) and genes mapped to regulatory elements by promoter- and Hi-C-based assignment is described in the bottom. **B.** Genes assigned to fetal active elements are prenatally enriched, while genes assigned to adult active elements are postnatally enriched. **C.** Genes assigned to fetal active elements are relatively more enriched in neurons in the adult (Adult-Neuron) and fetal brain (Developmental-Quies and Repl), while genes assigned to adult active elements are relatively more enriched in glia (astrocytes, endothelial cells, and oligodendrocytes). **D.** Gene ontology enrichment for genes that are assigned to fetal and adult active regulatory elements based on chromatin interactions. Fetal active elements were assigned to genes associated with neuronal differentiation and synaptic formation, while adult active elements were assigned to genes involved in gliogenesis and synaptic maturation.

We constructed background interaction profiles from all SNPs with the imputation score > 0.9 in the genome to fit null distribution of the expected interaction frequencies given the chromosome and distance (see section S6.3 for more details). Significance of interaction from each QTL was calculated as the probability of observing higher interaction frequencies under the fitted null distribution. Interactions with FDR<0.01 were selected as significant interactions, and the regions that significantly interact with QTLs were overlapped with genomic coordinates of promoter (defined as 2kb upstream of every TSS), exon coordinates (based on Gencode v19), and coordinates of chromatin marks used to detect cQTLs. When conducting chromatin interaction analysis for eQTL/isoQTL, we excluded QTLs that are located within promoter or exons (promoter/exonic QTLs) because there is a high probability that they are directly associated with the genes/chromatin marks in which they locate. We also excluded cQTLs within 20kb from chromatin marks as chromatin interactions within this range is undetectable.

An e-Gene/chromatin often has multiple QTLs due to the linkage disequilibrium (LD), which makes it difficult to identify causal variants. Therefore, instead of a direct comparison between eGenes/chromatin and genes/chromatin that physically interact with QTLs, we measured the fraction of eGenes/chromatin that also have Hi-C evidence. For this purpose, we grouped QTLs based on eGenes/chromatin and checked whether any of the QTLs for a given e-Gene/chromatin also physically interacts with the same e-Gene/chromatin.

According to this analysis, 31.9% of eQTLs and 12.4% of isoQTLs had Hi-C evidence, indicating that chromatin interactions may impact cis-regulatory relationships via gene regulation than isoform switching. We also found that 6.5% of cQTLs have Hi-C evidence. Although this overlap is lower than what we found from eQTLs and isoQTLs, we think this reflects the low power of cQTLs (292 samples for cQTL vs. 1,387 samples for eQTL). In details, 27.4% of eQTLs were supported by promoter-based interactions, while 30.9% were supported by exon-based interactions, suggesting that exon-level interactions also have potentials to affect gene regulation, which has not been previously studied. Given that 31.9% (< 27.4% promoter-based interactions + 30.9% exon-based interactions = 58.4%) of eQTLs are supported by either promoter or exon-level interactions, most of the exon-/promoter-based interactions are redundant, indicating a complex gene regulatory network. On the contrary, 10.9% of sQTLs were supported by promoter-based interactions, while 3.7% were supported by exon-based interactions, which are largely non-redundant (12.4% total Hi-C supported sQTL ~ 10.9% promoter-based interactions + 3.7% exon-based interactions = 14.6%). In total, 32% of the eGenes showed evidence of chromatin interactions, accounting for 239,837 eQTLs, 3,235 isoQTLs.

We then compared the significance of associations for Hi-C supported QTLs, promoter/exonic QTLs, and non-supported QTLs (intronic/intergenic QTLs that do not have Hi-C evidence). We grouped QTLs based on these three categories and compared the significance of associations for each group. We compared the distribution of -log10(P-values) for each group using a (pairwise) Wilcoxon test. When there are more than two groups to compare, multiple testing correction was performed using FDR.

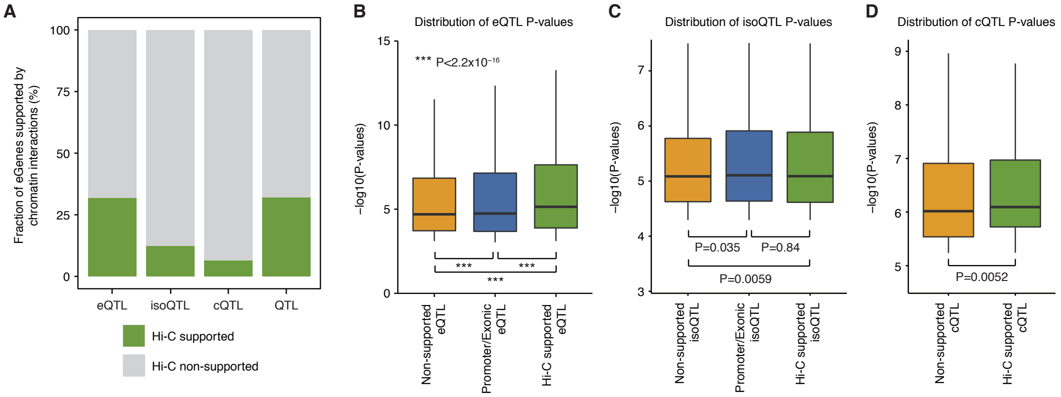


Fig. S6.4 Chromatin interactions mediate cis- and trans-regulatory relationships. A. A proportion of QTL-associated genes (eQTLs), isoforms (isoQTLs) and chromatin marks (cQTLs) that have Hi-C evidence. **B.** eQTLs supported by Hi-C evidence show stronger associations not only to eQTLs without genomic annotations (non-supported), but also to exonic and promoter eQTLs. **C-D.** isoQTLs (C) and cQTLs (D) supported by Hi-C evidence show stronger associations than those without genomic annotations (non-supported).

S6.5 Imputed gene regulatory networks (TFs)

We integrated and imputated all possible regulatory relationships in the frontal cortex including the enhancers, transcription factors (TFs), miRNAs and target genes in this resource. The first step involved inferring the positions of the TF binding sites (TFBSs) within the key regulatory elements in our model, namely, promoters and enhancers in TADs. To do this, we started with a previously generated genome-wide map of all the TFBSs using a list of 786 TF motif position weight matrices (PWMs) downloaded from CIS-BP (build 1.02, Weirauch et al., 2014), with TFBS locations on the hg19 genome build found using the program FIMO from the MEME suite (version 4.11.4, Grant et al., 2011) with a threshold of 0.00001.

Next, we defined the promoter regions by a window of ±1.25 kb (=2.5 kb in total) relative to the transcription start site (TSS), while the PEC enhancer regions of uniform length 1 kb were used. The ENCODE DNase hypersensitivity site (DHS) datasets for the frontal cortex (in .bed format) were then used to find open chromatin regions within the promoters, and the TFs with TFBSs within these open chromatin regions of the regulatory elements were linked to the corresponding elements. Since the PEC enhancers were already defined within regions of open chromatin, there was no need to further filter them out using DHS data, hence the TFs within the enhancers were directly linked to them. Finally, we tentatively link all enhancers and promoters within the same TADs determined from the Hi-C data on the reference brains (pooled data from three reference brains). The net result is a set of preliminary linkages in the form of [Enhancer TFs]=>Enhancers=>Promoters<=[Promoter TFs].

There are some noteworthy points on this analysis. Firstly, when the PEC enhancers were expanded to a uniform size of 1 kb, there were some overlaps between adjacent enhancers. With regard to the TF linkages, we resolved these overlaps by assigning a TF within the overlap region only to the first enhancer encountered in the sorted enhancer list. Secondly, there are two experimental DHS files for the frontal cortex from the ENCODE consortium, resulting in two different sets of TF linkages for the promoters. The results from the two replicates were merged into a single consensus set of linkages.

In total, we included 675,061 enhancer-target-promoter in TADs and 823,946 TF-target-promoter binding linkages, providing a reference wiring network on gene regulation in brain, which consists of the regulatory factors and elements (e.g., TFs, enhancers) and target genes. An associated data file with the reference TF network is available on the website (adult.psychencode.org).

To identify activated regulatory wires for a particular phenotype or disorder, we further used the method to determine such activated regulation. Given a gene and a phenotype/disorder, we applied the Elastic net regression, linearly combining the L_1 and L_2 regularizations to predict its gene expression data from the expression data of the TFs that have the binding sites on the gene's enhancers and promoter and overlap the QTLs; i.e., the QTLs break the binding sites. We then identified the activated TF-target regulatory relationships if TFs have large regression coefficients. In detail, suppose *Y* is an *N*-dimensional vector with elements being the gene's expression levels across samples, where *N* is the sample number for the phenotype/disorder. *X* is an *N* by *M* matrix whose columns are the TFs' expression levels, where *M* is the number of potential TFs. The Elastic net regression estimates the coefficients of *M* TFs, denoted by an M-dimensional vector, $B= \text{argmin}_B \|Y\text{-}XB\|^2+\text{alpha}\|B\|_{L2}+\text{beta}*\|B\|_{L1}$, where alpha and beta are parameters to adjust the contributions from L_2 and L_1 regularizations of B. The mean square error of Elastic net regression is equal to $\|Y\text{-}XB\|^2/N$ based on ⅔ training and ⅓ test data. For each gene and its TFs, we used the gene expression data across all adult samples (N=1866) in the resource to run the Elastic net regression. For example, we identified a strong regulatory relationship between four promoter TFs (NKX2-4, FOXE3, FOXI1, TFAP2B, coefficients >0.2) and three enhancer TFs (FOXA2, FOXI2, HMX2, coefficients > 1) with CHD8, a chromatin remodeler strongly associated with ASD. In total, we could predict the expression level of CHD8 with mean square error < 0.034.

We compared the HiC enhancer-promoter interactions and the interactions between eGenes and associated e/isoQTLs on enhancers with TF activity to determine a highly confident, overlapped enhancer-target-promoter linkages. In summary, there were 43,181 TF-to-target and 37,052 enhancer-to-target-promoter linkages among the top 5% Elastic net regression coefficients (absolute value >0.2). from at least two of these types: (i) activity relationships (~448k enhancer-to-target-promoter linkages), (ii) physical chromatin interactions (~91k Hi-C enhancer-promoter interactions), and (iii) 36,293 QTLs (e/isoQTL-SNP on brain enhancers to eGene). Associated data files with the final, Elastic-Net-Based TF network and HiC-derived enhancer-promoter linkages are on the website (adult.psychencode.org).

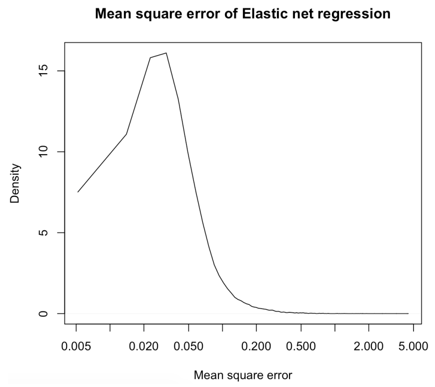


FIg. S6.5 Mean square error distribution of Elastic net regression predicting target gene expression from TF expression. The x-axis is the mean square error range across protein-coding target genes. The y-axis is the density of target genes. An associated data file with the mean square error values for each gene with an Elastic Net prediction is available on the website (adult.psychencode.org).

S7. Supp. content to main text section "Linking GWAS variants"

S7.1 Identification of GWAS associated genes for schizophrenia

We used 5,996 schizophrenia (SCZ)-associated autosomal putative causal (credible) SNPs reported in the original study (Pardiñas et al., 2018) and categorized them into promoter/exonic and intergenic/intronic SNPs. Promoter/exonic SNPs were directly assigned to the target genes based on the genomic coordinates, while intergenic/intronic SNPs were annotated based on chromatin interactions and enhancer-target-gene linkages supported by activity relationships from Elastic net regression. We used promoter-based interactions defined by Hi-C and enhancer-target-gene linkages to assess whether credible SNPs reside in (1) regions that physically interact with promoters of any genes (see Section S6.3) and/or (2) enhancer regions supported by activity relationships (see Section S6.5).

Credible SNPs colocalize with 2,064 eQTLs associated with 282 eGenes, 91 of which overlap with those identified by the Hi-C driven approach. To confirm this overlap is mediated by the shared causal variants in GWAS and eQTLs, we performed a colocalization test (Giambartolomei et al., 2014), from which we identified 293 genes across 79 loci in which GWAS and eQTLs share causal variants.

Collectively, we identified 176 genes across 83 loci from the direct assignment, 597 genes across 92 loci from the Hi-C driven approach, 388 genes across 37 loci from enhancer-target links, 293 genes across 79 loci from eQTL associations, and 29 genes across 23 loci from isoQTL associations. In total, this leads to 1,097 genes across 119 loci, which are referred as SCZ genes. We also selected risk genes that are identified by two or more metrics to obtain SCZ high-confidence (HC) genes (304 genes). Associated data files with the full list of 1,097 SCZ risk genes and the filtered list of 304 high-confidence SCZ genes are available on the website (adult.psychencode.org).

We compared SCZ risk genes defined by each metric (QTL=eQTL and isoQTL, Hi-C, and enhancer-target links) by performing an over-representation test. One key thing for an over-representation test is to define a background gene set, because each metric has different background genes. For example, 13,304 genes have enhancer-target links (hereby referred as E-T genes), 33,217 genes have QTLs, while Hi-C has the genome-wide search space. Therefore, we defined a background gene list by taking an intersect of eGenes and E-T genes. For each metric, we took an intersect of SCZ risk genes and the background gene set and used them for the Fisher's exact test.

To assess what fraction of SCZ genes have distal regulatory relationships with putative causal SNPs, we compared SCZ genes with the genes that locate within the LD regions with the index SNPs (r2>0.6, includes genes partly overlapping with LDs). We also ran the colocalization test using the currently largest public dataset of eQTLs from the CMC (Fromer et al., 2016), assigning 137 genes to 68 loci. Notably, our newly generated eQTLs identified twice more genes than CMC eQTLs.

S7.2 Functional enrichment analysis

To assess whether SCZ genes and SCZ HC genes are dysregulated in neuropsychiatric disorders, we performed enrichment analysis by logistic regression on (1) differentially expressed genes (DEGs) in three types of disorders (ASD=autism spectrum disorder, SCZ=schizophrenia, BD=bipolar disorder) identified by (Gandal, M.J. et al., *submitted*), (2) genes affected by rare LoF variants in SCZ (TADA<0.3; Singh et al., 2016), and (3) genes located in recurrent SCZ copy number variation (CNVs) (Marshall et al., 2017). For the enrichment analysis on SCZ rare variants and CNVs, we used protein-coding genes for a background gene list and regressed exon lengths out. For the enrichment analysis on DEG, we used a union of eGenes and E-T genes detected in our study as a background gene list.

We analyzed GO enrichment for SCZ genes and SCZ HC genes using GOElite. We used the union of detected eGenes and E-T genes as a background gene list.

We used cell-level centered expression values to get average centered expression values for SCZ and SCZ HC genes in each cell type. Cell types were grouped into the clusters neurons, astrocytes, OPC, oligodendrocytes, microglia, endothelial cells, fetal neurons, and the neuronal subcluster (excitatory and inhibitory neurons) and measured relative expression levels in a given cluster by a *scale* function in R.

S7.3 Identification of TFs associated with schizophrenia risk genes

TF-target regulatory relationships (see Section S6.5) were used to detect TFs that are enriched either in (1) promoters of SCZ genes or (2) enhancers that overlap with SCZ credible SNPs. We calculated the significance of the enrichment by P=Pbinom(k>=s, n=n, p=p), when p = fraction of promoters/enhancers associated with credible SNPs, n = the number of total binding sites of a TF A (TFBSA) in promoters/enhancers, s = the number of total promoter/enhancer TFBSA associated with credible SNPs (Fig. S7.1).

For promoter enrichment, p = the number of SCZ genes / the number of genes that have TF-target-promoter links from the elastic net; s = the number of TFBSA within promoters of SCZ risk genes. For enhancer enrichment, p = the length of enhancers that harbor SCZ credible SNPs / the length of enhancers that have TF-enhancer-target links from the elastic net; s = the number of TFBSA within enhancers that harbor SCZ credible SNPs. For promoter enrichment, we calculated an enrichment P-value for each TF, which was subsequently corrected for the number of TFs bound to gene promoters. For enhancer enrichment, an enrichment P-value for each TF was subsequently corrected for the number of TFs within enhancers that harbor SCZ credible SNPs.

S7.4 Partitioned heritability

We assessed heritability explained by brain regulatory elements (enhancers) and variants (eQTLs) for different GWAS using partitioned LD score regression (LDSC, Finucane et al., 2015; https://github.com/bulik/ldsc/wiki/Partitioned-Heritability). We included 9 brain disorder GWAS and 3 non-brain disorder GWAS (GWAS sets and sources described below) in an attempt to test that partitioned heritability estimates of brain disorders are more strongly enriched in brain enhancers and eQTLs than in non-brain disorders. For eQTLs, we included all eQTLs in the model, since LD scores count for LD. We also used top SNPs (pruned for LD r2>0.5) to ensure that the enrichment signal doesn't come from the spurious LD structures, where we got similar enrichment results.

Disorders	**Source**
ADHD	Demontis et al. 2017
ASD	Grove et al. 2017
Bipolar disorder	Ruderfer et al. 2014
Depression (Broad General Practice)	Howard et al. 2017
Schizophrenia	Pardinas et al. 2018
Educational attainment	Okbay et al. 2016
Intelligence	Sniekers et al. 2017
Alzheimer's disease	Lambert et al. 2013
Parkinson's disease	Nalls et al. 2014

Type 2 diabetes (T2D)	Morris et al. 2012
Coronary artery disease (CAD)	Schunkert et al. 2011
Inflammatory bowl disease (IBD)	Liu et al. 2015

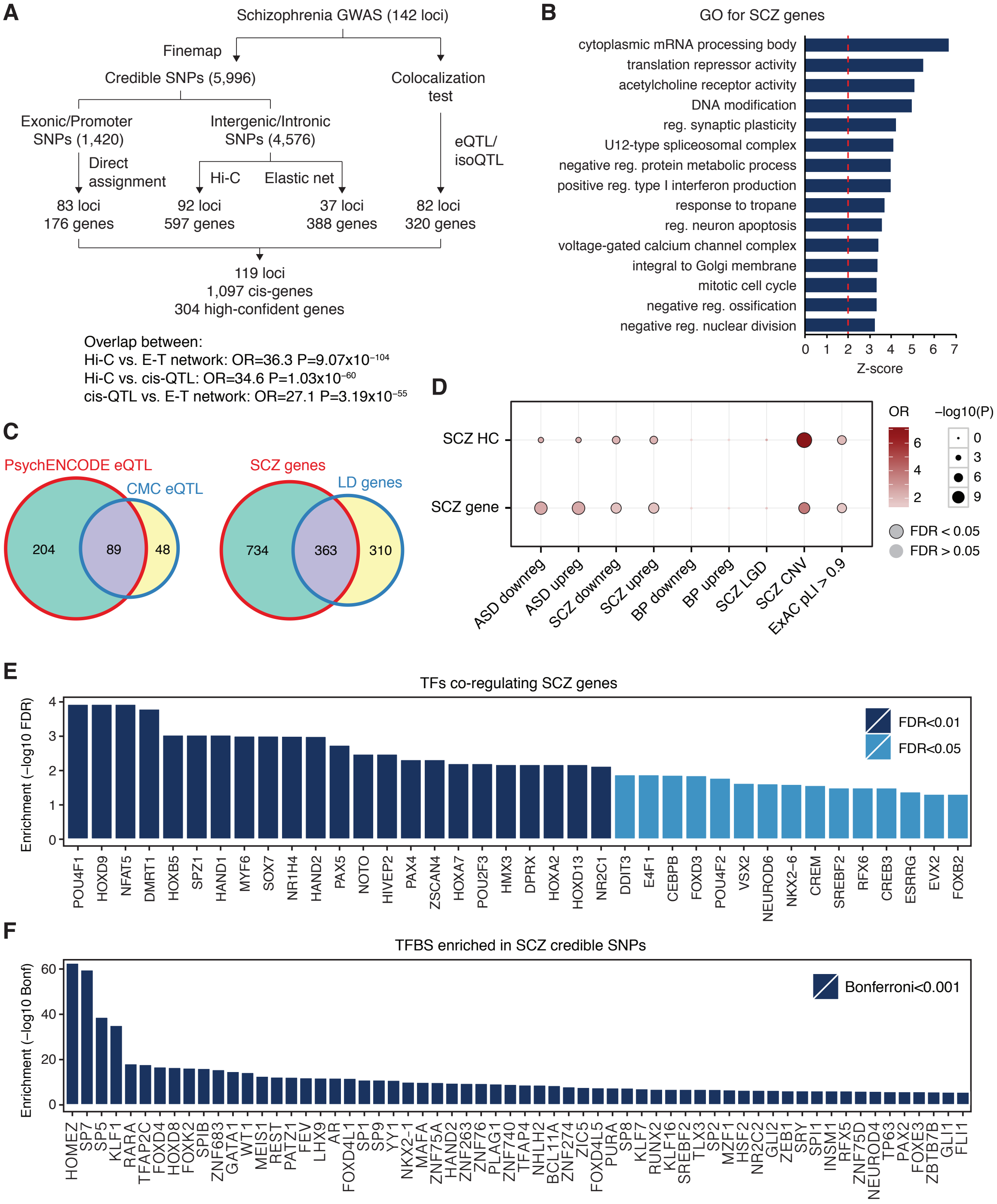


Fig. S7.1 Identification of schizophrenia risk genes. A. A schematic depicting how SCZ GWAS loci were assigned to putative genes. **B.** Gene ontology enrichment for SCZ-genes demonstrates that cholinergic receptors, synaptic genes, calcium channels, immune response-related genes, translational regulators, and RNA splicing regulators are associated with SCZ GWAS. **C.** Left, Colocalization analysis with eQTLs identified 2.13 fold more genes than the

CMC eQTLs (Fromer et al., 2016). Right, Most SCZ genes (66.2%) are not located in the genome-wide significant loci (LD defined as r2>0.6). **D.** SCZ risk genes are enriched for dysregulated genes in ASD and SCZ, genes affected by recurrent copy number variations (CNV) in SCZ (SCZ CNV), and genes intolerant to loss-of-function mutations (ExAC pLI>0.9). SCZ LGD, genes that harbor likely gene disrupting (LGD) mutations in SCZ; HC, SCZ high-confidence genes; Downreg, downregulation; Upreg, upregulation. **E.** TFs that are significantly enriched in promoter regions of SCZ genes. **F.** TFs that are significantly enriched in enhancers that harbor SCZ credible SNPs.

S8. Supp. content to main text section "Deep-learning model"

S8.1 Data

We integrate data of the kinds described above into a single model connecting genotype, functional genomics and phenotype data from PsychENCODE in the Prefrontal Cortex. We build separate models for the phenotypes Schizophrenia (SCZ), Bipolar disorder (BPD), Autism spectrum disorder (ASD), age (AGE), gender (GEN) and reported ethnicity (ETH). For each phenotype, we created 10 balanced train / test splits as described below, and we report the performance of all models averaged across these 10 splits of the data. For the disease conditions, these splits contain equal numbers of cases and controls, while for age, gender and ethnicity, only control samples are used. As inputs to the model during training, we use the imputed genotypes; intermediate phenotype data including gene expression, enhancer h3k27ac activation levels, cell fraction estimates, and co-expression module mean expression; and high-level phenotype data corresponding to the categories above. Normalization of the gene expression and enhancer activation data was identical to that used in the QTL calculations. Also, for the cRBM, cDBM and DSPN models, all functional genomics data was binarized by thresholding at the median value (per gene/enhancer/cell-type/module). Further, DSPN model connectivity was constrained by using the estimated eQTLs, cQTLs and fQTLs, along with the Gene Regulatory Network (GRN) TF-gene and enhancer-gene linkages estimated in the elastic net analysis.

8.1.1 Balanced Datasets

We first describe how the balanced datasets are created for SCZ, and then describe how the balanced datasets are created for the other high-level phenotypes using a similar process with small modifications. For SCZ, we divide the PEC data into subsets, each containing samples from a common assay (BipSeq, brainGVEX, CMC, CMC-HBCC, Libd, UCLA-ASD, Yale-ASD or GTEx-DFC), the same gender (M or F), the same ethnicity (Caucasian (CC) or African American (AA), to which most samples belonged), and the same age range (1-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+). For each subset, we found all SCZ and control (CTR) samples within that subset, with counts m and n for the number of cases and controls respectively. We then sampled uniformly without replacement $N_{subset} = \min(m, n)$ SCZ samples and N_{subset} CTR samples from the subset to add to a 'pool' of samples for the current data split. After having done this for all such subsets so that the pool contains N_{pool} SCZ and N_{pool} CTR samples, we partition the case samples randomly into groups of size $t_1 = \lceil \tau_{split} \cdot N_{pool} \rceil$ and $t_2 = N_{pool} - t_1$ for training and testing respectively ($\tau_{split} = 0.9$), and do likewise to add equal numbers (t_1, t_2) of controls to each partition. We repeat the whole process 10 times to generate 10 data splits; the above process ensures that each training and test partition contains a 50/50 split of SCZ/CTR samples, and additionally that the distribution of covariates (assay, gender, ethnicity and age) is approximately the same for cases and controls in the training and testing partitions.

Exactly the same method is used to create balanced data splits for BPD. For ASD, due to the limited number of cases, we set $\tau = 0.8$, and balance only for assay and gender (not for ethnicity and age range). For the non-disease phenotypes (AGE, GEN and ETH), a similar method is used but with the following modifications. Here, we use CTR samples only, and split the PEC data into subsets containing samples from a common assay, and which are matched on all covariates as above except the high-level phenotype being modeled. Then, equal numbers of samples are randomly selected for each binary value of the modeled phenotype to be added to the training/testing partitions (respectively t_1 and t_2 for training and testing as above); for GEN the binary values are M/F, for ETH they are CC/AA, and for AGE we binarize the trait as 0/1 such that 1 indicates that a sample is older than the median age of 51 (NB the median age binarization is used only when AGE is the modeled phenotype; for all other phenotypes age is balanced using the decade age bins as above). The above method generates 10 data splits each of the following sizes (training/testing): SCZ (640/70); BPD (170/18); ASD (50/12); AGE (244/26); ETH (284/30); GEN (312/34).

S8.2 Model descriptions, training and inference with observed intermediate phenotypes

8.2.1 Logistic regression (LR)

We train LR models to predict a binary phenotype from a single level of predictors (either genotype or an intermediate phenotype). The model has the form:

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x} + b), \tag{1}$$

where y is the phenotype, $\mathbf{x}$ is a vector of predictors, $\mathbf{w}$ is a weight vector, b is the bias term, and σ is the logistic function, $\sigma(a) = 1/(1 + e^{-a})$. Since training and test sets are both balanced, for a test sample we use the predictor $y_{test} = [\mathbf{w} \cdot \mathbf{x}_{test} + b > 0.5]$, where $[a]$ is the Iverson bracket, which is 1 if a is true, and 0 otherwise.

For each data split, we initially perform feature selection by calculating the correlation of each predictor with the high-level phenotype:

$$s_j = \mathrm{corr}([y_1, y_2, \dots y_N], [x_{1j}, x_{2j} \dots x_{Nj}]), \tag{2}$$

where y_i is the phenotype of the i'th training sample, x_{ij} is the value of the j'th predictor at the i'th training sample, and corr is the Pearson correlation function, $\mathrm{corr}(\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b}/(||\mathbf{a}|| \cdot ||\mathbf{b}||)$. To perform feature selection, we rank the predictors by the absolute value of s_j in descending order for a given training split, and include only predictors 1 ... $\lceil \pi N \rceil$ in the model for that data split. We learn two LR models for each phenotype, the first using the imputed genotypes at the eSNPs as predictors, and the second using PFC gene expression levels (transcriptome) as predictors. We set $\pi = 0.01$ and $\pi = 0.0001$ for the genotype and transcriptome models respectively. For optimization, we use the Matlab Statistics and Machine Learning toolbox (glmfit).

8.2.2 Conditional Restricted Boltzmann Machine (cRBM)

A Restricted Boltzmann Machine (RBM) models the joint distribution of a set of visible and hidden units; we will denote the visible units as $\mathbf{x}$ and y corresponding to the intermediate and high-level phenotypes respectively, and the hidden units as $\mathbf{h}$, all of which are binary variables (multivariate in the case of $\mathbf{x}$ and $\mathbf{h}$). An RBM has the form $p(\mathbf{x}, \mathrm{y}, \mathbf{h}) = \exp\left(-E_{RBM}(\mathbf{x}, \mathrm{y}, \mathbf{h})\right)/Z$, where Z is a normalizing partition function, and $E_{RBM}(\mathbf{x}, \mathrm{y}, \mathbf{h})$ is the RBM energy function, which has the form $E_{RBM}(\mathbf{x}, y, \mathbf{h}) = -[\mathbf{x}^{\mathrm{T}}, y]\mathbf{W}\mathbf{h} - [\mathbf{x}^{\mathrm{T}}, y]\mathbf{b_1} - \mathbf{h}^{\mathrm{T}}\mathbf{b_2}$, where $\mathbf{W}$ is a matrix of interaction weights between the visible and hidden units, and $\mathbf{b_1}$ and $\mathbf{b_2}$ are the visible and hidden bias terms respectively. A Conditional RBM (cRBM) models the conditional distribution of a set of visible and hidden units on a further set of conditioning (visible) units (see Mnih et al., 2012), which we will denote $\mathbf{z}$, and which are assumed to be discrete:

$$p(\mathbf{x}, \mathrm{y}, \mathbf{h}|\mathbf{z}) = \exp\left(-E_{cRBM}(\mathbf{x}, \mathrm{y}, \mathbf{h}|\mathbf{z})\right)/Z(\mathbf{z}),$$

$$E_{cRBM}(\mathbf{x}, \mathrm{y}, \mathbf{h}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{x} - [\mathbf{x}^{\mathrm{T}}, y]\mathbf{W}\mathbf{h} - [\mathbf{x}^{\mathrm{T}}, y]\mathbf{b_1} - \mathbf{h}^{\mathrm{T}}\mathbf{b_2}, \tag{3}$$

$$Z(\mathbf{z}) = \sum_{\mathbf{x},\mathrm{y},\mathbf{h}} \exp\left(-E_{cRBM}(\mathbf{x}, \mathrm{y}, \mathbf{h}|\mathbf{z})\right),$$

where $\mathbf{V}$ is a matrix of interaction weights between the conditioning and visible units (which are restricted here to exclude interactions involving y, and hence model only dependencies between genotype $\mathbf{z}$ and phenotype y which are mediated by the intermediate phenotypes $\mathbf{x}$).

Both the RBM and cRBM may be trained using Contrastive Divergence (CD). In the case of the cRBM, CD finds an approximate gradient to the conditional log-likelihood of the training data:

$$\frac{\partial \log\left(p(\mathbf{x}, \mathrm{y}|\mathbf{z})\right)}{\partial w_{ij}} = < x_i h_j|\mathbf{z} >_0 - < x_i h_j|\mathbf{z} >_\infty \approx < x_i h_j|\mathbf{z} >_0 - < x_i h_j|\mathbf{z} >_1 = \mathrm{CD}(w_{ij}), \quad (4)$$

where $< a >_n$ denotes the expected value of a after performing n steps of alternating Gibbs sampling, starting with the visible units fixed to the training data (see Hinton, 2012 for the RBM case). Approximate gradients for interactions involving y and $\mathbf{z}$ and the bias terms may be found similarly by estimating the expected statistics for $x_i y$, $z_i x_j$, x_i and z_i after one step of alternating Gibbs sampling. The step size for the change in w_{ij} at iteration t, $\Delta_t(w_{ij})$, may then be calculated as:

$$\Delta_t(w_{ij}) = \alpha\Delta_{t-1}(w_{ij}) - \epsilon \mathrm{CD}(w_{ij}) - C w_{ij}, \quad (5)$$

where α is a momentum parameter, ϵ is the learning rate, and C is a weight cost to encourage sparsity. At each iteration, we evaluate Eq. 5 using a subset of the training samples (a mini-batch), hence performing stochastic gradient descent (SGD). We cycle once through the training data in disjoint mini-batches to form an epoch, and use early stopping after τ_{stop} epochs to control for overfitting.

Given a test sample, we wish to predict y given $\mathbf{x}$ and $\mathbf{z}$ (or $\mathbf{z}$ alone for imputation based inference, see below). This can be achieved by maximizing the conditional probability of y and $\mathbf{x}$ given $\mathbf{z}$, or equivalently minimizing the free-energy (see Hinton, 2012 for the RBM case):

$$\mathrm{argmax}_y\left(p(\mathbf{x}, \mathrm{y}|\mathbf{z})\right) = \mathrm{argmin}_y\left(F(\mathbf{x}, \mathrm{y}|\mathbf{z})\right)$$

$$F(\mathbf{x}, \mathrm{y}|\mathbf{z}) = -\sum_i b_{1i}x_i - b_{1y}y - \sum_{ij} v_{ij}z_i x_j - \sum_j \log\left(1 + \exp\left(b_{2j} + \sum_i x_i w_{ij} + y w_{yj}\right)\right). \quad (6)$$

We use the 10 balanced data split above to train a series of models for each phenotype. We initially perform feature selection (for each data split) using the method in Eq. 2 to identify a subset of genes to include as transcriptome predictors in $\mathbf{x}$ (setting $\pi = 0.05$), and include all eSNPs associated with these genes in $\mathbf{z}$. We also enforce sparsity on the matrix $\mathbf{V}$ during training, so that only connections supported by eQTLs are allowed to be non-zero. Further, we set $N_h = 400$ (the number of hidden nodes), $\alpha = 0.1$, $\epsilon = 1e - 4$, and used mini-batches of size 61, 10, 17, 71, 39 and 64 for AGE, ASD, BPD, ETH, GEN and SCZ models respectively. For τ_{stop}, we used either a variable setting which was set independently for each model trained, or a fixed setting which was held constant across all data-splits for a given phenotype. In each case, we trained all models for 100 epochs. For the variable setting, we chose τ_{stop} to minimize the test error for each data-split separately, while for the fixed setting we chose the τ_{stop} which had the minimum mean test error across data-splits. Results are shown using both variable (Fig. 6D) and fixed (Table. S8.1) settings for all phenotypes except ASD; for ASD we use only the fixed setting to control for the smaller number of samples in the ASD cohort. Performance for each phenotype is calculated as an average across data splits for the accuracy of a model on its corresponding test partition.

8.2.3 Conditional Deep Boltzmann Machine (cDBM)

A Deep Boltzmann Machine (DBM) may be defined as in (Salakhutdinov and Hinton, 2012) as a Boltzmann machine with additional structure such that it can be viewed as a stack of RBMs. The model with two hidden layers has the form: $p(\mathbf{x}, \mathbf{y}, \mathbf{h_1}, \mathbf{h_2}) = \exp\left(-E_{DBM}(\mathbf{x}, \mathbf{y}, \mathbf{h_1}, \mathbf{h_2})\right)/Z$, where Z is a normalizing partition function, and $E_{DBM}(\mathbf{x}, \mathbf{y}, \mathbf{h_1}, \mathbf{h_2})$ is the DBM energy function, which has the form $E_{DBM}(\mathbf{x}, \mathbf{y}, \mathbf{h_1}, \mathbf{h_2}) = -\mathbf{x}^\mathrm{T}\mathbf{W_1}\mathbf{h_1} - \mathbf{h}_\mathbf{1}^\mathrm{T}\mathbf{W_2}\mathbf{h_2} - \mathbf{h}_\mathbf{2}^\mathrm{T}\mathbf{W}_{lab}\mathbf{y} - \left[\mathbf{x}^\mathrm{T}, \mathbf{h}_\mathbf{1}^\mathrm{T}, \mathbf{h}_\mathbf{2}^\mathrm{T}, \mathbf{y}\right]\mathbf{b}$. Here, $\mathbf{W_1}$, $\mathbf{W_2}$, $\mathbf{W}_{lab}$ are matrices of interaction weights between the visible and first-layer hidden units, the first and second layer hidden units, and the 'labels' and second-layer hidden units respectively. For the DBM, we write $\mathbf{y}$ as a vector, since for convenience we assume the class variables (high-level phenotypes) are represented using one-of-n encoding (i.e. for a binary trait, either $[1,0]^{\mathrm{T}}$ or $[0,1]^{\mathrm{T}}$ for the two classes), and we write $\mathbf{b}$ for a single vector combining all the bias terms.

As for the cRBM, we can use a family of DBMs to model a conditional distribution which depends on a further set of variables, $\mathbf{z}$. This is equivalent to converting the DBM from a Markov Random Field (MRF) into a Conditional Random Field (CRF, see Koller and Friedmann, 2009). We can thus define a conditional DBM analogously to the cRBM:

$$\begin{gathered} p(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}|\mathbf{z}) = \exp\left(-E_{cDBM}(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}|\mathbf{z})\right)/Z(\mathbf{z}), \\ E_{cDBM}(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W_1}\mathbf{h_1} - \mathbf{h}_1^{\mathrm{T}}\mathbf{W_2}\mathbf{h_2} - \mathbf{h}_2^{\mathrm{T}}\mathbf{W}_{lab}\mathbf{y} - \left[\mathbf{x}^{\mathrm{T}},\mathbf{h}_1^{\mathrm{T}},\mathbf{h}_2^{\mathrm{T}},\mathbf{y}\right]\mathbf{b}, \\ Z(\mathbf{z}) = \sum_{\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}} \exp\left(-E_{cDBM}(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}|\mathbf{z})\right). \end{gathered} \tag{7}$$

The cDBM can be trained by adapting the Persistent Markov Chain Monte Carlo algorithm used in (Salakhutdinov and Hinton, 2012). In this approach, following a pre-training phase which uses CD to train adjacent layers as RBMs, the weights for the whole network are optimized jointly by approximating the gradient to the full data log-likelihood of the model. For the cDBM, we can write the approximation as:

$$\frac{\partial \log\left(p(\mathbf{x},\mathbf{y}|\mathbf{z})\right)}{\partial w_{1ij}} \approx < x_i h_{1j}|\mathbf{z} >_{MF} - < x_i h_{1j}|\mathbf{z} >_{pMCMC} = \mathrm{pMCMC}(w_{1ij}), \tag{8}$$

where for convenience we show only the gradient for a weight in matrix $\mathbf{W_1}$. The first term $<.>_{MF}$ uses a mean-field approximation to evaluate the conditional expectation of $x_i h_{1j}$ when $\mathbf{x}$ and $\mathbf{z}$ are clamped to their observed values (due to this clamping, the unimodal form of the mean-field distribution is expected to hold approximately). Mean-field updates in the cDBM may be calculated straightforwardly by incorporating terms involving $\mathbf{V}$ into the energy. The second term approximates the model statistics with $\mathbf{x}$ unclamped; in the case of the DBM a set of N_{pMC} persistent Markov Chains are maintained for this purpose, each tracking the trajectory of a 'fantasy particle' consisting of a joint setting of the model variables $(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2})$. The fantasy particles make a fixed number of updates at each gradient iteration using the current model weight settings, and are not re-initialized (hence 'persisting') between gradient updates (each can be thought of as a series of Markov chains with changing parameters, or a single Markov chain over the model variables and weight parameters). The fantasy particles can then be used to estimate the required model expectations for the gradient. A similar approach can be used for the cDBM, only because the required term in the gradient is now a conditional expectation, it cannot be estimated by calculating expectations over a set of fantasy particles all evolving according to the same Markov process. Rather, a set of fantasy particles is required for each training sample ($N_{pMC} = N_{fantasy} \cdot N_{train}$), each evolving according to a Markov process conditioned on that sample's $\mathbf{z}$ value, and the expectation is calculated across the entire collection. Stochastic gradient updates are then made to the weights as in Eq. 5 (substituting $\mathrm{pMCMC}(.)$ for $\mathrm{CD}(.)$). Finally, as in (Salakhutdinov and Hinton, 2012) back-propagation can be applied for fine-tuning, and we use a single forward pass through the network for prediction. Settings of the parameters above are described in the context of the DPSN in the following section.

8.2.4 Deep Structured Phenotype Network (DSPN)

We define a Deep Structured Phenotype Network (DSPN) as a conditional Deep Boltzmann Machine, with extra structure added to the visible units to reflect regulatory relationships between various intermediate phenotypes. The general form of the model is:

$$\begin{gathered} p(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}|\mathbf{z}) = \exp\left(-E_{DSPN}(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}|\mathbf{z})\right)/Z(\mathbf{z}), \\ E_{DSPN}(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{U}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W_1}\mathbf{h_1} - \mathbf{h}_1^{\mathrm{T}}\mathbf{W_2}\mathbf{h_2} - \mathbf{h}_2^{\mathrm{T}}\mathbf{W}_{lab}\mathbf{y} - \left[\mathbf{x}^{\mathrm{T}},\mathbf{h}_1^{\mathrm{T}},\mathbf{h}_2^{\mathrm{T}},\mathbf{y}\right]\mathbf{b}, \\ Z(\mathbf{z}) = \sum_{\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}} \exp\left(-E_{DSPN}(\mathbf{x},\mathbf{y},\mathbf{h_1},\mathbf{h_2}|\mathbf{z})\right), \end{gathered} \tag{9}$$

which is identical to the cDBM, except for the introduction of a matrix of interaction terms $\mathbf{U}$ between the visible units. However, we also require that $\mathbf{x}$, $\mathbf{U}$ and $\mathbf{V}$ have specific forms, such that:

$$\begin{gathered}\mathbf{x} = \left[\mathbf{x}_{\mathrm{gene}}^{\mathrm{T}}, \mathbf{x}_{\mathrm{enh}}^{\mathrm{T}}, \mathbf{x}_{\mathrm{frac}}^{\mathrm{T}}, \mathbf{x}_{\mathrm{mod}}^{\mathrm{T}}\right]^{\mathrm{T}}, \\ \mathbf{x}^{\mathrm{T}}\mathbf{U}\mathbf{x} = \mathbf{x}_{\mathrm{gene}}^{\mathrm{T}}\mathbf{U}_{\mathrm{GRN}}\mathbf{x}_{\mathrm{gene}} + \mathbf{x}_{\mathrm{enh}}^{\mathrm{T}}\mathbf{U}_{\mathrm{ET-links}}\mathbf{x}_{\mathrm{gene}} + \mathbf{x}_{\mathrm{frac}}^{\mathrm{T}}\mathbf{U}_{\mathrm{markers}}\mathbf{x}_{\mathrm{gene}} + \mathbf{x}_{\mathrm{mod}}^{\mathrm{T}}\mathbf{U}_{\mathrm{WGCNA}}\mathbf{x}_{\mathrm{gene}}, \\ \mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{x} = \mathbf{z}^{\mathrm{T}}\mathbf{V}_{\mathrm{eQTL}}\mathbf{x}_{\mathrm{gene}} + \mathbf{z}^{\mathrm{T}}\mathbf{V}_{\mathrm{cQTL}}\mathbf{x}_{\mathrm{enh}} + \mathbf{z}^{\mathrm{T}}\mathbf{V}_{\mathrm{fQTL}}\mathbf{x}_{\mathrm{frac}} + \mathbf{z}^{\mathrm{T}}\mathbf{V}_{\mathrm{modQTL}}\mathbf{x}_{\mathrm{mod}}, \qquad (10)\end{gathered}$$

where $\mathbf{x}_{\mathrm{gene}}^{\mathrm{T}}$, $\mathbf{x}_{\mathrm{enh}}^{\mathrm{T}}$, $\mathbf{x}_{\mathrm{frac}}^{\mathrm{T}}$, $\mathbf{x}_{\mathrm{mod}}^{\mathrm{T}}$ are (binarized) representations of the gene expression, enhancer activity (h3k27ac level), cell-type fraction and co-expression module net activation respectively; $\mathbf{U}_{\mathrm{GRN}}$ is a sparse matrix where non-zero entries are allowed only between genes having a TF-target relationship determined by the elastic net model; $\mathbf{U}_{\mathrm{ET-links}}$ is a sparse matrix where non-zeros are allowed only between enhancers and genes when an enhancer-target link is determined by the elastic net model; $\mathbf{U}_{\mathrm{markers}}$ and $\mathbf{U}_{\mathrm{WGNA}}$ are sparse matrices where non-zero entries are allowed only between a cell-type/co-expression module and the marker-genes/member-genes associated with it respectively; and $\mathbf{V}_{\mathrm{eQTL}}, \mathbf{V}_{\mathrm{cQTL}}, \mathbf{V}_{\mathrm{fQTL}}, \mathbf{V}_{\mathrm{modQTL}}$are sparse matrices with non-zero elements allowed only between SNPs and genes/enhancers/cell-types/modules supported by a QTL linkage. We note that the results of previous analyses (e.g. elastic net and QTL analyses) are used only to establish the sparse structure of the $\mathbf{U}$ and $\mathbf{V}$ matrices, but not the actual linkage values of the non-zero entries, which are learned during joint training of the DSPN model (along with the $\mathbf{W}$ and $\mathbf{b}$ parameters). In general, we do not expect the magnitudes established independently for these linkages in the previous analyses to relate in a straightforward way to their optimal settings in a joint model, and hence we use only the connectivity structure as prior information during training.

The DSPN model can be trained similarly to the cDBM using persistent MCMC as described above. Mean-field approximate inference and Gibbs sampling steps are straightforwardly adapted to incorporate the additional linkages between the visible units. Because of the dependencies within the visible units, the mean-field and sampling steps cannot be made in parallel for the visible layer unlike the cDBM; for this reason, we choose a random update schedule of the nodes within the visible layer on each iteration, and update all other layers in parallel as before. In principle, the approach described learns a model representing the joint distribution of intermediate and high-level phenotypes conditioned on genotypes, and so can be used for prediction of high-level phenotypes either directly from the intermediate phenotypes, or from the genotype with imputation when the intermediate layers are unobserved. However, we adopt a slightly different training process when the goal is to provide a model for inference with imputed intermediate phenotypes, as described below, to optimize performance for this scenario. We summarize here the parameter settings for the model with direct observations: we perform feature selection as in Eq. 2 for each intermediate phenotype (setting $\pi = 0.05$); additionally, we set $N_{h_1} = 400$ and $N_{h_2} = 100$ (the number of hidden nodes in layers 1 and 2 respectively), $N_{fantasy} = 5$, $\alpha = 0.1$, $\epsilon = 1e - 4$, and use variable/fixed settings of τ_{stop} and mini-batch sizes as described above for the cRBM.

S8.3 Imputation of intermediate phenotypes

8.3.1 Deep Structured Phenotype Network with Imputation (DSPN-imput)

To optimize performance for prediction of high-level phenotypes from genotype data with imputation of intermediate phenotypes, we adopt a specialized training process. We assume that during training, we have access to fully observed genotype and intermediate phenotype data. Additionally, we split the training data for each data split evenly into training and validation partitions.

First, we train logistic regression models independently to predict each intermediate phenotype (e.g. gene expression level, enhancer activation) from the genotype at each of its QTLs using the training partition. We then fix the $\mathbf{V}$ matrices of the DSPN directly to the coefficients of the logistic regression models, and train $\mathbf{U}'$ and $\mathbf{W_1}'$ matrices (along with the biases for the visible layer and first hidden layer; primes indicate that these parameters are initial estimates only) by optimizing $\boldsymbol{p}(\mathbf{x}|\mathbf{z})$ on the validation partition, while fixing all hidden nodes at the second layer to 0; since we only allow one level of hidden nodes to vary, this model is equivalent to a cRBM (with additional structure on the visible nodes), and hence we use the Contrastive Divergence (Eq. 5) for optimization. Additionally, we perform feature selection at this stage by only including in the model the top $\pi_{\text{gene}}, \pi_{\text{enh}}, \pi_{\text{frac}}, \pi_{\text{mod}}$ proportion of intermediate phenotypes for each respective type as order by their predictive accuracy using the initial logistic predictor. We then use the partial cRBM model over $(\mathbf{z}, \mathbf{x}, \mathbf{h_1})$ to jointly infer estimated intermediate phenotype data for the validation samples, which we label $\mathbf{x_{imput}}$ (we infer $\mathbf{x_{imput}}$ by initializing it to the maximum likelihood outputs of the logistic predictors, and performing Gibbs sampling according to the cRBM energy function to refine this estimate). Finally, we train a full DSPN (with $\mathbf{V}$ still fixed) on the validation data, but optimized using the imputed rather than the original intermediate phenotype data, i.e. using $(\mathbf{z}, \mathbf{x_{imput}}, \boldsymbol{y})$ as training samples.

At test time, we do not make use of the intermediate phenotype data. Instead, we follow a similar path to training, by first imputing the intermediate phenotype data using the partial cRBM model with parameters $\mathbf{V}$, $\mathbf{U}'$ and $\mathbf{W_1}'$ (initialized using the individual logistic predictors used to form $\mathbf{V}$). We then treat the imputed phenotype data as fixed, and predict the associated high-level phenotype data from the full DSPN model using a forward pass as described for prediction in the cDBM model. We train the imputation based DSPN model using the same hyper-parameters as for the DSPN above, while setting $\pi_{\text{gene}} = 0.01$, $\pi_{\text{enh}} = 0.01$, $\pi_{\text{frac}} = 0.5$, $\pi_{\text{mod}} = 0.05$.

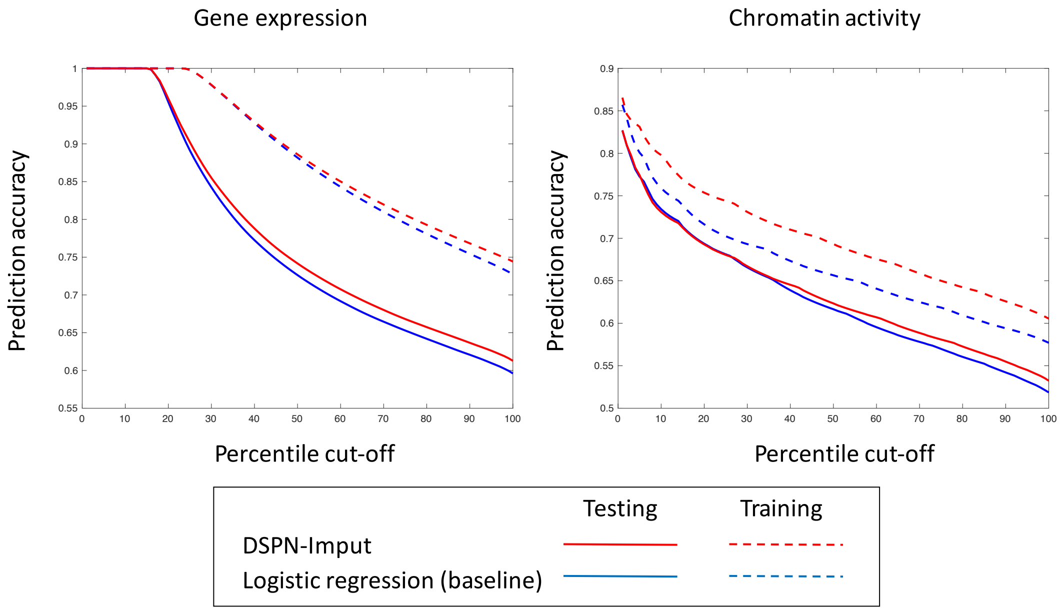


Fig. S8.1 Accuracy of intermediate phenotype imputation using DSPN-imput model Figure compares prediction accuracy for gene expression and chromatin activity using the full DSPN-imput model (with GRN structure included) vs prediction with a logistic model (independent prediction). Performance on training and testing partitions is shown.

S8.4 Variance explained on liability scale

To covert predictive performance of all models onto the liability scale, we use the following conversion due to Falconer (see International Schizophrenia Consortium, 2009; Falconer and MacKay, 1996):

$$v_{liab} = 2p_{pos}(1 - p_{pos})(\text{GRR} - 1)^2/\text{i}^2, \tag{11}$$

Here, v_{liab} is the variance explained on the liability scale, p_{pos} is the probability the model predicts a genotype to be a case, GRR is the genotype relative risk, and $i = z/K$, where K is the disease prevalence, and z the height of a standard normal distribution when the cumulative distribution has height $(1 - K)$. Letting a, b, c, d be the true negatives, false negatives, false positives and true positives respectively for a given model on test data, we estimate $p_{pos} = (c + \left(\frac{K}{1-K}\right) d)/(a + c + \left(\frac{K}{1-K}\right)(b + d))$, and GRR $=$ $(\frac{d}{c+d})/(\frac{b}{a+b})$. We set $K = 0.011, 0.01, 0.015$ for SCZ, BPD and ASD respectively.

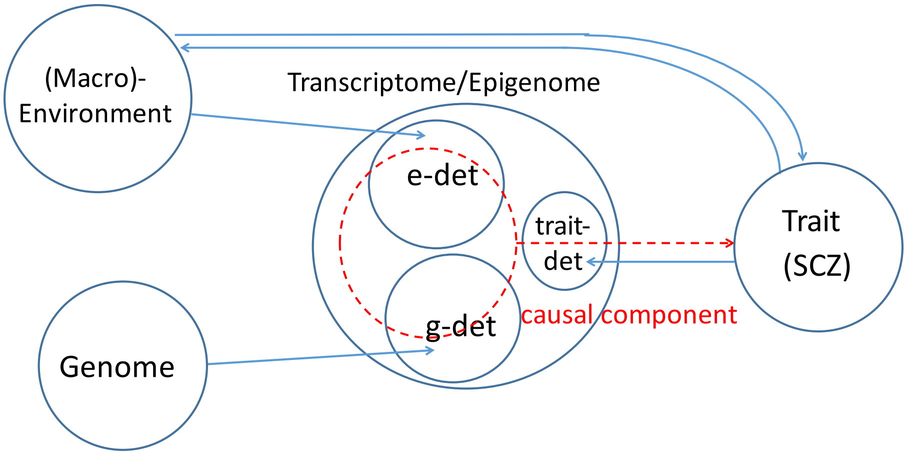


Fig. S8.2 Potential causal relationships between genome, transcriptome/epigenome, macro-environment and high-level traits A schematic of possible decomposition of variation in the indicated variables. Large circles represent total entropy of each variable, and smaller circles (e-det, g-det, trait-det) represent multivariate mutual information shared between variables linked by arrows (directionality represents causation). The red dotted circle and arrow represent causal influence of transcriptome/epigenome on the high-level trait, only part of which need intersect the g-det circle; hence, the trait variance explained by the transcriptome/epigenome is an upper-bound on the genetically determined trait variance. Only three-way intersections involving trait interactions are shown

S8.5 Enrichment analysis for prioritized modules and higher-order groupings

To provide interpretation of the DSPN model we develop a multilevel prioritization scheme, which, given a node of interest and a lower ‘projection layer’, defines positive and negative subsets of nodes on the projection layer which are most ‘important’ in influencing the value of the node of interest. In our analysis, we take the node of interest to be either a high-level trait (e.g. SCZ), or a hidden-layer node, and the projection layer to be an intermediate phenotype; we then use the prioritized subsets either to look for intermediate phenotypes prioritized for a given trait, or to functionally annotate hidden-layer nodes by looking for functionally enriched categories in the prioritized subsets.

In general, we assume we have a neural network with layers $L_l = \{n_{l,1}, n_{l,2}, \ldots n_{l,N_l}\}$, $l = 0 \ldots N_L$, with L_0 the lowest (input) layer and L_{N_L} the highest (output) layer. We fix a node of interest on layer m, $n^* \in L_m$, and a ‘branching factor’ B, which will determine the maximum size of the prioritized sets associated with n^*. Given these, we recursively define the positive and negative sets $S_{(l,+)}$ and $S_{(l,-)}$ associated with n^* for all $l \leq m$. We start by defining $S_{(m,+)} = \{n^*\}$ and $S_{(m,+)} = \{\}$. Then, for all $l < m$:

$$S_{(l,+)} = \left(\bigcup_{n \in S_{(l+1,+)}} B_{(n,+)} \right) \cup \left(\bigcup_{n \in S_{(l+1,-)}} B_{(n,-)} \right), \qquad (12)$$
$$S_{(l,-)} = \left(\bigcup_{n \in S_{(l+1,+)}} B_{(n,-)} \right) \cup \left(\bigcup_{n \in S_{(l+1,-)}} B_{(n,+)} \right),$$

where we define the sets $B_{(n,+)}, B_{(n,-)} \subset L_l$ for $n \in L_{l+1}$ as $B_{(n,+)} = \{n' \mid \text{rank}_n^+(n') \leq B\}$ and $B_{(n,-)} = \{n' \mid \text{rank}_n^-(n') \leq B\}$, where the function $\text{rank}_n^+(n')$ returns the rank of n' when the nodes of layer L are ranked in descending order by the network weights $w_{n,n'}$, and $\mathrm{rank}_n^-(n')$ returns the rank when the nodes are ranked in ascending order by the same weights. We note that $S_{(l,+)}$ and $S_{(l,-)}$ may contain common elements (i.e. nodes that contribute both positively and negatively to variation in a higher-level node).

To find prioritized modules for a given trait, we fix the 'projection layer' l to be the co-expression module sublayer in the DSPN (L_{1b-ii} in Fig. 6A), and find the sets $S_{(l,+)}$ and $S_{(l,-)}$ when n^* is set to the output trait node. We repeat this analysis for models trained on the 10 splits of the data for the given trait, generating 10 positive and negative projected sets. For module $n_{l,i}$, we then calculate the counts $c_{(i,+)} = \sum_t [n_{l,i} \in S_{(l,+)}^t]$, where $S_{(l,+)}^t$ is the positive projected set from the model trained on data split t, and $c_{(i,-)} = \sum_t [n_{l,i} \in S_{(l,-)}^t]$. For our final list of positive and negative prioritized modules we use $S^+ = \{n_{l,i} \mid c_{(i,+)} > \tau_{\mathrm{prioritize}}\}$ and $S^- = \{n_{l,i} \mid c_{(i,-)} > \tau_{\mathrm{prioritize}}\}$ respectively. The threshold $\tau_{\mathrm{prioritize}}$ is set such that, $p\left(c_{(i,+)} > \tau_{\mathrm{prioritize}} \mid B\right) < \alpha$ under a null distribution where the network weights are sampled from a standard normal distribution and the same branching factor B is used. We set $\alpha = 0.001$, and evaluate $\tau_{\mathrm{prioritize}}$ using 10,000 simulations. Setting $B = 4$, we found that this implied an estimate of $\tau_{\mathrm{prioritize}} = 3$, and generated ~ 30 positive and negative prioritized modules per trait (out of ~ 5000).

To annotate 'typical' ancestor nodes of module $n_{l,i}$ at layers $l+1$ and $l+2$ in the DSPN (hidden layers L_{2a} and L_{2b} respectively in Fig. 6A), for each data split we find nodes $n_{l+1,j}$ and $n_{l+2,k}$ such that $(n_{l,i}, n_{l+1,j}, n_{l+2,k})$ forms the 'best path' from module $n_{l,i}$ to the trait output node in the sense that it minimizes the score:

$$\mathrm{Sc} = \sum_{(l',i',j') \in \{(l,i,j),(l+1,j,k),(l+2,k,0)\}} \min\left(\mathrm{rank}_{n_{l'+1,j'}}^+\left(n_{l',i'}\right), \mathrm{rank}_{n_{l'+1,j'}}^-\left(n_{l',i'}\right)\right), \qquad (13)$$

across all 'positive' paths, meaning:

$$\prod_{(l',i',j')} (-1)^{\left[\mathrm{rank}_{n_{l'+1,j'}}^-\left(n_{l',i'}\right) \le B\right]} \cdot \left[\min\left(\mathrm{rank}_{n_{l'+1,j'}}^+\left(n_{l',i'}\right), \mathrm{rank}_{n_{l'+1,j'}}^-\left(n_{l',i'}\right)\right) \le B\right] = 1, \qquad (14)$$

and ties are broken arbitrarily (a similar annotation can be made for negative paths by placing -1 on the RHS of Eq. 13). Writing $n_{l+1,j}^t$, $n_{l+2,k}^t$ for the nodes on the best path from module $n_{l,i}$ in the model from data split t, we evaluate the counts for all modules, $c_{(i',+)} = \sum_t [n_{l,i'} \in S_{(l,+)}^t(n_{l+1,j}^t)]$ and $d_{(i',+)} = \sum_t [n_{l,i'} \in S_{(l,+)}^t(n_{l+2,k}^t)]$, where we write $S_{(l,+)}^t(n)$ for the positive projected set at level l for data split t when we set the node of interest $n^* = n$. We then evaluate $S_c^+ = \{n_{l,i'} \mid c_{(i',+)} > \tau_{\mathrm{prioritize}}\}$ and $S_d^+ = \{n_{l,i'} \mid d_{(i',+)} > \tau_{\mathrm{prioritize}}\}$ where $\tau_{\mathrm{prioritize}}$ is defined as above, and annotate a typical (positive) ancestor of $n_{l,i}$ at layer $l+1$ (respectively $l+2$) by finding the functional annotations enriched in the gene-set formed by taking the union of the co-expression modules in S_c^+ (respectively S_d^+).

We perform functional enrichment analysis using the R package 'clusterProfiler' (Yu et al., 2012) using KEGG pathway annotations, and setting the p-value and q-value cutoffs to 0.05 and 0.1 respectively. Further, we perform enrichment analysis for the cell-type marker genes corresponding to the cell-types used in our single-cell deconvolution analysis. Here, we threshold the marker gene expression matrix for each gene independently at its 0.75 quantile value to define a collection of subsets of marker genes for each cell-type. We test for enrichment of cell type markers using the hypergeometric test with a p-value cutoff of 0.1. Finally, we also compare the modules prioritized for our SCZ model using the above approach with those prioritized using a gradient-based approach, following (Simonyan et al., 2013) where the magnitude of the gradient of the response of a node of interest (in our case, the trait node responses across the training set) is use to prioritize salient input nodes (modules). We provide the results of this analysis in the enrichment analysis data file on the website (adult.psychencode.org), but found it to exhibit a strong bias towards prioritizing smaller modules, which may be due to the underestimation of the contribution of saturated nodes in gradient approaches (see Shrikumar et al., 2017, which attempts to circumvent these problems, but requires definition of a 'reference' state which is unclear in our model), causing us to prefer the prioritization scheme developed above, in which we did not observe such a bias.

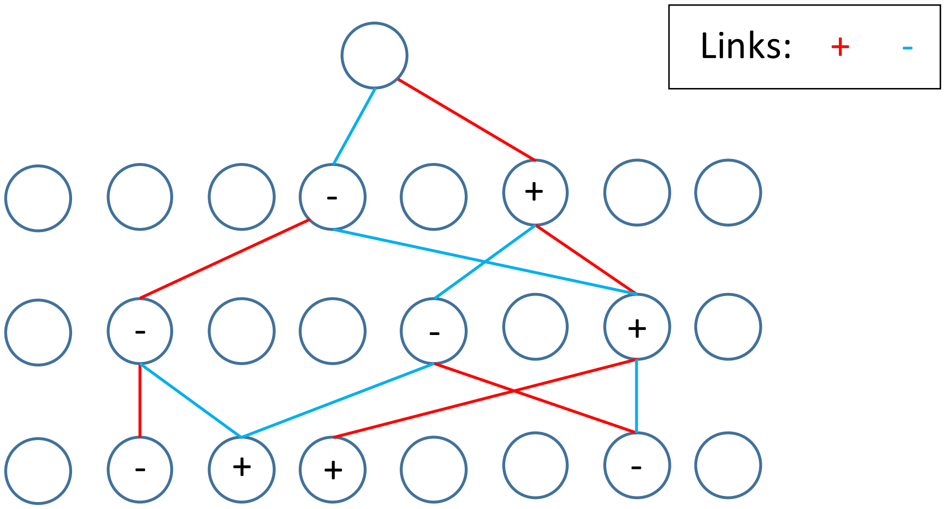


Fig. S8.3 Schematic representation of prioritization scheme for interpreting DSPN latent nodes and modules Circles represent nodes on three layers within the DPSN, along with a 'node of interest' on the upper layer. The prioritization shown uses a branching factor of 2, where red and blue links indicate the largest positive and negatively weighted edge respectively connected to each node from below. + and - signs represent the positive and negative prioritized sets for the node of interest at each of the lower levels, which are assigned based on the multiplication of signs along connecting paths (conflicting pathways would result in assignment to both positive and negative sets; not shown).

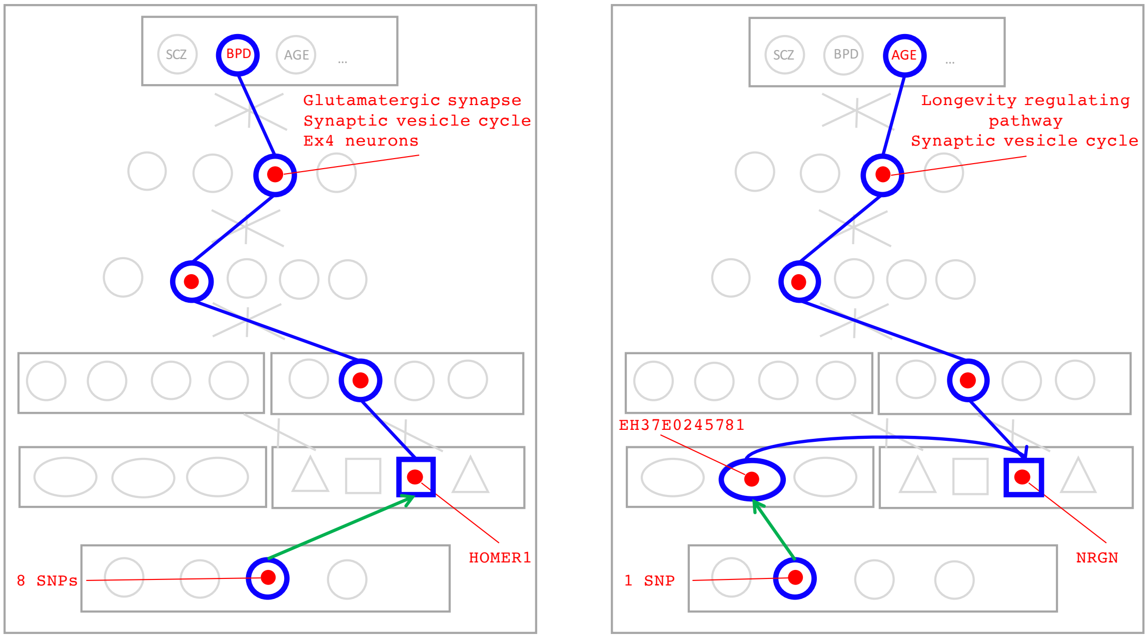


Fig. S8.4 Further DSPN traces for functional enrichment of prioritized modules in DSPN models Examples are shown of genes belonging to prioritized modules in BPD (left) and AGE (right) related DSPN models. HOMER1 has previously been associated with BPD, and NRGN was strongly associated with age in our differential expression analysis (as well as being a SCZ associated gene). An associated data file summarizing the functional and cell-type enrichments in the prioritized modules of all phenotypes can be found on the website (adult.psychencode.org).

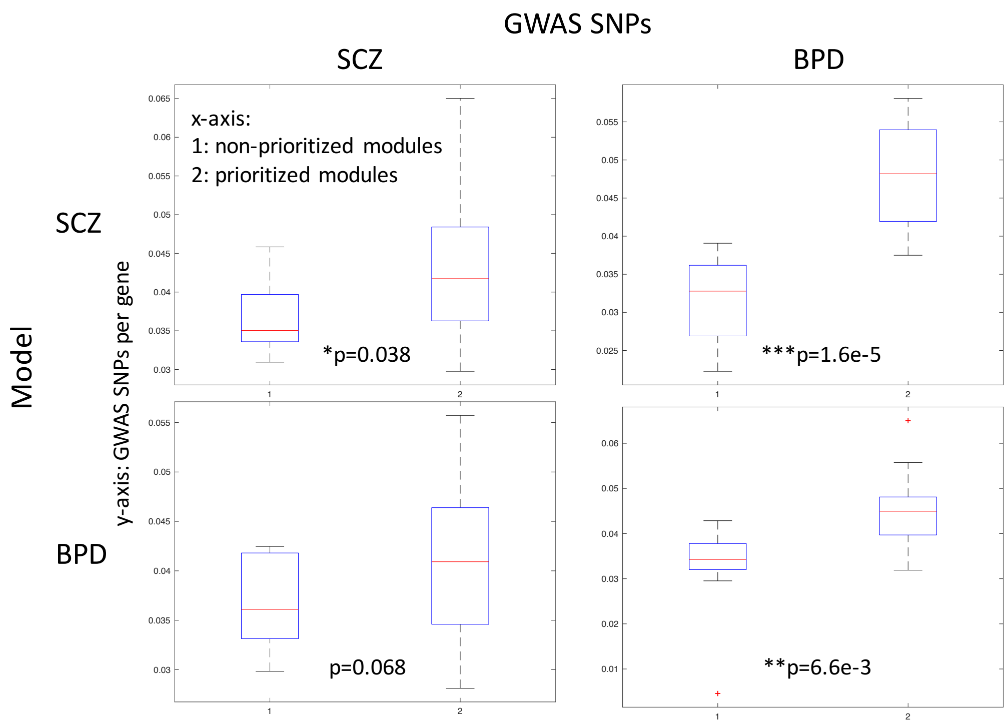


Fig. S8.5 Enrichment of GWAS SNPs in DSPN prioritized modules. Figure shows enrichment of GWAS SNPs associated with SCZ and BPD in the DSPN modules prioritized in the SCZ and BPD models. SNPs are linked with prioritized modules using all eQTLs associated with genes they contain. Enrichment is tested using a 1-tailed Mann-Whitney test for an increase in the number of GWAS SNPs per gene in prioritized versus non-prioritized modules. We observe enrichments for both disease modules with their respective GWAS SNPs, and also an enrichment of BPD GWAS SNPs in the SCZ modules, consistent with an overlap in disease etiology.

Method	SCZ	BPD	ASD	GEN	ETH	AGE
DSPN-mod (a)	62.3% (4.6%)	57.2% (2.6%)	70.0% (8.5%)	52.4%	70.0%	76.2%
DSPN-mod (b)	66.1% (7.8%)	66.1% (8.4%)	-	60.6%	75.7%	81.9%
cRBM (a)	67.1% (16.0%)	65.6% (7.6%)	56.7% (3.8%)	67.4%	85.7%	81.5%
DSPN-imput (a)	56.4% (1.2%)	61.7% (5.4%)	62.5% (2.6%)	-	-	-
DSPN-full (a)	67.9% (16.3%)	66.1% (30.0%)	68.3% (14.4%)	69.7%	92.7%	83.9%

Table S8.1 Performance of DSPN-mod and comparison of stopping criteria. Performance of DSPN-mod and other models are shown using (a) fixed and (b) variable early stopping thresholds, as described in the supplemental text (S8.2). Test accuracy is shown for all models along with corresponding liability scores in brackets averaged across 10-fold cross validation data splits. A fixed threshold only is used for the ASD model (due to small sample size); variable threshold settings for cRBM, DSPN-imput and DSPN-full models are as in Fig. 6D for all phenotypes except ASD.

S9. Resource website and raw data

S9.1 Resource website: http://adult.psychencode.org/

The website contains much supplementary information related to the project, including the raw and processed data files them. For convenience, we reproduce below some sections of the site and from the PsychENCODE Synapse website related to the descriptions of the data files.

S9.2 RNA-seq, ChIP-seq and genotype data (The text in this section up to Study 8 was directly adapted from the Psychencode/Synapse Website).

We processed data from 9 studies: UCLA-ASD, Yale-ASD, BrainGVEX, the The Lieber Institute for Brain Development (LIBD), GTEx, the CommonMind Consortium (CMC), the CMC's NIMH Human Brain Collection Core (CMC HBCC) and Bipseq a Bipolar cohorts. The detailed descriptions of PsychENCODE related 8 studies were listed below and may also be found on supplemental Table S2.1, as well as in the PsychENCODE Knowledge Portal (https://www.synapse.org/#!Synapse:syn4921369/wiki/390659).

Study 1 - BrainGVEX
RNA-seq: RNAseq data was generated from 427 postmortem prefrontal cortex from subjects with schizophrenia (n=95), bipolar disorder (n=73), and non-psychiatric controls (n=259), as part of the BrainGVEX study (Synapse accession doi:10.7303/syn4590909) within the PsychEncode Consortium (https://www.synapse.org/pec) (72). BrainGVEX study includes RNA samples collected as part of the "Array Collection", "Consortium Collection", "New Collection" and "Extra Collection" from the Stanley Medical Research Institute (SMRI). Array collection and Consortium collection were from superior frontal gyrus (BA9) whereas those labelled EXTRA or NEW were from the middle frontal gyrus (BA46). Another 184 controls were obtained as fresh-frozen brain tissue from the Banner Sun Health Research Institute (BSHRI). All BSHRI samples were from frontal cortex. RNA were extracted from BSHRI samples by first homogenizing 20-50 mg of tissue in QIAzol (Qiagen) using the Lysin Matrix D and FastPrep®-24 system (MPBiomedicals). Total RNA were then isolated using the miRNeasy Kit (Qiagen) according to manufacturer's instructions. RNA integrity was assessed with Agilent Technologies RNA 600 nano kit. Samples with RNA Integrity Number (RIN) lower than 5.5 were excluded from the study. RNA sequencing libraries were prepared using TruSeq Stranded Total RNA sample prep kit with RiboZero Gold HMR (Illumina). Libraries were multiplexed (3 per lane) for paired-end 100 bp sequencing on Illumina HiSeq2000 with read depth >70 million reads on average.

Genotyping: DNA genotyping were done using two different platforms. 144 samples (SMRI Consortium and Array Collections) were genotyped using the Affymetrix GeneChip Mapping 5.0K Array. Genotypes were called with the BRLMM-p algorithm (Affymetrix) with all arrays simultaneously (Zhang et al., 2010). The rest of samples (SMRI New and Extra Collection, and BSHRI Collection) were genotyped with the Human PsychChip, which is a custom version of the Illumina Infinium CoreExome-24 v1.1 BeadChip (#WG-331-1111) supplemented with content derived from GWASs and DNA sequencing studies of multiple psychiatric disorders by the Psychiatric Genomics Consortium (PGC). Genotypes were called using Illumina's GenomeStudio software, Birdseed and Zcall, as described (Code found at: https://github.com/Nealelab/ricopili/blob/master/rp_bin/mergecall_10) (Pedersen et al., 2018).

GenomeStudio and Birdseed were used separately to initially call variants in 288 individuals. Accepted variants had a call frequency greater than 97% and a Hardy-Weinberg Equilibrium (HWE) p-value > 1 x 10^{-6}. 24 of the 288 individuals were immediately excluded because they were missing calls for >5% of genotyped SNPs, when either caller was used. Birdseed and GenomeStudio variant calls were then merged by consensus. If both programs returned a different result for a single variant, the final call for that variant was set to "missing." When a call was made with only one of the two programs, that successful call was deemed the consensus.

The resulting merged consensus data was filtered again according to the same call frequency, sample missingness, MAF and HWE criteria described above. Finally, valid rare variant calls were refined using Zcall. Meaning, genotype calls for variants with MAF < 0.01 in the merged and filtered dataset were replaced with zCall results when, in zCall, their HWE p-values > 1 x 10^{-6}, missingness rates were below 3% and MAF < 0.05. Note that zCall only refines GenomeStudio calls, so zCall results are independent of Birdseed calls. Ultimately 577,643 variants were called, 242,272 being rare.

Study 2- BrainSpan

RNA-seq: RNA was extracted using RNeasy Plus Mini Kit (Qiagen) for mRNA. Either approximately 30 mg of pulverized tissue (12 PCW – 40 Y specimens) or entire amount of dissected brain piece (8 – 9 PCW, smaller than 30 mg) was processed. Tissue was pulverized with liquid nitrogen in a chilled mortar and pestle and transferred to a chilled safe-lock microcentrifuge tube (Eppendorf). Per tissue mass, equal mass of chilled stainless steel beads (Next Advance, cat# SSB14B) along with two volumes of lysis buffer were added. Tissue was homogenized for 1 min in Bullet Blender (Next Advance # SSB14B) at speed 6 and incubated at 37°C for 5 min. Lysis buffer up to 0.6 ml was again added, tissue homogenized for 1 min and incubated at 37°C for 1 min. Extraction was further carried out according to manufacturer's protocol. Genomic DNA was removed by a proprietary column provided in RNeasy Plus Mini Kit (Qiagen) or by DNase treatment using TURBO DNA-free Kit (Ambion/ Life technologies). 260:A280 ratio and RNA Integrity Number (RIN) were determined for each sample with NanoDrop (Thermo Scientific) and Agilent 2100 Bioanalyzer system, respectively.

The mRNA-sequencing (mRNA-seq) Sample preparation Kit (Illumina) was used to prepare cDNA libraries per manufacturer instructions with some modifications. Briefly, polyA RNA was purified from 1 to 5 µg of total RNA using Oligo (dT) beads. Quaint-IT RiboGreen RNA Assay Kit (Invitrogen) was used to quantitate purified mRNA with the NanoDrop 3300. Following mRNA quantitation, 2.5 µl spike-in master mixes, containing five different types of RNA molecules at varying amounts (2.5 × 10-7 to 2.5 × 10-14 mol), were added per 100 ng of mRNA. Spike-in RNAs were synthesized by the External RNA Control Consortium (ERCC) by in vitro transcription of de novo DNA sequences or DNA derived from B. subtilis or the deep-sea vent microbe M. jannaschii and were a generous gift of Dr. Mark Salit at The National Institute of Standards and Technology (NIST). Each sample was tagged by adding two spike-in RNAs unique to the region from which the sample was taken. Besides, three common spike-in RNAs with gradient concentrations were added to each sample, aiming at the assessment of sequencing quality. Spike-in sequences are available at http://archive.gersteinlab.org/proj/brainseq/spike_in/spike_in.fa. The mixture of mRNA and spike-in RNAs was subjected to fragmentation, reverse transcription, end repair, 3' end adenylation, and adapter ligation to generate libraries of short cDNA molecules, followed by PCR amplification. The PCR enriched product was assessed for its size distribution and concentration using Bioanalyzer DNA 1000 Kit.

Single Cell RNA-seq: Neurotypical control tissue samples used in this study were obtained from various sources. Tissue was collected after obtaining parental or next of kin consent and with approval by the institutional review boards at the Yale University School of Medicine, and at each institution from which tissue specimens were obtained. Tissue was handled in accordance with ethical guidelines and regulations for the research use of human brain tissue set forth by the NIH and the WMA Declaration of Helsinki. Fresh tissue samples were received in Hibernate E solution. Tissues were then dissected depending on their ages. Embryonic samples were dissected under microscope and the whole pallial wall was sampled. Samples from later stages were placed on ventral side up onto a chilled aluminum plate (1 cm thick) on ice. The brainstem and cerebellum were removed from the cerebrum by making a transverse cut at the junction between the diencephalon and midbrain. Next, the cerebrum was divided into left and right hemispheres by cutting along the midline using a Tissue-Tek Accu-Edge trimming blade, 260 mm. The regions of interest were dissected using a scalpel blade and immediately processed. The sampled area corresponds to dorsolateral prefrontal cortex (DLPFC) and it was sampled from the middle third of the dorsolateral surface of the anterior third of the cerebral hemisphere. These specimens contained the marginal zone, cortical plate, and part of the underlying subplate. Dissected tissue was dissociated to cell suspension using Papain-Protease-DNase (PPD) and gentleMACS dissociator (Miltenyi Biotec). Cell suspension was then processed on Fluidigm C1 machine to capture single cells, according to manufacturer's protocol. RNA extraction from each single cell was carried out on Fluidigm C1 machine, according to manufacturer's protocol.

Genotype data was not used in this study due to the small adult sample size.

Study 3 - CommonMind
Full details of the CommonMind study have been published (*19*). Data is available through the Sage Bionetworks Synapse system (https://www.synapse.org/cmc; doi:10.7303/syn2759792). Samples were acquired through brain banks at three institutions:The Mount Sinai NIH Brain Bank and Tissue Repository, University of Pennsylvania Brain Bank of Psychiatric illnesses and Alzheimer's Disease Core Center, and the University of Pittsburgh NIH NeuroBioBank Brain and Tissue Repository. Details about brain banks, inclusion/exclusion criteria, and sample collection and processing are described here: https://www.synapse.org/#!Synapse:syn2759792/wiki/71104

RNA-seq: RNA-seq data from 613 total human postmortem dorsolateral prefrontal cortex (DLPFC) brain samples were obtained from 603 subjects with schizophrenia (n=263), bipolar disorder (n=47), affective disorder (8), and neurotypical controls (n=285), where 10 neurotypical controls were sequenced as biological replicates). Total RNA was extracted from 50 mg of homogenized DLPFC brain tissue using RNeasy kit. Samples with RIN < 5.5 (n=51) were excluded. The remaining samples had a mean RIN of 7.7. RNAseq library preparation was performed using ribosomal RNA depletion, with the Ribozero Magnetic Gold Kit. Samples were barcoded, multiplexed (n=10/lane), and sequenced across two lanes as 100 bp paired end sequencing on the Illumina HiSeq 2500 with an average of 85 million reads. Data is provided for those samples that passed all of the following QC filters: samples were required to have had a minimum of 50 million total reads and less than 5% rRNA alignment.

ChIP-seq: ChIP-seq data of H3K27ac and H3K4me3 of NeuN+ cells were generated on a subset of the CommonMind Samples in PsychENCODE Epidiff study. Full details of the data generation protocal could be found in one of the published methods paper (Kundakovic M et al., 2017) We used H3K27ac from Dorsolateral Prefrontal Cortex of 117 neurotypical controls and 109 schizophrenia individuals.

Genotyping: DNA was isolated from approximately 10 mg dry homogenized tissue coming from the same dissected samples as the RNA isolation using the Qiagen DNeasy Blood and Tissue Kit according to manufacturer's protocol. Genotyping was performed using the Illumina Infinium HumanOmniExpressExome platform (Catalog #: WG-351-2301). All data were checked for discordance between nominal and genetically-inferred sex using Plink software to calculate the mean homozygosity rate across X-chromosome markers and to evaluate the presence or absence of Y-chromosome markers. In addition, pairwise comparison of samples across all genotypes was done to identify potentially duplicate samples (genotypes > 99% concordant) or related individuals using Plink.

Study 4 - Yale-ASD
RNA-seq: Total RNA was extracted using mirVana kit (Ambion) with some modifications to the manufacturer's protocol. Approximately 60 mg of tissue was pulverized with liquid nitrogen in a prechilled mortar and pestle and transferred to a chilled safe-lock microcentrifuge tube (Eppendorf). Per tissue mass, equal mass of chilled stainless steel beads (Next Advance, catalog # SSB14B) along with one volume of lysis/binding buffer were added. Tissue was homogenized for 1 min in Bullet Blender (Next Advance) and incubated at 37°C for 1 min. Another nine volumes of the lysis/binding buffer were added,

homogenized for 1 min, and incubated at 37°C for 2 min. One-tenth volume of miRNA Homogenate Additive was added and extraction was carried out according to the manufacturer's protocol. RNA was treated with DNase using TURBO DNA-free Kit (Ambion/ Life Technologies) and RNA integrity was measured using Agilent 2200 TapeStation System. Barcoded libraries for RNA-seq were prepared with 5ng of RNA using TruSeq Stranded Total RNA with Ribo-Zero Gold kit (Illumina) per manufacturer's protocol. Paired-end sequencing (100bp x 2) was performed on HiSeq 2000 sequencers (Illumina) at Yale Center for Genome Analysis.

Genotype data is not available yet for this study.

Study 5 - UCLA-ASD
Full details of the UCLA-ASD study have been published (Parikshak, et al., 2016).

RNA-seq: RNA-seq data for replication was generated from 251 postmortem cortex brain samples from subjects with ASD and non-psychiatric controls, across frontal cortex (BA9/46), temporal cortex (BA41/42/22), and cerebellum.

Brain samples were obtained from the Harvard Brain Bank as part of the Autism Tissue Project (ATP). An ASD diagnosis was confirmed by the Autism Diagnostic Interview-Revised (ADIR) in 48 of the subjects. In the remaining two subjects, diagnosis was supported by clinical history. Frozen brain regions were dissected on dry ice in a dehydrated dissection chamber to reduce degradation effects from sample thawing or humidity. Approximately 50-100mg of tissue across the cortical region of interest was isolated from each sample using the miRNeasy kit with no modifications (Qiagen). For each RNA sample, RNA quality was quantified using the RNA Integrity Number (RIN) on an Agilent Bioanalyzer. Strand-specific, rRNA-depleted RNAseq libraries were prepared using TruSeq Stranded Total RNA sample prep kit with RiboZero Gold (Illumnia) kits. Libraries were randomly pooled to multiplex 24 samples per lane using Illumina TruSeq barcodes. Each lane was sequenced five times on an Illumina HiSeq 2500 instrument using high output mode with standard chemistry and protocols for 50 bp paired-end reads to achieve a target depth of 70 million reads.

ChIP-seq: For each ChIP-seq experiment approximatively 100mg of frozen brain tissue per sample was aliquoted and thawed on ice in 1ml PBS buffer. Tissue was then homogenized using a manual glass douncer with 7-15 slow strokes on ice. The cell suspension was filtered with a 40uM cell strainer (Falcon) by spinning at 2000rpm for 1 minute at 4C in a swing bucket centrifuge (Eppendorf Centrifuge 5810R). Pellets were then washed twice with cold PBS, crosslinked with 1% formaldehyde for 15 minutes at room temperature and excess formaldehyde quenched by addition of glycine (0.625M). Cells were lysed with FA and nuclei were collected and re-suspended in 300 µl SDS lysis buffer (1% SDS, 1% Triton X 100, 2 mM EDTA, 50 mM Hepes-KOH [pH 7.5], 0.1% Na dodecyl-sulfate, Roche 1X Complete protease inhibitor). Nuclei were lysed for 15 minutes, after which sonication was used to fragment chromatin to an average size of 200–500 bp (Bioruptor Next gen, Diagenode). Protein-DNA complexes were immuno-precipitated using 3 µg of H3K27acetyl antibody of the same lot for all ChIP experiments (catalogue number 39133; Actif motif) coupled to 50µl protein G Dynal beads (Invitrogen) overnight. The beads were washed and protein-DNA complexes were eluted with 150 µl of elution buffer (1% SDS, 10 mM EDTA, 50 mM Tris.HCl [pH 8]), followed by protease treatment and de-crosslinking at 65°C overnight. After phenol/chloroform extraction, DNA was purified by ethanol precipitation. Library preparation was performed as in (Quail et al., 2008). After 15 cycles of PCR using indexing primers, libraries were size selected for 300-500 bp on low melting agarose gel and 4 libraries were pooled and sequenced in one lane of 2 x 100bp using the same Illumina HiSeq 2000 with V3 reagents.

Genotyping: Genotyping was performed using Illumina Omni 2.5 arrays.

Study 6 - BipSeq
RNAseq: same as below **Study 8**

Genotyping: same as below **Study 8**

Study 7 - CMC_HBCC
Brain specimens for the CMC_HBCC study were obtained from the the NIMH Human Brain Collection Core (HBCC) (https://www.nimh.nih.gov/labs-at-nimh/research-areas/research-support-services/hbcc/human-brain-collection-core-hbcc.shtml) under protocols approved by the CNS IRB (NCT00001260), with the permission of the next-of-kin through the Offices of the Chief Medical Examiners in the District of Columbia, Northern Virginia and Central Virginia. All specimens were characterized neuropathologically, clinically and toxicologically. A clinical diagnosis was obtained through family interviews and review of medical records by two psychiatrists based on DSMIV criteria. Non-psychiatric controls were defined as having no history of a psychiatric condition or substance use disorder.

RNAseq: Samples were dissected at the NIMH Human Brain Collection Core and shipped to Ichan School of Medicine - Mt Sinai (ISMMS) for sample preparation and RNA-sequencing. Samples for the study were dissected from either the left or right hemisphere of fresh frozen coronal slabs cut at autopsy from the dorsolateral prefrontal cortex. Total RNA from 468 HBCC samples was isolated from approximately 100 mg homogenized tissue from each sample by TRIzol/chloroform extraction and purification with the Qiagen RNeasy kit (Cat#74106) according to manufacturer's protocol. Samples were processed in randomized batches of 12. The order of extraction was assigned randomly with respect to diagnosis and all other sample characteristics. The mean total RNA yield was 24.2 ug. The RNA Integrity Number (RIN) was determined by fractionating RNA samples on the 4200 Agilent TapeStation System. 69 samples with RIN <5.5 were excluded from the study. An additional 12 samples were removed post sequencing due to evidence of sample swap or contamination, resulting in a final dataset of 387 samples with a mean RIN of 7.5 and a mean ratio of 260/280 of 2.0. (Bipolar Disorder n=70, Schizophrenia n=97, neurotypical controls n=220) RNA sequencing raw and quantified expression data is provided for 387 samples consisting of data from 387 unique individuals. Data was generated, QCed, processed and quantified as follows: All samples submitted to the New York Genome Center for RNAseq were prepared for sequencing in randomized batches of 94. The sequencing libraries were prepared using the KAPA Stranded RNAseq Kit with RiboErase (KAPA Biosystems). rRNA was depleted from 1ug of RNA using the KAPA RiboErase protocol that is integrated into the KAPA Stranded RNAseq Kit. The insert size and DNA concentration of the sequencing library was determined on Fragment Analyzer Automated CE System (Advanced Analytical) and Quant-iT PicoGreen (ThermoFisher) respectively. A pool of 10 barcoded libraries were layered on a random selection of two of the eight lanes of the Illumina flow cell at appropriate concentration and bridge amplified to ~ 250 million raw clusters. One-hundred base pair paired end reads were obtained on a HiSeq 2500.

Genotyping: Genotyping was done on the Illumina_1M, Illumina_h650, and Illumina_Omni5 platform.

Study 8 - LIBD_szControl + BipSeq
RNAseq: Post-mortem tissue homogenates of dorsolateral prefrontal cortex grey matter (DLPFC) approximating BA46/9 in postnatal samples and the corresponding region of PFC in fetal samples were obtained from all subjects. Total RNA was extracted from ~100 mg of tissue using the RNeasy kit (Qiagen) according to the manufacturer's protocol. The poly-A containing RNA molecules were purified from 1 µg DNAse treated total RNA and sequencing libraries were constructed using the Illumina TruSeq© RNA Sample Preparation v2 kit. Sequencing indices/barcodes were inserted into Illumina adapters allowing samples to be multiplexed in across lanes in each flow cell. These products were then purified and enriched with PCR to create the final cDNA library for high throughput sequencing using an Illumina HiSeq 2000 with paired end 2x100bp reads. More details are available in: https://www.biorxiv.org/content/early/2017/11/22/124321

Genotyping: SNP genotyping with HumanHap650Y_V3, Human 1M-Duo_V3, and Omni5 BeadChips (Illumina, San Diego, CA) was carried out according to the manufacturer's instructions with DNA extracted from cerebellar tissue. Genotype data were processed and normalized with the crlmm R/Bioconductor package separately by platform.

There is an overlap in the donors and samples used for CMC_HBCC and LIBD_scControl and BipSeq came from, because they originate from the same brain bank (the NIMH human brain collection core). There is therefore a set of biological replicates from the same brain region where the samples have been processed separately. The same individual ID has been used on all 3 studies. The CMC data also has a set of 10 biological replicates (all controls). The individual IDs are the same (starting with CMC_..). We included all samples (including replicates) and accounted for them using random effect mixed model.

An initial quality control step was taken in which all datasets were first pre-processed to remove outliers using a hierarchical clustering based global outlier detection. Samples from UCLA were subdivided into three different brain regions (vermis, Brodmann area 9, and Brodmann area 41).

The gene expression data from these 9 centers were merged into one gene expression matrix, and subsequently normalized using the protocol detailed by GTEx (GTEx Consortium, 2017).

S10. References in supplement

Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Gephart MG, Barres BA, Quake SR. A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences. 2015 Jun 9;112(23):7285-90.

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze S, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. *Nature Genetics* 2016 (48):1284–1287

Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. A complete tool set for molecular QTL discovery and analysis. Nature communications. 2017 May 18;8:15452.

Demontis, D., R. K. Walters, J. Martin, M. Mattheisen, T. D. Als, E. Agerbo, R. Belliveau, J. Bybjerg-Grauholm, M. Bækved-Hansen, F. Cerrato, K. Chambert, C. Churchhouse, A. Dumont, N. Eriksson, M. Gandal, J. Goldstein, J. Grove, C. S. Hansen, M. Hauberg, M. Hollegaard, D. P. Howrigan, H. Huang, J. Maller, A. R. Martin, J. Moran, J. Pallesen, D. S. Palmer, C. B. Pedersen, M. G. Pedersen, T. Poterba, J. B. Poulsen, S. Ripke, E. B. Robinson, F. K. Satterstrom, C. Stevens, P. Turley, H. Won, O. A. Andreassen, C. Burton, D. Boomsma, B. Cormand, S. Dalsgaard, B. Franke, J. Gelernter, D. Geschwind, H. Hakonarson, J. Haavik, H. Kranzler, J. Kuntsi, K. Langley, K.-P. Lesch, C. Middeldorp, A. Reif, L. A. Rohde, P. Roussos, R. Schachar, P. Sklar, E. Sonuga-Barke, P. F. Sullivan, A. Thapar, J. Tung, I. Waldman, M. Nordentoft, D. M. Hougaard, T. Werge, O. Mors, P. B. Mortensen, M. J. Daly, S. V. Faraone, A. D. Børglum and B. M. Neale (2017). "Discovery Of The First Genome-Wide Significant Risk Loci For ADHD." bioRxiv.

Falconer, D.S. and MacKay T.F.C. (1996) Introduction to Quantitative Genetics, Ed 4. Longmans Green, Harlow, Essex, UK.

Finucane, H. K., B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P. R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, S. Ripke, F. R. Day, C. ReproGen, C. Schizophrenia Working Group of the Psychiatric Genomics, R. Consortium, S. Purcell, E. Stahl, S. Lindstrom, J. R. Perry, Y. Okada, S. Raychaudhuri, M. J. Daly, N. Patterson, B. M. Neale and A. L. Price (2015). "Partitioning heritability by functional annotation using genome-wide association summary statistics." Nat Genet 47(11): 1228-1235.

Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, Shah HR, Klei LL. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nature neuroscience. 2016 Nov;19(11):1442.

Gandal, M.J. et al. Dysregulation of cortical splicing, isoform and noncoding gene regulatory networks in ASD, schizophrenia, and bipolar disorder. *(Submitted)*.

Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS genetics. 2014 May 15;10(5):e1004383.

Grant C.E., Bailey T.L., Noble W.S. FIMO: scanning for occurrences of a given motif (2011). *Bioinformatics* 27 (7): 1017–1018.

Grove, J., S. Ripke, T. D. Als, M. Mattheisen, R. Walters, H. Won, J. Pallesen, E. Agerbo, O. A. Andreassen, R. Anney, R. Belliveau, F. Bettella, J. D. Buxbaum, J. Bybjerg-Grauholm, M. Bækved-Hansen, F. Cerrato, K. Chambert, J. H. Christensen, C. Churchhouse, K. Dellenvall, D. Demontis, S. De Rubeis, B. Devlin, S. Djurovic, A. Dumont, J. Goldstein, C. S. Hansen, M. E. Hauberg, M. V. Hollegaard, S. Hope, D. P. Howrigan, H. Huang, C. Hultman, L. Klei, J. Maller, J. Martin, A. R. Martin, J. Moran, M. Nyegaard, T. Nærland, D. S. Palmer, A. Palotie, C. B. Pedersen, M. G. Pedersen, T. Poterba, J. B. Poulsen, B. St Pourcain, P. Qvist, K. Rehnström, A. Reichenberg, J. Reichert, E. B. Robinson, K. Roeder, P. Roussos, E. Saemundsen, S. Sandin, F. K. Satterstrom, G. D. Smith, H. Stefansson, K. Stefansson, C. Steinberg, C. Stevens, P. F. Sullivan, P. Turley, G. B.

Walters, X. Xu, D. Geschwind, M. Nordentoft, D. M. Hougaard, T. Werge, O. Mors, P. B. Mortensen, B. M. Neale, M. J. Daly, A. D. Børglum and a. R. Team (2017). "Common risk variants identified in autism spectrum disorder." bioRxiv.

GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017 Oct;550(7675):204.

Hill WG, Mackay TF. DS Falconer and Introduction to quantitative genetics. Genetics. 2004 Aug 1;167(4):1529-36.

Hinton GE. A practical guide to training restricted Boltzmann machines. In Neural networks: Tricks of the trade 2012 (pp. 599-619). Springer, Berlin, Heidelberg.

Howard, D. M., M. J. Adams, M. Shirali, T.-K. Clarke, R. E. Marioni, G. Davies, J. R. I. Coleman, C. Alloza, X. Shen, M. C. Barbu, E. M. Wigmore, S. Hagenaars, C. M. Lewis, D. J. Smith, P. F. Sullivan, C. S. Haley, G. Breen, I. J. Deary and A. M. McIntosh (2017). "Genome-wide association study of depression phenotypes in UK Biobank (n = 322,580) identifies the enrichment of variants in excitatory synaptic pathways." bioRxiv.

International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009 Aug;460(7256):748.

Jaffe AE, Gao Y, Deep-Soboslay A, Tao R, Hyde TM, Weinberger DR, Kleinman JE. Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. Nature neuroscience. 2016 Jan;19(1):40.

Johnstone IM, Lu AY. On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association. 2009 Jun 1;104(486):682-93.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, Guennel T. Spatio-temporal transcriptome of the human brain. Nature. 2011 Oct;478(7370):483.

Koller, D. and Friedmann, N.. Probabilistic Graphical Models, 2009.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb;518(7539):317.

Kundakovic M, Jiang Y, Kavanagh DH, Dincer A, Brown L, Pothula V, Zharovsky E, Park R, Jacobov R, Magro I, Kassim B, Wiseman J, Dang K, Sieberts SK, Roussos P, Fromer M, Harris B, Lipska BK, Peters MA, Sklar P, Akbarian S. Practical Guidelines for High-Resolution Epigenomic Profiling of Nucleosomal Histones in Postmortem Human Brain Tissue. Biol Psychiatry. 2017;81(2):162-70. doi: 10.1016/j.biopsych.2016.03.1048.

Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, Wildberg A, Gao D, Fung HL, Chen S, Vijayaraghavan R. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 2016 Jun 24;352(6293):1586-90.

Lambert, J. C., C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, A. L. DeStafano, J. C. Bis, G. W. Beecham, B. Grenier-Boley, G. Russo, T. A. Thorton-Wells, N. Jones, A. V. Smith, V. Chouraki, C. Thomas, M. A. Ikram, D. Zelenika, B. N. Vardarajan, Y. Kamatani, C. F. Lin, A. Gerrish, H. Schmidt, B. Kunkle, M. L. Dunstan, A. Ruiz, M. T. Bihoreau, S. H. Choi, C. Reitz, F. Pasquier, C. Cruchaga, D. Craig, N. Amin, C. Berr, O. L. Lopez, P. L. De Jager, V. Deramecourt, J. A. Johnston, D. Evans, S. Lovestone, L. Letenneur, F. J. Moron, D. C. Rubinsztein, G. Eiriksdottir, K. Sleegers, A. M. Goate, N. Fievet, M. W. Huentelman, M. Gill, K. Brown, M. I. Kamboh, L. Keller, P. Barberger-Gateau, B. McGuiness, E. B. Larson, R. Green, A. J. Myers, C. Dufouil, S. Todd, D. Wallon, S. Love, E. Rogaeva, J. Gallacher, P. St George-Hyslop, J. Clarimon, A. Lleo, A. Bayer, D. W. Tsuang, L. Yu, M. Tsolaki, P. Bossu, G. Spalletta, P. Proitsi, J. Collinge, S. Sorbi, F. Sanchez-Garcia, N. C. Fox, J. Hardy, M. C. Deniz Naranjo, P. Bosco, R. Clarke, C. Brayne, D. Galimberti, M. Mancuso, F. Matthews, I. European Alzheimer's Disease, Genetic, D. Environmental Risk in Alzheimer's, C. Alzheimer's Disease Genetic, H. Cohorts for, E. Aging Research in Genomic, S. Moebus, P. Mecocci, M. Del Zompo, W. Maier, H. Hampel, A. Pilotto, M. Bullido, F. Panza, P. Caffarra, B. Nacmias, J. R. Gilbert, M. Mayhaus, L. Lannefelt, H. Hakonarson, S. Pichler, M. M. Carrasquillo, M. Ingelsson, D.

Beekly, V. Alvarez, F. Zou, O. Valladares, S. G. Younkin, E. Coto, K. L. Hamilton-Nelson, W. Gu, C. Razquin, P. Pastor, I. Mateo, M. J. Owen, K. M. Faber, P. V. Jonsson, O. Combarros, M. C. O'Donovan, L. B. Cantwell, H. Soininen, D. Blacker, S. Mead, T. H. Mosley, Jr., D. A. Bennett, T. B. Harris, L. Fratiglioni, C. Holmes, R. F. de Bruijn, P. Passmore, T. J. Montine, K. Bettens, J. I. Rotter, A. Brice, K. Morgan, T. M. Foroud, W. A. Kukull, D. Hannequin, J. F. Powell, M. A. Nalls, K. Ritchie, K. L. Lunetta, J. S. Kauwe, E. Boerwinkle, M. Riemenschneider, M. Boada, M. Hiltuenen, E. R. Martin, R. Schmidt, D. Rujescu, L. S. Wang, J. F. Dartigues, R. Mayeux, C. Tzourio, A. Hofman, M. M. Nothen, C. Graff, B. M. Psaty, L. Jones, J. L. Haines, P. A. Holmans, M. Lathrop, M. A. Pericak-Vance, L. J. Launer, L. A. Farrer, C. M. van Duijn, C. Van Broeckhoven, V. Moskvina, S. Seshadri, J. Williams, G. D. Schellenberg and P. Amouyel (2013). "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease." Nat Genet **45**(12): 1452-1458.

Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2007 Nov 16;24(5):719-20.

Li, M. et al., Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risk. *(Submitted)*.

Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJ, Kong SL, Chua C, Hon LK, Tan WS, Wong M. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nature genetics. 2017 May;49(5):708.

Liu, J. Z., S. van Sommeren, H. Huang, S. C. Ng, R. Alberts, A. Takahashi, S. Ripke, J. C. Lee, L. Jostins, T. Shah, S. Abedian, J. H. Cheon, J. Cho, N. E. Dayani, L. Franke, Y. Fuyuno, A. Hart, R. C. Juyal, G. Juyal, W. H. Kim, A. P. Morris, H. Poustchi, W. G. Newman, V. Midha, T. R. Orchard, H. Vahedi, A. Sood, J. Y. Sung, R. Malekzadeh, H. J. Westra, K. Yamazaki, S. K. Yang, C. International Multiple Sclerosis Genetics, I. B. D. G. C. International, J. C. Barrett, B. Z. Alizadeh, M. Parkes, T. Bk, M. J. Daly, M. Kubo, C. A. Anderson and R. K. Weersma (2015). "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations." Nat Genet 47(9): 979-986.

Maaten LV, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(Nov):2579-605.

Marshall, CR., et al. "Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects." Nature genetics 49.1 (2017): 27.

McCarthy DJ, Campbell KR, Lun AT, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017 Jan 14;33(8):1179-86.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology. 2010 May;28(5):495.

Mnih V, Larochelle H, Hinton GE. Conditional restricted boltzmann machines for structured output prediction. arXiv preprint arXiv:1202.3748. 2012 Feb 14.

Morris, A. P., B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segre, V. Steinthorsdottir, R. J. Strawbridge, H. Khan, H. Grallert, A. Mahajan, I. Prokopenko, H. M. Kang, C. Dina, T. Esko, R. M. Fraser, S. Kanoni, A. Kumar, V. Lagou, C. Langenberg, J. Luan, C. M. Lindgren, M. Muller-Nurasyid, S. Pechlivanis, N. W. Rayner, L. J. Scott, S. Wiltshire, L. Yengo, L. Kinnunen, E. J. Rossin, S. Raychaudhuri, A. D. Johnson, A. S. Dimas, R. J. Loos, S. Vedantam, H. Chen, J. C. Florez, C. Fox, C. T. Liu, D. Rybin, D. J. Couper, W. H. Kao, M. Li, M. C. Cornelis, P. Kraft, Q. Sun, R. M. van Dam, H. M. Stringham, P. S. Chines, K. Fischer, P. Fontanillas, O. L. Holmen, S. E. Hunt, A. U. Jackson, A. Kong, R. Lawrence, J. Meyer, J. R. Perry, C. G. Platou, S. Potter, E. Rehnberg, N. Robertson, S. Sivapalaratnam, A. Stancakova, K. Stirrups, G. Thorleifsson, E. Tikkanen, A. R. Wood, P. Almgren, M. Atalay, R. Benediktsson, L. L. Bonnycastle, N. Burtt, J. Carey, G. Charpentier, A. T. Crenshaw, A. S. Doney, M. Dorkhan, S. Edkins, V. Emilsson, E. Eury, T. Forsen, K. Gertow, B. Gigante, G. B. Grant, C. J. Groves, C. Guiducci, C. Herder, A. B. Hreidarsson, J. Hui, A. James, A. Jonsson, W. Rathmann, N. Klopp, J. Kravic, K. Krjutskov, C. Langford, K. Leander, E. Lindholm, S. Lobbens, S. Mannisto, G. Mirza, T. W. Muhleisen, B. Musk, M. Parkin, L. Rallidis, J. Saramies, B. Sennblad, S. Shah, G. Sigurethsson, A. Silveira, G. Steinbach, B. Thorand, J. Trakalo, F. Veglia, R. Wennauer, W. Winckler, D.

Zabaneh, H. Campbell, C. van Duijn, A. G. Uitterlinden, A. Hofman, E. Sijbrands, G. R. Abecasis, K. R. Owen, E. Zeggini, M. D. Trip, N. G. Forouhi, A. C. Syvanen, J. G. Eriksson, L. Peltonen, M. M. Nothen, B. Balkau, C. N. Palmer, V. Lyssenko, T. Tuomi, B. Isomaa, D. J. Hunter, L. Qi, C. Wellcome Trust Case Control, G. Meta-Analyses of, I. Insulin-related traits Consortium, A. T. C. Genetic Investigation of, C. Asian Genetic Epidemiology Network-Type 2 Diabetes, C. South Asian Type 2 Diabetes, A. R. Shuldiner, M. Roden, I. Barroso, T. Wilsgaard, J. Beilby, K. Hovingh, J. F. Price, J. F. Wilson, R. Rauramaa, T. A. Lakka, L. Lind, G. Dedoussis, I. Njolstad, N. L. Pedersen, K. T. Khaw, N. J. Wareham, S. M. Keinanen-Kiukaanniemi, T. E. Saaristo, E. Korpi-Hyovalti, J. Saltevo, M. Laakso, J. Kuusisto, A. Metspalu, F. S. Collins, K. L. Mohlke, R. N. Bergman, J. Tuomilehto, B. O. Boehm, C. Gieger, K. Hveem, S. Cauchi, P. Froguel, D. Baldassarre, E. Tremoli, S. E. Humphries, D. Saleheen, J. Danesh, E. Ingelsson, S. Ripatti, V. Salomaa, R. Erbel, K. H. Jockel, S. Moebus, A. Peters, T. Illig, U. de Faire, A. Hamsten, A. D. Morris, P. J. Donnelly, T. M. Frayling, A. T. Hattersley, E. Boerwinkle, O. Melander, S. Kathiresan, P. M. Nilsson, P. Deloukas, U. Thorsteinsdottir, L. C. Groop, K. Stefansson, F. Hu, J. S. Pankow, J. Dupuis, J. B. Meigs, D. Altshuler, M. Boehnke, M. I. McCarthy, D. I. G. Replication and C. Meta-analysis (2012). "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes." Nat Genet 44(9): 981-990.

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nature methods. 2015 May;12(5):453.

Ng B, White CC, Klein HU, Sieberts SK, McCabe C, Patrick E, Xu J, Yu L, Gaiteri C, Bennett DA, Mostafavi S, De Jager PL. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci. 2017;20(10):1418-26.

Okbay, A., J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G. B. Chen, V. Emilsson, S. F. Meddens, S. Oskarsson, J. K. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, T. S. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, J. H. Brandsma, M. Pina Concas, J. Derringer, N. A. Furlotte, T. E. Galesloot, G. Girotto, R. Gupta, L. M. Hall, S. E. Harris, E. Hofer, M. Horikoshi, J. E. Huffman, K. Kaasik, I. P. Kalafati, R. Karlsson, A. Kong, J. Lahti, S. J. van der Lee, C. deLeeuw, P. A. Lind, K. O. Lindgren, T. Liu, M. Mangino, J. Marten, E. Mihailov, M. B. Miller, P. J. van der Most, C. Oldmeadow, A. Payton, N. Pervjakova, W. J. Peyrot, Y. Qian, O. Raitakari, R. Rueedi, E. Salvi, B. Schmidt, K. E. Schraut, J. Shi, A. V. Smith, R. A. Poot, B. St Pourcain, A. Teumer, G. Thorleifsson, N. Verweij, D. Vuckovic, J. Wellmann, H. J. Westra, J. Yang, W. Zhao, Z. Zhu, B. Z. Alizadeh, N. Amin, A. Bakshi, S. E. Baumeister, G. Biino, K. Bonnelykke, P. A. Boyle, H. Campbell, F. P. Cappuccio, G. Davies, J. E. De Neve, P. Deloukas, I. Demuth, J. Ding, P. Eibich, L. Eisele, N. Eklund, D. M. Evans, J. D. Faul, M. F. Feitosa, A. J. Forstner, I. Gandin, B. Gunnarsson, B. V. Halldorsson, T. B. Harris, A. C. Heath, L. J. Hocking, E. G. Holliday, G. Homuth, M. A. Horan, J. J. Hottenga, P. L. de Jager, P. K. Joshi, A. Jugessur, M. A. Kaakinen, M. Kahonen, S. Kanoni, L. Keltigangas-Jarvinen, L. A. Kiemeney, I. Kolcic, S. Koskinen, A. T. Kraja, M. Kroh, Z. Kutalik, A. Latvala, L. J. Launer, M. P. Lebreton, D. F. Levinson, P. Lichtenstein, P. Lichtner, D. C. Liewald, S. LifeLines Cohort, A. Loukola, P. A. Madden, R. Magi, T. Maki-Opas, R. E. Marioni, P. Marques-Vidal, G. A. Meddens, G. McMahon, C. Meisinger, T. Meitinger, Y. Milaneschi, L. Milani, G. W. Montgomery, R. Myhre, C. P. Nelson, D. R. Nyholt, W. E. Ollier, A. Palotie, L. Paternoster, N. L. Pedersen, K. E. Petrovic, D. J. Porteous, K. Raikkonen, S. M. Ring, A. Robino, O. Rostapshova, I. Rudan, A. Rustichini, V. Salomaa, A. R. Sanders, A. P. Sarin, H. Schmidt, R. J. Scott, B. H. Smith, J. A. Smith, J. A. Staessen, E. Steinhagen-Thiessen, K. Strauch, A. Terracciano, M. D. Tobin, S. Ulivi, S. Vaccargiu, L. Quaye, F. J. van Rooij, C. Venturini, A. A. Vinkhuyzen, U. Volker, H. Volzke, J. M. Vonk, D. Vozzi, J. Waage, E. B. Ware, G. Willemsen, J. R. Attia, D. A. Bennett, K. Berger, L. Bertram, H. Bisgaard, D. I. Boomsma, I. B. Borecki, U. Bultmann, C. F. Chabris, F. Cucca, D. Cusi, I. J. Deary, G. V. Dedoussis, C. M. van Duijn, J. G. Eriksson, B. Franke, L. Franke, P. Gasparini, P. V. Gejman, C. Gieger, H. J. Grabe, J. Gratten, P. J. Groenen, V. Gudnason, P. van der Harst, C. Hayward, D. A. Hinds, W. Hoffmann, E. Hypponen, W. G. Iacono, B. Jacobsson, M. R. Jarvelin, K. H. Jockel, J. Kaprio, S. L. Kardia, T. Lehtimaki, S. F. Lehrer, P. K. Magnusson, N. G. Martin, M. McGue, A. Metspalu, N. Pendleton, B. W. Penninx, M. Perola, N. Pirastu, M. Pirastu, O. Polasek, D. Posthuma, C. Power, M. A. Province, N. J. Samani, D. Schlessinger, R. Schmidt, T. I. Sorensen, T. D. Spector, K. Stefansson, U. Thorsteinsdottir, A. R. Thurik, N. J. Timpson, H. Tiemeier, J. Y. Tung, A. G. Uitterlinden, V. Vitart, P. Vollenweider, D. R. Weir, J. F. Wilson, A. F. Wright, D. C. Conley, R. F. Krueger, G. Davey Smith, A. Hofman, D. I. Laibson, S. E. Medland, M. N. Meyer, J. Yang, M. Johannesson, P. M. Visscher, T. Esko, P. D. Koellinger, D. Cesarini and D. J. Benjamin (2016). "Genome-wide association study identifies 74 loci associated with educational attainment." Nature 533(7604): 539-542.

Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH. Functional organization of the transcriptome in human brain. Nature neuroscience. 2008 Nov;11(11):1271.

Parikshak NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, Gandal MJ, Hartl C, Leppa V, Ubieta LT, Huang J, Lowe JK, Blencowe BJ, Horvath S, Geschwind DH. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism, Nature. 2016 Dec 15;540(7633):423-427.

Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, Legge SE, Bishop S, Cameron D, Hamshere ML, Han J. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nature genetics. 2018 Feb 26:1.

Pedersen CB, Bybjerg-Grauholm J, Pedersen MG, Grove J, Agerbo E, Baekvad-Hansen M, Poulsen JB, Hansen CS, McGrath JJ, Als TD, Goldstein JI. The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. Molecular psychiatry. 2018 Jan;23(1):6.

Pedersen BS, Quinlan AR. Who's Who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. The American Journal of Human Genetics. 2017 Mar 2;100(3):406-13.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC. PLINK: a toolset for whole-genome association and population-base linkage analysis. American Journal of Human Genetics, 2007;81.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008;5(12):1005-10.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander and E. L. Aiden (2014). "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." Cell 159(7): 1665-1680.

Ruderfer, D. M., A. H. Fanous, S. Ripke, A. McQuillin, R. L. Amdur, C. Schizophrenia Working Group of the Psychiatric Genomics, C. Bipolar Disorder Working Group of the Psychiatric Genomics, C. Cross-Disorder Working Group of the Psychiatric Genomics, P. V. Gejman, M. C. O'Donovan, O. A. Andreassen, S. Djurovic, C. M. Hultman, J. R. Kelsoe, S. Jamain, M. Landen, M. Leboyer, V. Nimgaonkar, J. Nurnberger, J. W. Smoller, N. Craddock, A. Corvin, P. F. Sullivan, P. Holmans, P. Sklar and K. S. Kendler (2014). "Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia." Mol Psychiatry 19(9): 1017-1024.

Salakhutdinov, R. and Hinton, G. Deep Boltzmann Machines. AISTATS, 2009.

Schunkert, H., I. R. Konig, S. Kathiresan, M. P. Reilly, T. L. Assimes, H. Holm, M. Preuss, A. F. Stewart, M. Barbalic, C. Gieger, D. Absher, Z. Aherrahrou, H. Allayee, D. Altshuler, S. S. Anand, K. Andersen, J. L. Anderson, D. Ardissino, S. G. Ball, A. J. Balmforth, T. A. Barnes, D. M. Becker, L. C. Becker, K. Berger, J. C. Bis, S. M. Boekholdt, E. Boerwinkle, P. S. Braund, M. J. Brown, M. S. Burnett, I. Buysschaert, Cardiogenics, J. F. Carlquist, L. Chen, S. Cichon, V. Codd, R. W. Davies, G. Dedoussis, A. Dehghan, S. Demissie, J. M. Devaney, P. Diemert, R. Do, A. Doering, S. Eifert, N. E. Mokhtari, S. G. Ellis, R. Elosua, J. C. Engert, S. E. Epstein, U. de Faire, M. Fischer, A. R. Folsom, J. Freyer, B. Gigante, D. Girelli, S. Gretarsdottir, V. Gudnason, J. R. Gulcher, E. Halperin, N. Hammond, S. L. Hazen, A. Hofman, B. D. Horne, T. Illig, C. Iribarren, G. T. Jones, J. W. Jukema, M. A. Kaiser, L. M. Kaplan, J. J. Kastelein, K. T. Khaw, J. W. Knowles, G. Kolovou, A. Kong, R. Laaksonen, D. Lambrechts, K. Leander, G. Lettre, M. Li, W. Lieb, C. Loley, A. J. Lotery, P. M. Mannucci, S. Maouche, N. Martinelli, P. P. McKeown, C. Meisinger, T. Meitinger, O. Melander, P. A. Merlini, V. Mooser, T. Morgan, T. W. Muhleisen, J. B. Muhlestein, T. Munzel, K. Musunuru, J. Nahrstaedt, C. P. Nelson, M. M. Nothen, O. Olivieri, R. S. Patel, C. C. Patterson, A. Peters, F. Peyvandi, L. Qu, A. A. Quyyumi, D. J. Rader, L. S. Rallidis, C. Rice, F. R. Rosendaal, D. Rubin, V. Salomaa, M. L. Sampietro, M. S. Sandhu, E. Schadt, A. Schafer, A. Schillert, S. Schreiber, J. Schrezenmeir, S. M. Schwartz, D. S. Siscovick, M. Sivananthan, S. Sivapalaratnam, A. Smith, T. B. Smith, J. D. Snoep, N. Soranzo, J. A. Spertus, K. Stark, K. Stirrups, M. Stoll, W.

H. Tang, S. Tennstedt, G. Thorgeirsson, G. Thorleifsson, M. Tomaszewski, A. G. Uitterlinden, A. M. van Rij, B. F. Voight, N. J. Wareham, G. A. Wells, H. E. Wichmann, P. S. Wild, C. Willenborg, J. C. Witteman, B. J. Wright, S. Ye, T. Zeller, A. Ziegler, F. Cambien, A. H. Goodall, L. A. Cupples, T. Quertermous, W. Marz, C. Hengstenberg, S. Blankenberg, W. H. Ouwehand, A. S. Hall, P. Deloukas, J. R. Thompson, K. Stefansson, R. Roberts, U. Thorsteinsdottir, C. J. O'Donnell, R. McPherson, J. Erdmann, C. A. Consortium and N. J. Samani (2011). "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease." Nat Genet 43(4): 333-338.

Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. arXiv preprint arXiv:1704.02685. 2017 Apr 10.

Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013 Dec 20.

Singh T, Kurki MI, Curtis D, Purcell SM, Crooks L, McRae J, Suvisaari J, Chheda H, Blackwood D, Breen G, Pietiläinen O. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nature neuroscience. 2016 Apr;19(4):571.

Sniekers, S., S. Stringer, K. Watanabe, P. R. Jansen, J. R. I. Coleman, E. Krapohl, E. Taskesen, A. R. Hammerschlag, A. Okbay, D. Zabaneh, N. Amin, G. Breen, D. Cesarini, C. F. Chabris, W. G. Iacono, M. A. Ikram, M. Johannesson, P. Koellinger, J. J. Lee, P. K. E. Magnusson, M. McGue, M. B. Miller, W. E. R. Ollier, A. Payton, N. Pendleton, R. Plomin, C. A. Rietveld, H. Tiemeier, C. M. van Duijn and D. Posthuma (2017). "Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence." Nat Genet 49(7): 1107-1112.

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440-5.

van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. BioRxiv. 2017 Jan 1:111591.

Weirauch M.T., Yang A., Albu M., Cote A.G., Montenegro-Montero A., Drewe P., Najafabadi H.S., Lambert S.A., Mann I., Cook K., Zheng H,. Goity A., van Bakel H., Lozano J.C., Galli M., Lewsey M.G., Huang E., Mukherjee T., Chen X., Reece-Hoyes J.S., Govindarajan S., Shaulsky G., Walhout A.J., Bouget F.Y., Ratsch G., Larrondo L.F., Ecker J.R., Hughes T.R. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158(6): 1431-1443. doi: 10.1016/j.cell.2014.08.009.

Won, H., L. de la Torre-Ubieta, J. L. Stein, N. N. Parikshak, J. Huang, C. K. Opland, M. J. Gandal, G. J. Sutton, F. Hormozdiari, D. Lu, C. Lee, E. Eskin, I. Voineagu, J. Ernst and D. H. Geschwind (2016). "Chromosome conformation elucidates regulatory relationships in developing human brain." Nature 538(7626): 523-527.

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics. 2011 Jan 7;88(1):76-82.

Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology. 2012 May 1;16(5):284-7.

Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology. 2005 Aug 12;4(1).

Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, Craig DW, Redman M, Gershon ES, Liu C. Genetic control of individual differences in gene-specific methylation in human brain. The American Journal of Human Genetics. 2010 Mar 12;86(3):411-9.