# 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[2,3*], 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], Dan 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, 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, 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

**Formatted:** Header

**Style Definition:** Heading 1

**Style Definition:** Heading 2

**Style Definition:** Heading 3

**Style Definition:** Heading 4

**Style Definition:** Heading 5

**Style Definition:** Heading 6

**Style Definition:** Title

**Style Definition:** Subtitle

**Deleted: (4481 + fig captions ~1400)**

**Deleted:** ¶

**Deleted:** 1

**Formatted:** Header

## Abstract

**Deleted: (190) (Ready for MG)**

Understanding how genomic variants relate to brain disorders remains challenging. To address this, the PsychENCODE consortium has generated functional genomic datasets on 1866 brains, including genotype, transcriptome, chromatin, and single-cell data. By uniformly processing and consistently analyzing these, we developed a comprehensive resource for the adult brain. In particular, we deconvolved the bulk tissue expression using single-cell data, finding that changing proportions of basic cell types explain >85% the across-population variation observed. Moreover, we used the chromatin and Hi-C data from reference brain samples to find ~79,000 active enhancers in the prefrontal cortex and link them to genes and transcription factors in a regulatory network. We identified ~2.5M eQTLs and many additional QTLs associated with chromatin, splicing and cell-type changes. In addition, we leveraged our QTLs, Hi-C data and regulatory network to find a variety of additional genes associated with GWAS variants with psychiatric disorders (e.g., 304 for schizophrenia). Finally, we developed a deep-learning model encapsulating the regulatory network to connect genotypes and phenotypes, achieving almost 5-fold improvement in disease-trait prediction above a conventional additive genetic model. This model enables highlighting key genes and functional modules and imputing missing transcriptome and epigenome from genotype data only.

**Deleted:** 112

**Deleted:** .

## Introduction

**Deleted: (460) (Ready for MG)**

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).

**Deleted:** \cite{19507169}.

**Deleted:** \cite{9603539, 24071849, 16461820, 19360079}.

**Deleted:** \cite{19339359}.

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), suggestive of 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).

**Deleted:** \cite{26404826}. Many of these lie in non-coding regions \cite{26404826}, suggestive of roles in gene regulation.

**Deleted:** \cite{27668389, 22031444},

**Deleted:** \cite{27760116,27339989, 26060301}.

**Deleted:** 2

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

**Formatted:** Header

**Deleted:** \cite{26605881}.

**Deleted: (196) (Ready for MG)**

**Formatted:** No underline

**Deleted:** Adult.PsychENCODE.org

**Deleted:** \cite{Supplement}.

**Deleted:** \cite{27339989, 26060301} .

**Deleted:** (schizophrenia, bipolar, autism

**Deleted:** \cite{Supplement}.

**Formatted:** No underline

**Deleted: Bulk &**

**Deleted: (660) (Ready for MG)**

**Commented [1]:** +wonhyejung87@gmail.com +shuang.liu@yale.edu +declan.clarke@yale.edu can we use 12K genes for trans-eQTL ?

**Deleted:** \cite{cap1, Supplement}.

**Deleted:** \cite{21614001, 29439242,18849986, 27409810}.

**Deleted:** 3

expression variation. Here, we address this question for expression over our cohort of 1,866 adult brains.

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≈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≈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,

**Formatted:** Header

**Deleted:** adults

**Deleted:** ,

**Deleted:** \cite{Supplement}.

**Deleted:** (

**Deleted:** )

**Deleted:** S5

**Deleted:** Discussion

**Deleted:** (

**Deleted:** )

**Deleted:** 2B and C, see \cite{Supplement}).

**Deleted:** (

**Deleted:** )

**Deleted:** (

**Deleted:** )

**Deleted:** \cite{Supplement}

**Deleted:** S6). We also compared our results with previous deconvolution methods \cite{Supplement}.

**Deleted:** \cite{Supplement}.

**Deleted:** SXXX

**Deleted:** autism

**Deleted:** .

**Commented [2]:** need to add Me stuff here. +fabio.navarro@yale.edu

**Deleted:** a gene involved in the early growth response was down-regulated in older groups, whereas ceruloplasmin was down-regulated in the middle-aged (Fig. 2F).

**Deleted:** 4

**Formatted:** Header

for genes involved in the early growth response (e.g., EGR1), expression and promoter methylation in older groups; in contrast, ceruloplasmin (e.g., CP) exhibits an opposite trend (Fig. 2F, S2.10 and 2.11) (*15*).

## Enhancers

**Deleted:** (440) (Ready for MG)

**Formatted:** No underline

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).

**Deleted:** \cite{Supplement}.

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 presents 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*).

**Deleted:** CBC) on a cohort of 50 individuals \cite{Supplement}.

**Deleted:** present

**Deleted:** \cite{27863250}.

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.

**Commented [3]:** what %?

**Deleted:** SX

## Consistent comparison: transcriptome and epigenome

**Formatted:** No underline

**Deleted:** &

**Deleted:** (270) (Ready for MG)

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 essentially reduces the dimensionality of the projections.

**Deleted:** \cite{Supplement}.

**Deleted:** 5

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 gene 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 (Fig. 4C), we identified ~2.5M cis-eQTLs (~386K independent after linkage-disequilibrium (LD) pruning) and ~33K eGenes (including non-coding ones) (Fig. 4A). 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 that can be expressed in brain. We evaluated the replication rate of GTEx and CMC eQTLs on PFC in our eQTLs set using $\pi_1$ statistic (23). GTEx and CMC $\pi_1$ values are 0.93 and 0.90 if we included associations with coding eGenes (0.95 and 0.84 for all eGenes). These results indicated high replication rates of GTEx and CMC brain PFC eQTLs in our study. 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).

**Formatted:** Header

**Deleted:** SXXX

**Deleted:** 3C) \cite{Supplement}.

**Deleted:** the

**Deleted:** (

**Deleted:** (408)(Ready for MG)

**Formatted:** No underline

**Deleted:** (2,097,741

**Formatted:** Font: Times New Roman, 12 pt, Font color: Auto, English (US)

**Deleted:** \cite{Supplement}. We expect

**Formatted:** English (US)

**Deleted:** the eQTL number is close to saturation, in terms of associating almost every possible variant

**Formatted:** Font color: Custom Color(RGB(34,34,34)), English (US), Pattern: Clear (White)

**Deleted:** an expression-modulating characteristic.

**Formatted:** Font: Times New Roman, 12 pt, Font color: Auto, English (US)

**Formatted:** English (US)

**Deleted:** . Finally, we also identified ~1.8M trans-eQTLs.

**Formatted:** Font: Times New Roman, 12 pt, Font color: Auto, English (US)

**Deleted:** \cite{25799442, 26300125}.

**Deleted:** \cite{supplement}.

**Deleted:** \cite{https://doi.org/10.1101/141986}.

**Deleted:** 6

Next, we determined if any SNPs were associated with changes in the relative fractions of cell types across individuals (fQTLs). In total, we identified 3720 distinct SNPs constituting 4186 different fQTLs. Of these, the proportions of microglia and excitatory neuron Ex8 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, appreciable number of eQTLs enriched on the promoter of a different gene than 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 also did π$_1$ statistics to evaluating the sharing among eQTLs with other QTLs. We found that sharing between eQTLs and cQTLs was the highest ($\pi_1$=0.89) while sharing between eQTLs and fQTLs was the lowest ($\pi_1$=0.11). There were 119 SNPs that functioned as QTLs in more than 3 different capacities (e.g. as eQTLs, cQTLs and isoQTLs). We dubbed these 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

**Formatted:** Header

**Commented [9]:** This is really interesting, as promoter-based Hi-C interactions are also enriched in promoters - supporting this claim that often times promoters act as enhancers.

**Deleted:** \cite{28581502}.

**Commented [10]:** No explanation of what this means.

**Deleted:** (Fig. 4C). There were 122

**Deleted:** Finally, we found XXX trans-eQTLs overlapping cis-eQTLs, necessarily for a different gene. Many of these cases in which a SNP is shared between a cis-eQTL and a trans-eQTL may constitute cis mediators, in which the SNP associated with a nearby gene (through a cis-eQTL) also constitutes a trans-eQTL if the nearby gene is involved in regulating the expression of a distant gene\cite{29021290}.

**Formatted:** Font: Times New Roman, 12 pt, Font color: Auto, English (US)

**Formatted:** No underline

**Deleted:** **(390) (Ready for MG)**

**Deleted:** \cite{27760116, Supplement}.

**Deleted:**

**Deleted:** 3

**Deleted:** \cite{27760116},

**Deleted:** 4.1-

**Deleted:** , Fig.

**Deleted:** .3

**Deleted:** .

**Deleted:** o r

**Deleted:** 7

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 mean square error (MSE< 0.05) (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 constituted a high-confidence set supported by more than two evidence sources (i.e., QTL and Hi-C, Fig. 5D-F, Fig. S7.1A); e.g., SCZ-gene, CACNA1C is regulated by multiple neuronal TFs via enhancers (Fig. 5E). The SCZ-genes represent a substantial increase from the previously reported 22 genes across 19 loci based on a smaller QTL set (8, 28). The majority of SCZ-genes were not in linkage disequilibrium with index SNPs (734 genes [~66%] with $r^2$<0.6, Fig. S7.1C), 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. S7.1B,D). 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-disease GWAS variants, not just schizophrenia. In particular, compared to non-brain related disorders, we found more significant heritability enrichments in cis-eQTL SNPs and GWAS SNPs for many brain disorders, with Schizophrenia having the strongest enrichment (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-

**Formatted:** Header

**Deleted:** \cite{Supplement}.

**Deleted:** 8930

**Deleted:** minimal

**Deleted:** Sxx

**Commented [12]:** This will be really interesting to ASD researchers. Is it possible to make a Figure related to this? What TFs mediate this?

**Commented [13]:** DW to follow up +daifeng.wang@yale.edu

**Commented [14]:** we need to make a supp fig & chg ASD to autism

**Deleted:** .

**Formatted:** Pattern: Clear

**Deleted:** ). It also had

**Formatted:** Pattern: Clear

**Deleted:** \cite{Supplement}.

**Deleted:** (266) **(Ready for MG)**

**Formatted:** No underline

**Deleted:** \cite{29483656},

**Deleted:** 320

**Deleted:** Sxx).

**Deleted:** \cite{29483656, 27668389}.

**Deleted:** >

**Deleted:** Sxx

**Deleted:** \cite{27760116}.

**Deleted:** in

**Deleted:** \cite{29483656},

**Deleted:** Sxx

**Deleted:** overlap

**Deleted:** of

**Deleted:** , Bipolar disorder, Major Depressive Disorder (MDD), Post-Traumatic Stress Disorder (PTSD), Alzheimers

**Deleted:** Parkinson's disease

**Formatted:** English (US)

**Formatted:** No underline

**Deleted:** (848) **(Ready for MG)**

**Deleted:** 8

learning framework, a Deep Structured Phenotype Network (DSPN) (15). This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from gene regulatory networks. 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 connectivity aspects, 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 genotype-to-phenotype models; 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 50% random baseline). In comparison, the DSPN was able to gain a larger 4.6X improvement (+17.4% 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 ~2.7X 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 a liability scale in order to compare 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 (26). These may be taken as upper bounds for additive predictive models using common variants, given unlimited data; by contrast, non-linear predictors can potentially exceed these limits. Our best liability scores based on 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 was similar order that explained by common SNPs for all three conditions (16.3%, 30%, and 14.4%); improved imputation may thus capture most of the variance due to common-SNP, narrow-sense heritability, although this is limited by the proportion of total variance in the imputed variables which is genetically determined (Fig. S8.2), as well as the sufficiency of the intermediate phenotypes used.

**Formatted:** Header

**Deleted:** , Supplement).

**Deleted:** 6

**Deleted:** ie

**Deleted:**

**Deleted:** 6D) \cite{Supplement}.

**Deleted:** schizophrenia, bipolar disorder, and autism, respectively \cite{bioRxiv_10.1101/048991}.

**Commented [17]:** correct

**Deleted:** . (By

**Deleted:** .)

**Deleted:** alone

**Deleted:** The

**Deleted:** of intermediate phenotypes

**Commented [18]:** todisc

**Deleted:** S8.1).

**Formatted:** Font: Times New Roman, 12 pt, Font color: Auto, English (US)

**Deleted:** 9

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 and examined which of these the DSPN prioritized, as well as new sets of genes associated with DSPN latent nodes that were uncovered at each hidden layer (Fig. S8.3) (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 (29). 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. S2.8).

## 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.

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, 30). 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-

**Formatted:** Header

**Deleted:** Supplement and Fig. S8.2).

**Deleted:** 6c, Fig.

**Deleted:** 3

**Deleted:** 4

**Deleted:** \cite{26814963}. Further,

**Deleted:** the prioritized modules/groupings for aging, the marker genes for

**Deleted:** types,

**Deleted:** SXXXX).

**Formatted:** Font: Times New Roman

**Deleted:**

**Deleted:** **(318)** **(Ready for MG)**

**Deleted:** Our integrative analyses here and with respect to the disease and developmental transcriptome \cite{cap1, devcap}, demonstrate that functional annotation of gene regulatory elements is necessary to understand the molecular mechanisms of disease-related genetic variation.¶
¶

**Deleted:** that

**Formatted:** Not Highlight

**Deleted:** arge

**Deleted:** \cite{26949524, 25053837}. This is particularly the case for specific cell sub-structures such as axons and dendrites \cite{25053837}.

**Deleted:** 10

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, 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, 31*) demonstrate that functional annotation of gene regulatory elements is useful for unraveling the molecular mechanisms in the brain.

# Figures

**Figure 1. Comprehensive data resource of functional genomics in 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).

**Figure 2. Deconvolution analysis of bulk and single-cell transcriptomics reveals cell fraction changes across tissue phenotypes and disorders**. **(A)** Genes had significantly higher expression variability across single cells than tissue samples. Left: dopamine gene, DRD3. **(B)** Top: the bulk tissue gene expression matrix (B, genes by individuals) can be decomposed by NMF to the product of two matrices: NMF component matrix (V, genes by top NMF components; i.e., NMF-TCs) and component fraction matrix (H, top NMF components by

**Formatted:** Header

**Deleted:**

**Commented [21]:** better last line

**Formatted:** Not Highlight

**Deleted:** **(1435)**

**Formatted:** Space After: 0 pt

**Deleted:** 1753

**Deleted:** 113

**Deleted:** 1866

**Formatted:** English (US)

**Deleted:** (

**Deleted:** ex

**Deleted:** (

**Deleted:** (

**Formatted:** Font: Times New Roman, 12 pt, Font color: Auto, English (US)

**Deleted:** 11

individuals); i.e., $B \approx VH$. 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., $B \approx CW$. 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 types for the biomarker genes (N=457). For example, NMF-7 highly correlated with the Ex3 cell type (r=0.66). **(D)** The estimate cell fractions contributed to >85% bulk tissue expression variations; i.e., $1-||B-CW||^2/||B||^2>0.85$. **(E)** The cell fractions changed across genders (control samples) and brain disorders. The neuronal cell types (e.g. In8) had significantly higher fractions 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)** The cell fractions, gene expression (EGR1) and methylation level (EGR1) changed across ages. The excitatory neuronal cell type Ex3 had a significant increase 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) were used to identify active enhancers, located in the open chromatin region (ATAC-seq peaks), with strong H3K27ac/H3K4me1 signal and lack 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 the reference brain enhancers in the population is shown as the Venn diagram. The histogram shows the overlapping 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 ellipse), Roadmap Epigenomics brain samples (light green ellipse), and other non-brain tissues (magenta ellipses) are plotted. The reference brain is shown as the purple dot (same in E and F). **(D)** The transcriptional diversity on 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 major tissue types including cerebellum, cortex, lung, skin, and testes, using PolyA RNA-seq data. **(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 ellipse) are shown on an RCA scatterplot of (E).

**Figure 4. Summary of QTLs of human adult brain PFC**. **(A)** Numbers of genes with at least one eQTL (eGenes) are shown compared to sample size in different studies. The number of

**Formatted** ... [1]

**Deleted:** -TCs and single cell types for the biomarker ... [2]

**Formatted** ... [3]

**Formatted** ... [4]

**Deleted:** . Enhancers were

**Deleted:** high

**Formatted** ... [6]

**Deleted:** signal

**Formatted** ... [5]

**Formatted** ... [7]

**Deleted:** a

**Formatted** ... [8]

**Deleted:** of

**Formatted** ... [9]

**Deleted:** peak signal

**Formatted** ... [10]

**Deleted:** cohort

**Formatted** ... [11]

**Deleted:** ,

**Formatted** ... [12]

**Deleted:** shallow to dark blue indicating low to high ... [13]

**Formatted** ... [14]

**Deleted:** .

**Formatted** ... [15]

**Deleted:** number of

**Formatted** ... [16]

**Deleted:** active

**Formatted** ... [17]

**Deleted:** achieved a saturation of over 70,000 with 20 ... [18]

**Moved down [1]: )** The coefficients (PC1 vs. PC2) of

**Formatted** ... [19]

**Deleted:** Other external brain samples (light green) ... [20]

**Formatted** ... [21]

**Deleted:** well. **(D)**

**Formatted** ... [22]

**Deleted:** center (cross) and ranges

**Formatted** ... [23]

**Deleted:** different tissue clusters (dashed ellipse) are ... [24]

**Formatted** ... [25]

**Deleted:** ). **(E**

**Formatted** ... [26]

**Deleted:** active brain enhancers

**Formatted** ... [27]

**Deleted:** The brain cluster (green ellipse) includes the

**Formatted** ... [28]

**Deleted:** tissue clusters

**Formatted** ... [29]

**Deleted:** consisting of Epigeomics Roadmap samples. **(**

**Formatted** ... [30]

**Deleted:** CB (lightblue),

**Formatted** ... [31]

**Deleted:** (xxx),

**Formatted** ... [32]

**Deleted:** (xxx), ovary (xxx),

**Formatted** ... [33]

**Deleted:** (xx),

**Formatted** ... [34]

**Deleted:** (xx),

**Formatted** ... [35]

eGenes increased as the sample size increased. The eGenes of PsychENCODE is close to saturation. The estimated replication $\pi_1$ values of GTEx and CMC eQTLs in PsychENCODE are 0.93 and 0.90 respectively.  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 QTLs, eGenes, enhancers, celltypes and QTL SNPs are shown in the left table. 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 in pie charts. Sharing of the QTLs vs. eQTLs are shown using $\pi_1$ values in the orange bar plot indicating the highest sharing is between cQTLs vs. eQTLs. An example on the right side shows the sharing SNPs in orange of eQTLs and cQTLs for gene MTOR. (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.

**Figure 5. Data integration and modeling predicts gene regulatory network, revealing additional GWAS genes for psychiatric disorders**. **(A)** The full Hi-C data from adult brain reveal the folding principle 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 a full gene regulatory network consisting of ~150,000 Hi-C interactions, ~2 million eQTL-eGene linkages, ~211k TF-to-target and ~577k 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. An overlap between SCZ-genes defined by QTL associations (QTL), chromatin interactions (Hi-C), and activity relationships (Activity) is depicted in 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 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 (excitatory neurons) than others.

**Figure 6. Deep-learning model predicts genotype-phenotype 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 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 structure learned previously is embedded in layers L0 and L1, with different types of regulatory

**Formatted** ... [39]
**Formatted** ... [40]
**Deleted:** B
**Formatted** ... [41]
**Deleted:** genes with at least one eQTL (
**Deleted:** )
**Formatted** ... [43]
**Deleted:** compared to sample size
**Deleted:** different studies. The number of eGenes ... [45]
**Deleted:** sample size increased. The eGenes of ... [47]
**Formatted** ... [42]
**Formatted** ... [44]
**Formatted** ... [46]
**Deleted:** sQTL
**Formatted** ... [49]
**Deleted:** is
**Formatted** ... [48]
**Formatted** ... [50]
**Deleted:** In total, 80% of eQTLs overlapped with other ... [51]
**Formatted** ... [52]
**Deleted:** of overlapped SNPs of eQTL with other QTLs ... [53]
**Formatted** ... [54]
**Deleted:** among all QTLs. A total of 31% of fQTL SNPs ... [55]
**Formatted** ... [56]
**Deleted:** (9%) (p<1e-15, odds ratio:2.07).
**Formatted** ... [57]
**Deleted:** Enrichment of
**Formatted** ... [58]
**Deleted:** SNPs on cis-eQTL SNPs is shown. ... [59]
**Deleted:** disorders on cis-
**Formatted** ... [61]
**Deleted:** SNPs were more significant
**Formatted** ... [62]
**Deleted:** those of
**Formatted** ... [60]
**Formatted** ... [63]
**Deleted:** SNPs. Schizophrenia GWAS SNPs had the ... [64]
**Formatted** ... [65]
**Formatted** ... [66]
**Deleted:** enhancer-promoter-based interactions. We ... [67]
**Formatted** ... [68]
**Deleted:** eQTLs (gold),
**Formatted** ... [69]
**Deleted:** 302
**Formatted** ... [70]
**Deleted:** schizophrenia
**Formatted** ... [71]
**Deleted:** is shown
**Formatted** ... [72]
**Deleted:**
**Formatted** ... [73]
**Formatted** ... [74]
**Deleted:** 13

Formatted: Header

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 across models of different complexity; using different predictors (genotype/transcriptome); and with or without imputation (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 60.2%, which can be expressed as a 2.7X improvement above chance (+10.2% vs. +3.8%, blue).  Corresponding improvements in liability variance scores are shown in brackets. Disorders are abbreviated as in the main text, GEN=Gender, ETH=Ethnicity, AOD=Age of death.

Formatted: Font: Times New Roman, 12 pt, Font color: Auto, English (US)

## Reference

Deleted: ¶

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).

Deleted: 14

Formatted: Header

15. M. a. m. a. a. a. s. 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. A. Sekar *et al.*, Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177-183 (2016).

30. S. Liu, C. Trapnell, Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5**,  (2016).

31. M. Li, e. al., Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risk. *submitted*.

Deleted: 15

**Page 1: [1] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

Header

**Page 12: [2] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

**Page 12: [2] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

**Page 12: [2] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

**Page 12: [2] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

**Page 12: [2] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

**Page 12: [3] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

Space After: 0 pt

**Page 12: [4] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [4] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [5] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [6] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [7] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [7] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [8] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [9] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [10] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [11] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [12] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [13] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

shallow to dark blue indicating low to high signal peak

**Page 12: [14] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [15] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [15] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [15] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [16] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [17] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [18] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

achieved a saturation of over 70,000 with 20 samples. **(C**

**Page 12: [19] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [19] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [19] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [20] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

Other external brain samples (light green) including GTEx and other tissue samples (magenta) are shown

**Page 12: [21] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [22] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [23] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [24] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

different tissue clusters (dashed ellipse) are shown on an RCA scatterplot of

**Page 12: [25] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [25] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [26] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [27] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [28] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [28] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [29] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [30] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Font: Not Bold, English (US)

**Page 12: [30] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Font: Not Bold, English (US)

**Page 12: [30] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Font: Not Bold, English (US)

**Page 12: [30] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Font: Not Bold, English (US)

**Page 12: [31] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [32] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [33] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [34] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [35] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [36] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [36] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [36] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 12: [36] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 12: [37] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

**Figure 4. Summary of QTLs of human adult brain PFC**. (**A**

**Page 12: [38] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

Line spacing: single, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

**Page 1: [39] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

Header

**Page 13: [40] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [41] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [42] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [43] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [44] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [45] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

different studies. The number of eGenes increased as

**Page 13: [46] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [47] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

sample size increased. The eGenes of PsychENCODE is close to saturation. (**C**)

**Page 13: [48] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [49] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [50] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [51] Deleted** **Daifeng Wang** **4/7/18 5:20:00 PM**

In total, 80% of eQTLs overlapped with other QTLs. The

**Page 13: [52] Formatted** **Daifeng Wang** **4/7/18 5:20:00 PM**

English (US)

**Page 13: [53] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

of overlapped SNPs of eQTL with other QTLs was

**Page 13: [54] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [55] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

among all QTLs. A total of 31% of fQTL SNPs overlapped with other QTLs, which was the lowest among all QTLs; 36% of sQTL and cQTL SNPs overlapped with other QTL SNPs. fQTLs overlapped more with sQTLs (17%) than

**Page 13: [56] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [57] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [58] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [59] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

SNPs on cis-eQTL SNPs is shown. Enrichment for GWAS SNPs of

**Page 13: [60] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [61] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [62] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [63] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [64] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

SNPs. Schizophrenia GWAS SNPs had the highest enrichment on cis-eQTLs SNPs among the three brain disorders analyzed

**Page 13: [65] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [65] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [66] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Space After: 0 pt

**Page 13: [67] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

enhancer-

**Page 13: [67] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

enhancer-

**Page 13: [67] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

enhancer-

**Page 13: [67] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

enhancer-

**Page 13: [67] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

enhancer-

**Page 13: [67] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

enhancer-

**Page 13: [67] Deleted Daifeng Wang 4/7/18 5:20:00 PM**

enhancer-

**Page 13: [68] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Font: Times New Roman, 12 pt, Font color: Auto

**Page 13: [68] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Font: Times New Roman, 12 pt, Font color: Auto

**Page 13: [69] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [70] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [71] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [72] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

English (US)

**Page 13: [73] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Font: Times New Roman, 12 pt, Font color: Auto, English (US)

**Page 13: [74] Formatted Daifeng Wang 4/7/18 5:20:00 PM**

Space After: 0 pt