Comprehensive functional genomic resource and integrative model for the human brain

Daifeng Wang[1,2,3*], Shuang Liu[1,2*], Jonathan Warrell[1,2*], Hyejung Won[4,5*], Xu Shi[1,2*], Fabio C.P. Navarro[1,2*], Declan Clarke[1,2*], Mengting Gu[1*], Prashant Emani[1,2*], Yucheng T. Yang[1,2], Min Xu[1,2], Michael J. Gandal[6], Shaoke Lou[1,2], Jing Zhang[1,2], Jonathan J. Park[1,2], Chengfei Yan[1,2], Suhn Kyong Rhie[13], Kasidet Manakongtreecheep[1,2], Holly Zhou[1,2], Aparna Nathan[1,2], Mette Peters[14], Eugenio Mattei[15], Dominic Fitzgerald[16], Tonya Brunetti[16], Jill Moore[15], Yan Jiang[17], Kiran Girdhar[18], Gabriel E. Hoffman[18], Selim Kalayci[18], Zeynep H. Gümüş[18], Gregory E. Crawford[19], PsychENCODE Consortium[‡], Panos Roussos[17,18], Schahram Akbarian[17,20], Andrew E. Jaffe[22], Kevin P. White[16,23], Zhiping Weng[15], Nenad Sestan[21], Daniel H. Geschwind[7-9†], James A. Knowles[10†], Mark B. Gerstein[1,2,11,12,24†]

Affiliations:
[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA
[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[3]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA
[4]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA
[5]UNC Neuroscience Center, University of North Carolina, Chapel Hill, NC 27599, USA
[6]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.
[7]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[8]Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[9]Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David
[10]SUNY Downstate Medical Center College of Medicine, Brooklyn, NY 11203, USA
[11]Department of Computer Science, Yale University, New Haven, CT 06520, USA
[12]Department of Computer Science, Yale University, New Haven, CT 06520, USA
[13]Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90007, USA
[14]Sage Bionetworks, Seattle, WA 98109, USA
[15]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA
[16]Institute for Genomics and Systems Biology, Department of Human Genetics, University of Chicago, Illinois 60637, USA
[17]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[18]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[19]Center for Genomic and Computational Biology, Department of Pediatrics, Duke University, Durham, NC 27708, USA

[20]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[21]Department of Neuroscience and Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT 06520, USA
[22]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
[23]Tempus Labs Inc., Chicago, IL 60654, USA
[24]Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

* These authors contributed equally to this work
‡ The consortium authors are listed at the end of the paper.
† Co-corresponding authors

Abstract

Despite progress in defining genetic risk for psychiatric disorders, their molecular mechanisms remain elusive. Addressing this, PsychENCODE has generated a comprehensive resource for the adult brain across 1866 individuals (resource.psychencode.org). It contains ~79K brain-active enhancers, sets of Hi-C linkages and TADs, single-cell expression profiles for many cell types, expression QTLs, and further QTLs associated with chromatin, splicing, and cell-type proportions. Integration shows varying cell-type proportions largely account for the cross-population variation in expression (with >88% reconstruction accuracy). It also allows building a gene-regulatory network, linking GWAS variants to genes (e.g., 321 for schizophrenia). We embed this network into an interpretable deep-learning model, which improves disease prediction ~6X vs. polygenic risk scores and identifies key genes and pathways in psychiatric disorders.

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 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 strongly associated with neuropsychiatric disease risk -- for instance, the Psychiatric Genomics Consortium (PGC) has identified 142 GWAS loci associated with schizophrenia (SCZ) (*6*) -- for most of these variants we have little understanding of the molecular mechanisms affecting the brain (*7*).

Many of these variants lie in non-coding regions, and large-scale studies have begun to elucidate the changes in genetic and epigenetic activity associated with these genomic alterations, suggesting potential molecular mechanisms. In particular, the Genotype-Tissue-Expression (GTEx) project has associated many non-coding variants with expression quantitative-trait loci (eQTLs), and the ENCODE and Roadmap Epigenomics (Roadmap)

projects have identified non-coding regions acting as enhancers and promoters (*8-10*). However, none of these projects have focused their efforts on the human brain. Initial work focusing on brain-specific functional genomics has provided greater insight but could be enhanced with larger sample sizes (*11, 12*). Moreover, new methodologies such as Hi-C and single-cell sequencing, have yet to be fully integrated, at scale, with brain genomics data (*13-16*).

Hence, the PsychENCODE Consortium has generated large-scale data to provide insight into the brain and psychiatric disorders, including data derived through genotyping, bulk and single-cell RNA-seq, ChIP-seq, ATAC-seq, and Hi-C (*17*). All data have been placed into a central, publicly available resource that also integrates relevant re-processed data from related projects, including ENCODE, CommonMind (CMC), GTEx, and Roadmap. Using this resource, we identified functional elements, QTLs and regulatory-network linkages specific to the adult brain. Moreover, we combined these elements and networks to build an integrated deep-learning model that predicts high-level traits from genotype via intermediate molecular phenotypes. Here, by "intermediate phenotypes" we mean the read out of functional genomic information on genomic elements (e.g., gene expression and chromatin activity). In some contexts, these are also referred to as “molecular endophenotypes” (*18*). However, we include additional low-level "phenotypes" such as cell fractions, so we use the more general term intermediate phenotype. We also refer to the high-level traits as “observed phenotypes,” which include both classical clinical variables and characteristics of healthy individuals, such as gender and age.

Resource construction

Resource.PsychENCODE.org is the central site for this paper. It organizes data hierarchically, with a base of raw data files, a middle layer of uniformly processed and easily shareable results (such as open chromatin regions and gene-expression quantifications), and a top-level, “cap” of an integrative, deep-learning model, based on regulatory networks and QTLs. To build the base layer we included all adult brain data from PsychENCODE and merged these with relevant data from ENCODE, CMC, GTEx, Roadmap, and recent single-cell studies (Table S1, Fig. 1). In total, the resource contains 3,810 genotype, transcriptome, chromatin and Hi-C datasets from PsychENCODE and 1,662 datasets using similar “bulk” assays merged from outside the consortium. Overall, the datasets from prefrontal cortex (PFC) involve sampling from 1,866 individuals. The resource also has single-cell RNA-seq for 18,025 cells from PsychENCODE and 14,012 from outside sources (*19*). These data represent a range of psychiatric disorders including schizophrenia (SCZ), Bipolar Disorder (BPD), and Autism Spectrum Disorder (ASD). Note, the individual genotyping and raw next-generation sequencing of transcriptomics and epigenomics are restricted for privacy protection but access can be obtained upon approval. The protocols for all associated data are readily available (Fig. S1). Finally, PsychENCODE has developed a reference brain project on PFC, utilizing matched assays on the same set of brain tissues, which we used to develop an anchoring annotation (*20*).

Transcriptome analysis: bulk & single-cell

To identify the genomic elements exhibiting transcriptional activities specific to the brain, we took a conservative approach and used the standardized and established ENCODE pipeline to uniformly process RNA-seq data from PsychENCODE, GTEx and Roadmap (Figs. S2 and S3). This consistency makes our expression data and subsequent results (including eQTLs and single-cell analyses) comparable with previous work. Using these data, we identified non-coding regions of transcription and sets of differentially expressed and co-expressed genes (*20, 21*).

Brain tissue is composed of a variety of basic cell types. Gene expression changes observed at the tissue level may be due to changes in the proportions of basic cell types (*22-27*). However, it is unclear how these changes in cell proportions can contribute to the variation in tissue-level gene expression observed across a population of individuals. To address this question, we used two complementary strategies across our cohort of 1,866 individuals.

First, we used standard pipelines to uniformly process single-cell RNA-seq data from PsychENCODE, in conjunction with other single-cell studies on the brain (*14, 16, 19*). Then we assembled profiles of brain cell types, including both excitatory and inhibitory neurons (denoted as Ex1 to Ex9, and In1-8, following previous conventions), major non-neuronal types (e.g., microglia and astrocyte), and additional cell types associated with development (*20*). Depending on the underlying sequencing and quantification, our profiles were of two fundamentally different formats, Transcripts Per Kilobase Million (TPM) and Unique Molecular Identifier counts (UMI). The former ("TPM-profiles") includes the uniformly processed PsychENCODE developmental single-cell data merged with published adult and developmental data (Fig. S4 and Table S2) (*14, 16*). In contrast, the "UMI-profiles" are built by merging PsychENCODE adult single-cell profiles with other recently published datasets (*14*). Both formats share common neuronal and major non-neuronal cell types and are used interchangeably in various analyses here (Fig. S5; Tables S3 and S4). Moreover, the expression values of biomarker genes for the same cell type were correlated between two formats (Figs. S6 and S7). Note, however, that our TPM-profiles have additional development-specific cell types, such as quiescent and replicating.

From both sets of profiles, we can generate a matrix **C** of expression signatures, comprising marker genes and their expression levels across various cells (Fig. S8). In this matrix, a number of genes had expression levels that varied more across cell types than they did in bulk-tissue measurements across individuals in a population (e.g., dopamine receptor DRD3; Fig. 2A). This suggests cell-type changes across individuals could contribute substantially to variation in individual bulk expression levels.

Second, we used an unsupervised analysis to identify the primary components of bulk expression variation. We decomposed the bulk gene-expression matrix using non-negative matrix factorization (NMF, $\mathbf{B} \approx \mathbf{VH}$), and determined whether the top components, capturing the majority of covariance (NMF-TCs, columns of **V**; Fig. 2B), were consistently associated with the single-cell signatures (Fig. 2C) (*20*). A number of NMF-TCs were, in fact, highly correlated with cell types from matrix **C** for both TPM and UMI data -- e.g., component NMF-17 is correlated with the Ex2 cell type (r=0.63, Figs. 2C and S9). This demonstrates that an unsupervised

analysis derived solely from bulk data can roughly recapitulates the single-cell signatures, partially corroborating them.

We then examined how variation in the proportions of basic cell types contributes to variation in bulk expression. To this end, we estimated the relative proportions of various cell types for each tissue sample (i.e., "cell fractions"). In particular, we deconvolved the bulk, tissue-level expression matrix using the single-cell signatures to estimate cell fractions across individuals (**W**), solving **B≈CW** (Fig. 2B) (*20*). As a validation, our estimated fractions of NEU+/- cells matched the experimentally determined fractions from reference brain samples (Median difference = 0.04, Fig. S10). Overall, our analyses demonstrated that variation in cell types contributed significantly to bulk variation. That is, weighted combinations of single-cell signatures could account for most of the population-level expression variation, with an accuracy of >88% (Fig. 2D, $1 - || \mathbf{B} - \mathbf{CW} ||^2 / || \mathbf{B} ||^2 > 88\%$), and when calculated on a per person basis, this quantity varies ± 4% over the 1866 individuals in our cohort (Figs. S11 and S12). Also, our results explained more variation than previous deconvolution approaches (Fig. S13) (*20*).

We identified cell-fraction changes associated with different traits (Figs. 2E, S14, S15, S16, and S17). For example, there are different fractions of particular types of excitatory and inhibitory neurons in male and female samples – e.g., the fraction of In6 (Fig. 2E). Also, in individuals with ASD, the fraction of Ex5 was higher and oligodendrocytes, lower, with some commensurate increase for microglia and astrocytes (Figs. 2E, and S18) (*23, 28*).

Finally, we observed an association with age. In particular, with increasing age the fractions of Ex3 and Ex4 significantly increased and some non-neuronal types decreased (Figs. 2F and S19). These changes may be associated with differential expression of specific genes, e.g., Somatostatin (SST), known to be associated with aging and neurotransmission (Fig. 2F) (*29*) . Also, SST exhibits increasing promoter methylation with age, perhaps explaining its decreasing expression. Other genes known to be associated with brain aging exhibit different trends – e.g., EGR1 and CP (Figs. 2F, S20, and S21) (*20, 30*).

Enhancers

To annotate brain-active enhancers, we used chromatin-modification data from the reference brain, supplemented by DNase and ChIP-seq data from Roadmap PFC samples. All data were processed by standard ENCODE ChIP-seq pipelines, to ensure maximal compatibility of our results (Fig. S22). Consistent with ENCODE, we define active enhancers as open chromatin regions enriched in H3K27ac and depleted in H3K4me3 (Figs. 3A and S23) (*20*). Overall, we annotated a reference set of 79,056 enhancers in PFC. (We also provide a filtered subset (*20*).)

Assessing the variability across individuals and tissues for enhancers is more difficult than for gene expression (*31*). Not only is the variability in chromatin-mark level at enhancers across different individuals and tissues high, but the boundaries of enhancers can grow and shrink, sometimes disappearing altogether (e.g., H3k27ac; Fig. 3A). To investigate this in more detail, we uniformly processed the H3K27ac data from PFC, temporal cortex (TC), and cerebellum

(CB) on a cohort of 50 individuals, primarily of European descent and sequenced to similar depth (*20*) (Fig. S24). Aggregating 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 -- where each peak is present in more than half of the individuals surveyed. Comparing aggregated sets for these three brain regions, PFC was more similar to TC than CB (~90% vs 34% overlap in peaks). This difference is consistent with previous reports and suggests potentially different cell-type composition in cerebellum from cortex (*32, 33*).

We also examined how many of the enhancers in the reference brain are active in each of the individuals in our cohort (i.e., having enriched H3K27ac). As expected, not every reference enhancer was active in each individual. On average, only ~70% ± 15% (~54,000) of the enhancers in the reference brain were active in an individual in the cohort, and a similar fraction of the reference enhancers was active in more than half the cohort (68%, Fig. 3B). To estimate the total number of enhancers in PFC, we calculated the cumulative number of active regions across the cohort (Fig. S25). This increased for the first 20 individuals sampled, but saturated at the 30th. Thus, we hypothesize that pooling PFC enhancers from ~30 individuals is sufficient to cover nearly all possible PFC enhancer regions, estimated at ~120,000.

Consistent comparison: transcriptome & epigenome

As we uniformly processed the transcriptomic and epigenomic data across the PsychENCODE, ENCODE, GTEx, and Roadmap datasets, we could compare the brain to other organs in a consistent fashion and also compare transcriptome variation to that of the epigenome (Figs. 3C-F). Several approaches, including PCA, t-SNE, and reference component analysis (RCA) were tested to determine the best method for comparison. We found that, although popular and interpretable, PCA de-emphasizes local structure and is overly influenced by outliers; in contrast, t-SNE preserves local relationships but “shatters” global structure. RCA is a compromise *(20)*: it captures local structure while maintaining meaningful distances globally. We used RCA to project gene expression from PsychENCODE samples against a reference panel of gene-expression for different tissues derived from GTEx, and then reduced the dimensionality of the projections with PCA. RCA thus allowed us to represent high-dimensional expression data in a simple two-coordinate diagram.

For gene expression, RCA revealed that the brain separates from the other tissues in the first component (Fig. 3E and S26). In particular, for brain, inter-tissue comparisons exhibit more differences than intra-tissue ones (Figs. S27, S28, S29, and S30). A different picture emerged for chromatin. The H3K27ac chromatin levels at all regulatory positions were, overall, less distinguishable between brain and other tissues (Fig. 3C) (*20*). At first glance, this is surprising as one expects great differences in enhancer usage between tissues. However, our analysis compares chromatin signals over all regulatory elements from ENCODE (including enhancers and promoters), which is logically consistent with our expression comparison across all protein-coding genes (Fig. 3F vs. 3C, Tables S5, S6, and S7). As the total number of human regulatory elements is much larger than brain-active enhancers (~1.3M vs. ~79K), our results likely reflect the fact that there are proportionately fewer brain-active regulatory elements than protein-coding genes (6% vs. 60%).

Up to this point, our analysis has focused on 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 (Fig. S30). In particular, testes and lung have the largest extent of transcription overall for protein-coding genes (i.e., the most genes transcribed, Fig. 3D). However, when we shift to unannotated regions the ordering changes: brain tissues, such as cortex and cerebellum, now have a greater extent of transcription than any other tissue.

QTL analysis

We used the data in the brain resource to identify QTLs affecting gene expression and chromatin activity. We calculated expression, splicing-isoform, chromatin and cell-fraction QTLs (eQTLs, isoQTLs, cQTLs and fQTLs, respectively). For eQTLs, we adopted a standard approach, closely adhering to the GTEx pipeline for maximal compatibility (Figs. S31, S32, and S33; (*34*)). (However, for maximal utility of the resource, we also provide alternate lists, filtered more conservatively.) In PFC, we identified ~2.5M cis-eQTLs involving ~33K eGenes (i.e. expressed genes, ~17K non-coding and ~16K coding, with FDR<0.05; Fig. 4A). We found 1,341,182 eQTL SNPs from ~5.3M total SNPs tested in 1 Mb windows around genes, comprising 238,194 independent SNPs after linkage-disequilibrium (LD) pruning. This estimate identified substantially more eQTLs and associated eGenes than previous studies, reflecting our large sample size (*8, 11, 20*). The number of eGenes, in fact, approaches the total number of genes estimated to be expressed in brain. That said, a very large fraction of the smaller GTEx and CMC brain eQTL sets were contained with our set (as evident from overlap testing with the π_1 statistic, Fig. 4A) (*35*). Moreover, as expected, our brain eQTL set showed higher π_1 similarity and SNP-eGene overlap to GTEx brain eQTLs than those from other tissues (Figs. 4B and S31). Finally, we applied the QTL pipeline to isoform levels to calculate a set of isoQTLs. We performed filtering in a variety of different ways, generating a number of different lists (*20*).

For cQTLs no established methods exist for large-scale data, although there have been previous efforts (*36, 37*). To identify cQTLs, we focused on our reference set of enhancers and examined how H3K27ac activity varied at these loci across 292 individuals (Fig. 4C) (*20*). Overall, we identified ~2,000 cQTLs in addition to 6,200 identified from individuals within the CMC cohort (*38*).

We next identified SNPs associated with changes in the relative abundances of specific cell types. We term such relationships “cell-fraction QTLs” (fQTLs). In total, we identified 1672 distinct SNPs constituting 4199 fQTLs (Fig. S34). Of these, the excitatory neurons Ex4 and Ex5 were associated with the most fQTLs (1060 and 896, respectively). The biological mechanism governing a fQTL may involve other QTL types, such as eQTLs. An illustrative example is FZD9 (Fig. 4D): we found the expression levels of this gene were associated with a neighboring non-coding SNP via an eQTL, and this same SNP was associated with the proportion of Ex3 cells via a fQTL. Perhaps connected to this, deletion variants upstream of FZD9 had been previously been associated with cell-fraction changes, related to Williams syndrome (*39*).

Next, we attempted to re-calibrate the observed gene-expression variation, considering fQTLs. In particular, our scheme, described above, for approximately deconvolving gene expression from heterogeneous bulk tissue (**B**) into single-cell signatures (**C**) and estimated cell fractions (**W**) enables us to calculate the residual gene expression (**Δ**) remaining after accounting for cell fraction changes (Fig. 2). Specifically, it is the component of the bulk tissue expression variation that cannot be explained by the changing cell fractions alone: **Δ** = **B** - **CW**. We can subsequently use this quantity to determine "residual QTLs" by directly correlating it with genotype. In total, this results in 202,940 SNPs involved in residual eQTLs. Potentially, one can elaborate on this further by allowing the correlations to be done in a cell-type specific fashion (Fig. S35).

To further dissect the associations between genomic elements and QTLs, we compared all of the different types of QTLs with each other and with genomic annotations (Fig. 4E). As expected, eQTLs tended to be enriched at promoters, and cQTLs, at enhancers and TF-binding sites; 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 the activity of an Epromoter, a regulatory element with dual promoter and enhancer functions *(40)*. For the overlap among different QTLs, we expected that most cQTLs and fQTLs would be a subset of the much larger number of eQTLs; somewhat surprisingly, an appreciable number of these did not overlap (Fig. 4F). To evaluate this precisely, we calculated π_1 statistics and found that the cQTL overlap was larger than fQTL overlap (0.89 vs 0.11). Moreover, eQTL-cQTL overlaps often suggested that the expression-modulating function of an eQTL derived from chromatin changes (e.g. for MTOR, Fig. 4F). Overall, the total number of overlapping QTLs was 2,477 (which we dub multi-QTLs, Fig 4F).

Regulatory networks

We next integrated the genomic elements described above into a regulatory network. 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, Table S8) (*13, 20*). In total, we identified 2,735 topologically associating domains (TADs) and ~90K enhancer-promoter interactions (Fig. S36). As expected, ~75% of enhancer-promoter interactions occurred within the same TAD, and genes with more enhancers tended to have higher expression (Figs. 5B and S36). We integrated the Hi-C data with QTLs; surprisingly, QTLs involving SNPs distal to eGenes but linked by Hi-C interactions showed significantly stronger associations (i.e. QTL p-value) than those with SNPs directly in the eGene promoter or exons (Figs. 5C and S37).

To gain insights on the brain chromatin, we compared the adult PsychENCODE Hi-C dataset to those from other tissues in a similar fashion to the transcriptomic and epigenomic comparisons above. In particular, we selected a set of tissues and cell types from ENCODE and Roadmap, consistently processed their associated Hi-C data at a low resolution and compared them to our reference-brain Hi-C data. As expected, we found that all the samples for adult brain regions tend to separate markedly from the other tissues in terms of A/B compartment similarity and other metrics (Figs. 5D and S38).

In addition to the adult brain, we also added PsychENCODE Hi-C data of fetal brain into the comparison, assessing the degree to which the chromatin differences between developmental stages relate to those between tissues (Fig. 5D). We found that while Hi-C datasets for adult brain clustered together, the Hi-C dataset for fetal brain was distinct (Figs. 5D and S39). Indeed, only ~31% of the interactions in our adult Hi-C data were detected in the fetal dataset (Figs. S39 and S40) (*13*). While hard to exactly quantify, this difference appears larger than that seen from cross-tissue transcriptome comparison, with fetal samples included (Fig. S41). We did a number of other comparisons between fetal and adult brain Hi-C datasets, analyzing the regulatory elements and genes linked by each. As expected, we found fetal-linked genes more highly expressed, prenatally, and adult-linked ones, postnatally (Fig. 5E). In addition, the fetal-linked genes were preferentially expressed in developmental cell types (Fig. 5F). They were also highly expressed in adult neurons, while the adult-linked ones were preferentially expressed in glia, reflecting known cell-type composition (Figs. 5D and 5F) (*41*).

In addition to Hi-C linkages, we tried to find further regulatory connections by relating the activity of TFs to target genes (Fig. 5A). In particular, for each potential target of a TF, we created a linkage if (i) it had a "good binding site" (matching the TF's motif) in gene-proximal open chromatin regions (either promoters or brain-active enhancers) and (ii) it had a high coefficient in a regularized, elastic net regression, relating TF activity to target expression (Fig. S42) (*20*). Elastic-net regression assumes that target-gene expression is determined by a linear combination of the expression levels of its regulating TFs, via regression coefficients (using sparsified L_1 and L_2 regularization). Overall, we found that a subset of regulatory connections could predict the expression of 8,930 genes with MSE< 0.05 (mean-square error, Fig. S43). For example, we could predict the expression of the ASD-associated gene CHD8 with MSE=0.034 (equivalent to R^2=0.77 over the population) (*20*). Finally, the enhancer-binding TFs with high regression coefficients -- implying a high chance for TF regulation of the target genes via particular bound enhancers -- provide a third set of putative enhancer-to-gene links.

Collectively, we generated a full regulatory network, linking enhancers, TFs, and target genes (Fig. S42). This includes 43,181 proximal and 42,681 distal linkages involving 11,573 protein-coding target genes (TF-to-target gene via promoter for proximal vs via enhancer-target-gene connection for distal; Fig. 5A; 15 (*20*)). As functioning regulatory connections reflect cell type, we also generated potential cell-type specific regulatory networks (Figs. 5F, 5G and S44). In these, we found a number of well-known TFs associated with brain development -- e.g., NEUROG1, DLGAP2, and MEF2A for excitatory neurons and GAD1, GAD2, and LHX6 for inhibitory neurons (Fig. 5G) (*42-45*). Finally, for broad utility, on the resource website, we also provide an expanded regulatory network with slightly different parameterization (Fig. S42).

Linking GWAS variants to genes

We used our regulatory network based on Hi-C, QTLs, and activity relationships to connect non-coding GWAS loci to potential disease genes. In particular, for the 142 SCZ GWAS loci, we identified a set of 1,111 putative SCZ-associated genes, covering 119 loci (the "SCZ-genes," Fig. 6A) (*46*). 321 of these constitute a “high-confidence” set supported by more than two

evidence sources (e.g., QTLs and Hi-C, Figs. 6A, 6B and S45); examples include CHRNA2 and CACNA1C (Fig. 6B-C). Overall, the SCZ-genes represent an increase from the 22 genes reported in an earlier QTL study and a larger number than can be linked simply by genomic proximity (176, Fig. 6A) (*11, 46*). In fact, the majority of SCZ-genes were not even in LD with the index SNPs (~67% or 748/1,111 genes with r^2<0.6, Fig. S45), consistent with the fact that regulatory relationships often do not follow linear genome organization (*13*).

We then looked at the characteristics of the 1,111 SCZ-genes (and high-confidence subset of 321). As expected, they shared many characteristics with known SCZ-associated genes, being enriched in translational regulators, cholinergic receptors, calcium channels, synaptic genes, SCZ differentially expressed genes, and loss-of-function intolerant genes (Fig. S45) (*46*). Next, we identified the TFs regulating the SCZ-genes (based on our regulatory network, either directly or via an enhancer; Fig. 6D). These include LHX9 and SOX7, transcription factors critical for early cortical specification and neuronal apoptosis, respectively (*47, 48*). Finally, we integrated the SCZ-genes with single-cell profiles and found that they are highly expressed in neurons, particularly excitatory ones, consistent with the recent findings (Fig. 6E) (*46*).

In addition to SCZ, we also looked at other diseases, linked by our regulatory network. In particular, we found aggregate associations between our brain eQTLs and enhancers and many brain-disorder GWAS variants, much more so than for GWAS variants for non-brain diseases (Fig. 6F, Table S9).

Integrative deep-learning model

The full interaction between genotype and phenotype involves many levels, beyond those encapsulated by the regulatory network. We addressed this by embedding our regulatory network into a larger multilevel model. In particular, we developed an interpretable deep-learning framework, the Deep Structured Phenotype Network (DSPN) (*20*). This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from the regulatory network (*49*). Traditional classification methods such as logistic regression predict phenotype directly from genotype, without using intermediates such as the transcriptome (Fig. 7A). In contrast, the DSPN is constructed via a series of intermediate models that add layers of structure. We included layers for intermediate molecular phenotypes associated with specific genes (i.e., their gene expression and chromatin state) and pre-defined gene groupings (cell-type marker genes and co-expression modules), multiple higher layers for inferred groupings (hidden nodes), and a top layer for observed traits (psychiatric disorders and other brain phenotypes). Finally, we used sparse inter- and intra-level connectivity to integrate our knowledge of QTLs, regulatory networks, and co-expression modules from the sections above (Fig. 7B). 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 traits.

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 trait prediction than traditional additive models (Fig. 7C). For instance, a logistic predictor was able to gain a 2.4X improvement when including the transcriptome vs. using the genome alone

(+9.3% for transcriptome vs. +3.8% for the genome, above a 50% random baseline). In contrast, 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. This result clearly manifests that the transcriptome carries additional information, which the DSPN is able to extract. Moreover, the DSPN allows us to perform joint inference and imputation of intermediate phenotypes (i.e., transcriptome and epigenome) and observed traits from just the genotype alone, achieving a ~3.1X improvement over a logistic predictor in this context (Figs. 7C and S46). Overall, 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 on samples for which intermediate phenotypes are imputed.

We transformed our results to the liability scale for comparison with narrow-sense heritability estimates (Fig. 7C) (*20*). Prior studies have estimated that common SNPs explain 25.6%, 20.5%, and 19% of the genetic variance for SCZ, BPD and ASD, respectively (*50*). These may be taken as theoretical upper bounds for additive models, given unlimited common-variant data. By contrast, non-linear predictors can exceed these limits. Our best liability scores (from just the genotype at QTL-associated variants) are substantially below these bounds, implying that additional data would be beneficial. In contrast, the variance explained by the full DSPN model exceeds that explained by common SNPs in SCZ and BPD, possibly reflecting the influence of rare variants and epistatic interactions (32.8% and 37.4% respectively -- the variance of 11.3% for ASD is slightly lower). Note, however, these estimates may be confounded by trait-associated variation which is environmental in origin (Fig. S47).

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. Here, we included co-expression modules in the model, referring to this modification as "DSPN-mod" (Fig. S48). Using it, we determined which modules were prioritized, as well as the sets of genes associated with latent nodes that were found at each hidden layer (Fig. 8A and Table S10; *15* (*20*)). Broadly, we take an unbiased view of all 5,024 modules and higher-order groupings constructed from these and then prioritize a subset of ~180 modules and groupings for each psychiatric disorder, showing these to being enriched in specific functional categories and to intersect substantially with the modules from more disease-focused analyses (Figs. 8B-C, S49) (*21*). (For completeness, we provide a full table showing the prioritization and functional categories for all possible modules associated with various traits (Fig. S50).) In particular, we found that cross-disorder prioritized modules are associated with functional categories such as “immune processes”, “synaptic activity” and “splicing”, consistent with the findings from more disease-focused analyses (Fig. 8C) (*21*). Also, we showed that prioritized SCZ and BPD modules are enriched for known GWAS SNPs (Fig. S51, for ASD, the lack of GWAS SNPs precludes similar analyses). For SCZ, which is the best characterized of the three disorders, we find enrichments for pathways and genes known to be associated with the disease, including: (i) glutamatergic-synapse pathway genes, such as GRIN1, (ii) calcium-signaling pathway and astrocyte-marker genes, and (iii) complement cascade pathway genes such as C4A, C4B, and CLU (Fig. 8D) (*21*). Other prioritized modules include well-characterized genes such as MIAT,

RBFOX1 and ANK2 (SCZ), RELA, NFkB2 and NIPBL (ASD) and HOMER1 (BPD), consistent with the results of (*21*). Finally, we identify modules associated with aging, finding that they are enriched in Ex4 neuronal cell-type genes, synaptic and longevity functions, and the gene NRGN -- all consistent with differential expression analysis (Fig. 8D and S20).

Conclusion

We have developed a comprehensive resource for functional genomics of the adult brain by integrating PsychENCODE data with a broad range of publicly available datasets. In closing, we review our main findings and ways that they can be improved in the future.

First, in terms of QTLs, we identified a set of eQTLs several fold larger than previous studies, targeting a saturating proportion of protein-coding genes. Moreover, we were able to identify a substantial number of cQTLs. PsychENCODE was, in fact, among the first efforts to generate ChIP-seq data across a large cohort of brain samples, with experiments primarily focused on H3K27ac. In the future, further increasing cohort size and performing additional chromatin assays, such as STARR-seq and ChIP-seq for other histone modifications, will improve the identification of enhancers and cQTLs (*51*). More fundamentally, one-dimensional fluctuations in chromatin signal reflect changes in three-dimensional chromatin architecture and new metrics beyond cQTLs may be needed.

Second, in terms of single-cell analysis, we found varying proportions of basic cell types (with different expression signatures) accounted for a large fraction of the expression variation across a population of individuals. However, this assumes that the expression levels characterizing a signature are fairly constant over a population of cells of a given cell type. In the future, larger-scale single-cell studies will allow us to examine this question in detail, perhaps quantifying and bounding environment-associated transcriptional variability. In addition, current single-cell techniques suffer from low sensitivity and dropouts; thus, it remains challenging to reliably quantify low-abundance transcripts (*15, 52*). This is particularly the case for specific brain cell sub-structures, such as axons and dendrites (*15*).

Third, we developed a comprehensive deep-learning model, the DSPN, and used it to illustrate how functional genomics data could improve the link between genotype and phenotype. In particular, by integrating regulatory-network connectivity and latent factors, the DSPN improves trait prediction over traditional additive models. Moreover, it takes into account dependencies between gene expression levels not modeled by univariate eQTL methods. Note, here we kept our eQTL methods very standard, closely following the GTEx paradigm. This separation we make between univariate eQTL detection and multivariate integrative modeling allows us to compare our eQTLs directly to previous analyses such as CMC. However, multivariate-based methods for QTLs have been used elsewhere and, in the future, may be combined with our approach (*53, 54*).

Further, in the future, we can envision how our DSPN approach can be extended to modeling additional intermediate phenotypes. In particular, we can naturally embed in the middle levels of

the model additional types of QTLs and phenotype-phenotype interactions - e.g., QTLs associated with miRNAs, neuroimaging, human/primate specific genes and developmental brain enhancers (*55-58*).

We expect that the DSPN will improve accuracy mainly for complex traits with a highly polygenic architecture, but not necessarily for traits that are strongly determined by only a few variants, such as Mendelian disorders, or closely correlated with population structure, such as ethnicity. However, even when the DSPN performance is low, it may still provide insights about intermediate phenotypes; for instance, in our analysis the PFC transcriptome appears substantially less predictive with respect to gender (after removing the sex chromosome genes) than age, but this very fact highlights the similarity of the transcriptome between sexes (*59*). Finally, although our focus has been on common SNPs, the DSPN may be able to capture the effects of rare variants, such as those known to be implicated in ASD (*50*), through their influence on intermediate phenotypes.

In summary, our integrative analyses demonstrate the usefulness of functional genomics for unraveling molecular mechanisms in the brain (*60, 61*) and the result of these analyses suggest directions for further research into the etiology of brain disorders.

Materials and Methods Summary

The materials and methods for each section of main text are available in the section with same heading of the supplement (*20*); i.e., supplementary content for a given main text section within the supplement is named in a parallel fashion. Detailed data protocols are available in the supplement. Moreover, associated and derived data files are available on the resource website, Resource.psychENCODE.org. Often we provide multiple versions of the derived summary files with different parameterizations (e.g., for the single cell profiles and for eQTLs).

ACKNOWLEDGMENTS

Funding: Data were generated as part of the PsychENCODE Consortium, supported by: U01MH103339, U01MH103365, U01MH103392, U01MH103340, U01MH103346, U01MH116492, R01MH105472, R01MH094714, R01MH105898, R01MH110920, R01MH110905, R01MH110926, R01MH110927, R01MH110928, R01MH109715, R01MH111721, R21MH102791, R01MH110921, R01MH109677, R21MH105881, R21MH103877, R21MH109956, R21MH105853, and P50MH106934 awarded to: Schahram Akbarian (Icahn School of Medicine at Mount Sinai), Andrew Chess (Icahn School of Medicine at Mount Sinai), Gregory E. Crawford (Duke University), Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (University of Southern California), Zhiping Weng (UMass Medical School), Mark Gerstein (Yale University), Daniel H. Geschwind (University of California, Los Angeles), Fernando Goes (Johns Hopkins University), Thomas M. Hyde (Lieber Institute for Brain Development), Andrew Jaffe (Lieber Institute for Brain Development), James A. Knowles (SUNY Downstate Medical Center), Chunyu Liu (SUNY Upstate Medical University), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Panos Roussos (Icahn School of Medicine at Mount Sinai), Nenad Sestan (Yale University), Pamela Sklar (Icahn School of

Medicine at Mount Sinai), Matthew State (University of California, San Francisco), Patrick Sullivan (University of North Carolina), Flora Vaccarino (Yale University), Daniel Weinberger (Lieber Institute for Brain Development), Sherman Weissman (Yale University), Kevin White (University of Chicago), A. Jeremy Willsey (University of California, San Francisco) and Peter Zandi (Johns Hopkins University), Lara Mangravite, Mette Peters (Sage Bionetworks), Alexander Arguello, Lora Bingaman, Thomas Lehner, David Panchision, Geetha Senthil (NIMH). A subset of the transcriptome (RNA-seq) data were generated as part of the CommonMind Consortium supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership: Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals Company Limited), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche Ltd), Lara Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH). Zeynep Hulya Gumus is supported by NIH/NIAID grant U19AI118610. **Author contributions**: All the named authors contributed significantly to the paper either in data generation or analysis: data generation - H.W., M.J.G., Y.J., G.E.H, G.E.C., P.R., S.A., A.E.J., K.P.W., N.S.,D.H.G.; data analysis - D.W., S.L., J.W., H.W., X.S., F.C.P.N., D.C., M.G., P.E., Y.T.Y., M.X., M.J.G., S.K.L., J.Z., J.J.P., C.Y., S.K.R., K.M., H.Z., A.N., M.P., E.M., D.F., T.B., J.M., K.G., S.K., Z.H.G., G.E.C., P.R., S.A., A.E.J., K.P.W., Z.W., N.S., D.H.G., J.A.K., M.B.G. All the starred first authors contributed analysis efforts equally. The three corresponding authors co-led the analysis. **Competing interests**: G.E.C. is a co-founder of Element Genomics, Inc. K.P.W. is associated with Tempus Labs Inc. The rest of the authors declare no competing interests. **Data and materials availability**: all data are available in the manuscript or the supplementary material.

Figures

Figure 1. Comprehensive data resource for functional genomics of the human brain.

The functional genomics data generated by the PsychENCODE consortium constitute a multidimensional exploration across tissue, developmental stage, disorder, species, assay, and sex. The central data cube represents the results of our data integration for the three dimensions of disorder, assay, and tissue, where the numbers of datasets in the analysis are depicted. Projections of the data onto each of these three parameters are shown as graphs for assay and disorder, and as a schematic for the primary brain regions of interest. **Assay:** Dataset numbers for a subset of assays are shown, including RNA-seq (2040 PsychENCODE + 1632 GTEx, used in multiple downstream analyses), genotypes (1362 PsychENCODE + 25

GTEx = 1387 individuals matched to RNA-seq samples for QTL analysis after QC-filtering), and H3K27ac ChIP-seq (408 PsychENCODE + 5 Roadmap). The number of cells assayed by scRNA-seq (right-hand y-axis) = 18025 PsychENCODE + 14012 external datasets. **Disorder:** Across all assays, there are 113 GTEx + 926 PsychENCODE control individuals, and 558 SCZ, 217 BPD, 44 ASD and 8 AFF individuals from the PsychENCODE, resulting in 1,866 individuals. Tissue: Three brain regions are considered: the prefrontal cortex (PFC, N = 26,769), temporal cortex (TC, N = 2,153), and cerebellum (CB, N = 348). See Table S11 and Resource.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 (See Fig. S52). 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**. The three major cell types analyzed are depicted with 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; *15*). **(D)** The estimated cell fractions can account >88% of the bulk tissue expression variation across the population. **(E)** Cell fraction changes across genders and brain disorders. (Differences significant (via KS-test) compared to control samples after accounting for age distributions are labeled (**)). See Table S12 for more detail. **(F)** Changing cell fractions (for Ex3), gene expression (for SST) and promoter methylation level (median level, for SST) across age groups are shown. With increasing age, the fractions of Ex3 and Ex4 significantly increase and some non-neuronal types decrease (Ex3 trend analysis, p<6.3e-10).

Figure 3. Comparative analysis of transcriptomics and epigenomics between brain and other tissues.

(A) Epigenetics signals of the reference brain (purple) were used to identify active enhancers using the ENCODE enhancer pipeline. The H3K27ac signal tracks at the corresponding enhancer region from each individual in the cohort are shown in green. with the gradient showing the normalized signal value for each H3K27ac peak **(B)** The overlap of the H3K27ac peaks from an individual in the population with the reference brain enhancers is shown as the Venn diagram. The histogram shows the varying percentage of overlapped 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), external 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). **(D)** The extent of transcription for coding (arrowhead) and non-coding (diamond) regions. Average transcription extent (x-axis) is shown compared to the cumulative extent of transcription across a cohort of individuals (y-axis) for select tissue types including cerebellum, cortex, lung, skin, and testis, using PolyA RNA-seq data. Finally, *Panels E and F are drawn similarly to C, but now for transcription rather than epigenetics*. **(E)** RCA coefficients for gene expression data of PsychENCODE, GTEx brains, and other tissue samples are shown in dark green, light green and in magenta, respectively. **(F)** The center (cross) and ranges of different tissue clusters (dashed ellipses) are shown on an RCA scatterplot of (**E**).

Figure 4. QTLs in the adult brain.

(A) Frequency of genes with at least one eQTL (eGenes) are shown across different studies. The number of eGenes increased as the sample size increased. PsychENCODE eGenes are close to saturation for protein coding genes. The estimated replication π_1 values of GTEx and CMC eQTLs versus PsychENCODE are shown (*35*). **(B)** The similarity between PsychENCODE brain DLPFC eQTLs and GTEx eQTLs of other tissues are evaluated by π_1 values and SNP-eGene overlap rates. Both π_1 values and SNP-eGene overlap rates are higher in brain DLPFC than the other tissues. **(C)** An example of an H3K27ac signal across individuals in a representative genomic region showing largely congruent identification of regions of open chromatin. The region in the dashed frame represents a cQTL; the signal magnitudes of individuals with a G/G or G/T genotype were lower than the ones with a T/T genotype. **(D)** An example of the mechanism by which an fQTL may work to impact phenotype. This fQTL overlaps with an eQTL for FZD9, a gene located in the 7q11.23 region that is deleted in Williams syndrome. The fQTL may affect the fraction of Ex3 through regulating FZD9 expression. Note that only Ex3 constitutes a statistically significant fQTL with this SNP (as designated by the asterisk). **(E)** The enrichment of QTLs in different genomic annotations are shown. Pink circles indicate highly significant enrichment (p<1e-25). **(F)** Numbers of identified QTLs associated elements (eGenes, enhancers, and cell types) and QTL SNPs are shown in the bottom left table. Asterisks (*) indicate that, for cQTLs, we only show the number of top SNPs for each enhancer. Overlaps of all QTL SNPs are shown in heatmaps (square rows). The linked circles show the overlap of QTL types. The intersections of other QTLs with eQTLs are evaluated using π_1 values in the orange bar plot. The greatest intersection is between cQTLs and eQTLs. An example is displayed on the right: the intersection of eQTL SNPs (for the MTOR gene) and cQTL SNPs (for the H3K27ac signal on an enhancer ~50kb upstream of the gene). Hi-C interactions (bottom) indicate that the enhancer interacts with the promoter of MTOR, suggesting that the cQTL SNPs potentially mediate the expression modulation manifest by the eQTL SNPs.

Figure 5. Building a gene regulatory network from Hi-C and data integration.

(A) A full Hi-C data from adult brain reveals the higher-order structure of the genome, ranging from contact maps (top), TADs, and promoter-based interactions. Bottom shows a schematic of how we leveraged gene regulatory linkages involving TADs, TFs, enhancers, and target genes

to build a full gene regulatory network (Fig. S42) and a high confidence subnetwork consisting of 43,181 TF-to-target-promoter and 42,681 enhancer-to-target-promoter linkages (*20*). **(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 (174,719) showed more significant P-values than those that were not (promoter/exonic QTLs, 130,155; non-supported QTLs, 1,065,311). **(D)** Cross-tissue comparison of chromatin architecture indicates that adult brains in PsychENCODE and Roadmap (e.g. DLPFC, Hippocampus) share chromatin architecture more than non-related tissue types. Fetal brain shows distinct chromatin architecture to adult brain, indicating extensive rewiring of chromatin structures during brain development. **(E)** Genes assigned to fetal active elements are prenatally enriched, while genes assigned to adult active elements are postnatally enriched. **(F)** Genes assigned to fetal active elements are relatively more enriched in neurons in the adult (Adult-Neuron) and fetal brain (Development), while genes assigned to adult active elements are relatively more enriched in glia (Adult-astrocytes, endothelial cells, and oligodendrocytes). **(G)** The circos plots show the linkages from the full regulatory network targeting the cell-type-specific biomarker genes. The biomarker genes for excitatory/inhibitory neuronal type are the shared biomarker genes by at least five excitatory/inhibitory subtypes (*19*). Selected TFs for particular cell types are highlighted.

Figure 6. Gene regulatory networks assign genes to GWAS loci for psychiatric disorders.

(A) A schematic depicting how SCZ GWAS loci were assigned to putative genes. The number of SCZ GWAS loci and their putative target genes (SCZ-genes) annotated by each assignment strategy is described (top). The overlap between SCZ-genes defined by QTL associations (QTL), chromatin interactions (Hi-C), and activity relationships (Activity) is depicted in a Venn diagram (bottom). SCZ-genes with more than 2 evidence sources were defined as high-confidence (high conf.) genes. **(B)** A gene regulatory network of TFs, enhancers, and 321 SCZ high-confidence genes, on the basis of TF activity linkages. A subnetwork for *CACNA1C* is highlighted on the right. **(C)** An example of the evidence depicting that GWAS SNPs that overlap with *CHRNA2* eQTLs also have chromatin interactions and activity correlations with the same gene. Orange dots refer to SNPs that overlap between eQTLs and GWAS plots. **(D)** TFs that are significantly enriched in enhancers (left) and promoters (right) of SCZ-genes. **(E)** SCZ-genes show higher expression levels in neurons (particularly excitatory neurons) than other cell types. **(F)** 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 7. DSPN deep-learning model links genetic variation to psychiatric disorders and other traits.

(A) The schematic outlines the structure of the following models: Logistic Regression (LR), conditional Restricted Boltzmann Machine (cRBM), conditional Deep Boltzmann Machine (cDBM), and Deep Structured Phenotype Network (DSPN). Nodes are partitioned into four layers (L0-L3) and colored according to their status as visible, visible or imputed (depending on whether observed or not at test time) or hidden. **(B)** 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. 5A) is embedded in layers L0 and L1, with different types of regulatory linkages and functional elements shown. **(C)** The performance of different models is summarized, comparing performance (i) across models of different complexity, and (ii) transcriptome vs. genome predictors, corresponding to with/without imputation for the DSPN (colors highlight relevant models for each comparison). Performance accuracy is shown first, with variance explained on the liability scale in brackets. All models were tested on identical data splits, which were balanced for predicted trait and covariates (including gender, ethnicity, age and assay). RNA-seq, cell fraction, H3k27ac data were binarized by thresholding at median values (per gene, cell-type, enhancer respectively), as was age (median 51 years) when predicted. LR-gene and LR-trans are logistic models using the genotype and transcriptome as predictors respectively; DSPN-impute and DSPN-full are models with imputed intermediate phenotypes (genotype predictors only) and fully observed intermediate phenotypes (transcriptome predictors) respectively. Differential performance is shown in terms of improvement above chance, with liability variance score increases in brackets. Abbreviations as in main text, with GEN=Gender, ETH=Ethnicity, AOD=Age of death.

Figure 8. Interpretation of the DSPN model highlights functional associations and shared disease mechanisms.

(A) Schematic illustrates module (MOD) and higher-order grouping (HOG) prioritization scheme. Red and blue lines represent positive and negative weights respectively, and full and dotted lines represent first and second ranks by absolute value (creating a DAG with branching factor 4, rooted at L3). Highlighted nodes (grey) in L1d show positive prioritized MODs, for which a positive path (containing an even number of negative links) exists connecting module to SCZ node. $\mathbf{a_1}$/$\mathbf{a_2}$ and $\mathbf{b_1}$/$\mathbf{b_2}$ highlight 'best positive paths' from **a** and **b** respectively to SCZ in terms of absolute rank score. Associated HOGs are defined for $\mathbf{a_1}$/$\mathbf{a_2}$, containing all nodes in L1d having a path in the DAG to $\mathbf{a_1}$ (resp. $\mathbf{a_2}$) which is identically signed to the best path from **a** to $\mathbf{a_1}$ (resp. $\mathbf{a_2}$) (*20*). Positive prioritized HOGs are associated with nodes on best paths from all positive prioritized MODs; negative prioritized MODs/HOGs are calculated similarly. **(B)** Panel summarizes functional annotation scheme: (i) 5024 WGCNA MODs (modules/submodules) are derived from multiple data splits.(ii) MODs are prioritized as in (A) for each disorder, and (iii) associated HOGs are calculated. (iv) Gene set enrichment analysis associates functional terms with all MODs/HOGs. (v) Terms are ranked per disorder by counting the number of prioritized MODs/HOGs they associate with, and broad functional categories are defined; (vi) prioritized MODs/HOGs are

linked to potentially interesting genes, enhancers and SNPs using GRN connectivity. **(C)** Chart shows upper segment of cross-disorder ranking of GO/KEGG functional terms, where cross-disorder ranks are assigned using the average per-disorder rank ordering. Ranking score levels and functional categories are as in the key in (B). Highlighted ranks and terms correspond to examples shown in (D). See Fig. S49 for extended ranking. **(D)** shows examples of associations between prioritized MODs/HOGs and genes, enhancers and SNPs for each disorder and age model. Associated functional terms and categories are as in (B). A table providing coordinates of eQTLs and cQTLs for all examples shown is provided in Table S13.

References and Notes

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. D. H. Geschwind, J. Flint, Genetics and genomics of psychiatric disease. *Science* **349**, 1489-1494 (2015).
7. G. C. C. C. Psychiatric *et al.*, Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry* **166**, 540-556 (2009).
8. G. T. Consortium *et al.*, Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
9. C. Roadmap Epigenomics *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
10. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
11. M. Fromer *et al.*, Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).
12. C. Colantuoni *et al.*, Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**, 519-523 (2011).
13. H. Won *et al.*, Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527 (2016).
14. B. B. Lake *et al.*, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590 (2016).
15. 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).
16. 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).
17. E. C. Psych *et al.*, The PsychENCODE project. *Nat Neurosci* **18**, 1707-1712 (2015).
18. J. T. Walters, M. J. Owen, Endophenotypes in psychiatric genetics. *Mol Psychiatry* **12**, 886-890 (2007).
19. B. B. Lake *et al.*, Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).

20. Materials and methods are available as supplementary materials.
21. M. J. Gandal, e. al., Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **in revision**.
22. I. Voineagu *et al.*, Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384 (2011).
23. M. J. Gandal *et al.*, Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693-697 (2018).
24. M. C. Oldham *et al.*, Functional organization of the transcriptome in human brain. *Nat Neurosci* **11**, 1271-1282 (2008).
25. T. E. Bakken *et al.*, A comprehensive transcriptional map of primate brain development. *Nature* **535**, 367-375 (2016).
26. 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).
27. K. W. Kelley, H. Nakao-Inoue, A. V. Molofsky, M. C. Oldham, Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat Neurosci* **21**, 1171-1184 (2018).
28. J. L. Rubenstein, M. M. Merzenich, Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes Brain Behav* **2**, 255-267 (2003).
29. B. C. McKinney *et al.*, Hypermethylation of BDNF and SST Genes in the Orbital Frontal Cortex of Older Individuals: A Putative Mechanism for Declining Gene Expression with Age. *Neuropsychopharmacology* **40**, 2604-2613 (2015).
30. R. Tacutu *et al.*, Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res* **46**, D1083-D1090 (2018).
31. M. Kasowski *et al.*, Extensive variation in chromatin states across humans. *Science* **342**, 750-752 (2013).
32. W. Sun *et al.*, Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* **167**, 1385-1397 e1311 (2016).
33. D. Purves, *Neuroscience*. (Oxford University Press, New York, ed. Sixth edition., 2018), pp. 1 volume (various pagings).
34. G. T. Consortium, Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).
35. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445 (2003).
36. R. C. del Rosario *et al.*, Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nat Methods* **12**, 458-464 (2015).
37. F. Grubert *et al.*, Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-1065 (2015).
38. J. Bryois *et al.*, Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nat Commun* **9**, 3121 (2018).
39. T. Chailangkarn *et al.*, A human neurodevelopmental model for Williams syndrome. *Nature* **536**, 338-343 (2016).
40. L. T. M. Dao *et al.*, Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet* **49**, 1073-1081 (2017).
41. L. de la Torre-Ubieta, H. Won, J. L. Stein, D. H. Geschwind, Advancing the understanding of autism disease mechanisms through genetics. *Nat Med* **22**, 345-361 (2016).
42. C. Fode *et al.*, A role for neural determination genes in specifying the dorsoventral identity of telencephalic neurons. *Genes Dev* **14**, 67-80 (2000).
43. A. H. Rasmussen, H. B. Rasmussen, A. Silahtaroglu, The DLGAP family: neuronal expression, function and role in brain disorders. *Mol Brain* **10**, 43 (2017).

44. M. G. Erlander, N. J. Tillakaratne, S. Feldblum, N. Patel, A. J. Tobin, Two genes encode distinct glutamate decarboxylases. *Neuron* **7**, 91-100 (1991).
45. P. Liodis *et al.*, Lhx6 activity is required for the normal migration and specification of cortical interneuron subtypes. *J Neurosci* **27**, 3078-3089 (2007).
46. 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).
47. V. S. Mangale *et al.*, Lhx2 selector activity specifies cortical identity and suppresses hippocampal organizer fate. *Science* **319**, 304-309 (2008).
48. C. Wang *et al.*, SOX7 interferes with beta-catenin activity to promote neuronal apoptosis. *Eur J Neurosci* **41**, 1430-1437 (2015).
49. R. Salakhutdinov, G. Hinton, Deep Boltzmann Machines. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* **5**, 448--455 (2009).
50. C. Brainstorm *et al.*, Analysis of shared heritability in common disorders of the brain. *Science* **360**, (2018).
51. Y. Liu *et al.*, Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol* **18**, 219 (2017).
52. S. Liu, C. Trapnell, Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5**, (2016).
53. H. Chun, S. Keles, Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics* **182**, 79-90 (2009).
54. M. P. Scott-Boyer *et al.*, An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Stat Appl Genet Mol Biol* **11**, (2012).
55. C. E. Bearden, P. M. Thompson, Emerging Global Initiatives in Neurogenetics: The Enhancing Neuroimaging Genetics through Meta-analysis (ENIGMA) Consortium. *Neuron* **94**, 232-236 (2017).
56. T. G. M. van Erp *et al.*, Cortical Brain Abnormalities in 4474 Individuals With Schizophrenia and 5098 Control Subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium. *Biol Psychiatry*, (2018).
57. A. M. M. Sousa *et al.*, Molecular and cellular reorganization of neural circuits in the human lineage. *Science* **358**, 1027-1032 (2017).
58. A. A. e. al., Integrative multi-omics analyses of iPSC-derived brain organoids identify early determinants of human cortical development. *Science* **in revision**.
59. O. V. Evgrafov *et al.*, Gene expression in patient-derived neural progenitors provide insights into neurodevelopmental aspects of schizophrenia. *bioRxiv*, (2017).
60. M. J. Gandal, Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *revision*.
61. M. e. a. Li, Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risk. *Science* **in revision**.
62. G. T. Consortium *et al.*, Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
63. B. B. Lake *et al.*, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590 (2016).
64. B. B. Lake *et al.*, Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).
65. 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).
66. D. J. McCarthy, K. R. Campbell, A. T. Lun, Q. F. Wills, Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179-1186 (2017).

67. D. van Dijk *et al.*, MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv*, (2017).
68. H. G. Maaten LV, Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
69. J. P. Brunet, P. Tamayo, T. R. Golub, J. P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**, 4164-4169 (2004).
70. M. C. Oldham *et al.*, Functional organization of the transcriptome in human brain. *Nat Neurosci* **11**, 1271-1282 (2008).
71. M. J. Gandal, e. al., Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **in revision**.
72. M. J. Gandal *et al.*, Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693-697 (2018).
73. A. M. Newman *et al.*, Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015).
74. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17 (2005).
75. P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719-720 (2008).
76. A. E. Jaffe *et al.*, Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* **19**, 40-47 (2016).
77. D. Purves, *Neuroscience*. (Oxford University Press, New York, ed. Sixth edition., 2018), pp. 1 volume (various pagings).
78. H. J. Kang *et al.*, Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011).
79. C. Roadmap Epigenomics *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
80. I. M. Johnstone, A. Y. Lu, On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *J Am Stat Assoc* **104**, 682-693 (2009).
81. H. Li *et al.*, Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* **49**, 708-718 (2017).
82. O. Delaneau *et al.*, A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8**, 15452 (2017).
83. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445 (2003).
84. B. Ng *et al.*, An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat Neurosci* **20**, 1418-1426 (2017).
85. H. Chun, S. Keles, Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics* **182**, 79-90 (2009).
86. M. P. Scott-Boyer *et al.*, An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Stat Appl Genet Mol Biol* **11**, (2012).
87. H. Won *et al.*, Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527 (2016).
88. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
89. C. Y. McLean *et al.*, GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).
90. A. D. Schmitt *et al.*, A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* **17**, 2042-2059 (2016).
91. J. R. Dixon *et al.*, Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).

92. K. K. Yan, G. G. Yardimci, C. Yan, W. S. Noble, M. Gerstein, HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* **33**, 2199-2201 (2017).
93. M. Li, e. al., Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risk. *submitted*.
94. M. T. Weirauch *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
95. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).
96. A. F. A. Smit, Hubley, R. and Green, P. (1996-2010).
97. 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).
98. C. Giambartolomei *et al.*, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
99. M. Fromer *et al.*, Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).
100. T. Singh *et al.*, Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci* **19**, 571-577 (2016).
101. C. R. Marshall *et al.*, Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* **49**, 27-35 (2017).
102. H. K. Finucane *et al.*, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-1235 (2015).
103. V. Mnih, H. Larochelle, G. E. Hinton, Conditional Restricted Boltzmann Machines for Structured Output Prediction. *https://arxiv.org/abs/1202.3748*, (2012).
104. R. Salakhutdinov, G. Hinton, Deep Boltzmann Machines. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* **5**, 448--455 (2009).
105. D. Koller, N. Friedman, *Probabilistic graphical models : principles and techniques*. Adaptive computation and machine learning (MIT Press, Cambridge, MA, 2009), pp. xxi, 1231 p.
106. C. International Schizophrenia *et al.*, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).
107. D. S. a. M. T. F. C. Falconer, *Introduction to Quantitative Genetics, Ed 4*. (Longmans Green, Harlow, Essex, UK, 1996).
108. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).
109. K. Simonyan, D Phil, University of Oxford, (2013).
110. G. P. Shrikumar A, Kundaje Learning important features through propagating activation differences. *arXiv* (2017).
111. D. Zhang *et al.*, Genetic control of individual differences in gene-specific methylation in human brain. *Am J Hum Genet* **86**, 411-419 (2010).
112. C. B. Pedersen *et al.*, The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol Psychiatry* **23**, 6-14 (2018).
113. M. Kundakovic *et al.*, Practical Guidelines for High-Resolution Epigenomic Profiling of Nucleosomal Histones in Postmortem Human Brain Tissue. *Biol Psychiatry* **81**, 162-170 (2017).
114. N. N. Parikshak *et al.*, Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423-427 (2016).
115. M. A. Quail *et al.*, A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-1010 (2008).

116. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
117. S. Das *et al.*, Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
118. C. Brainstorm *et al.*, Analysis of shared heritability in common disorders of the brain. *Science* **360**, (2018).
119. D. Demontis, 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 Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. *bioRxiv*, (2017).
120. J. Grove *et al.*, Common risk variants identified in autism spectrum disorder. *bioRxiv*, (2017).
121. D. Bipolar, d. r. v. e. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address, D. Bipolar, C. Schizophrenia Working Group of the Psychiatric Genomics, Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705-1715 e1716 (2018).
122. N. R. Wray *et al.*, Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668-681 (2018).
123. A. Okbay *et al.*, Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539-542 (2016).
124. J. E. Savage *et al.*, Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* **50**, 912-919 (2018).
125. J. C. Lambert *et al.*, Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-1458 (2013).
126. M. A. Nalls *et al.*, Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* **46**, 989-993 (2014).
127. A. P. Morris *et al.*, Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981-990 (2012).
128. H. Schunkert *et al.*, Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* **43**, 333-338 (2011).
129. J. Z. Liu *et al.*, Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).

‡The PsychENCODE Consortium:

Allison E Ashley-Koch, Duke University; Gregory E Crawford, Duke University; Melanie E Garrett, Duke University; Lingyun Song, Duke University; Alexias Safi, Duke University; Graham D Johnson, Duke University; Gregory A Wray, Duke University; Timothy E Reddy, Duke University; Fernando S Goes, Johns Hopkins University; Peter Zandi, Johns Hopkins University; Julien Bryois, Karolinska Institutet; Andrew E Jaffe, Lieber Institute for Brain Development; Amanda J Price, Lieber Institute for Brain Development; Nikolay A Ivanov, Lieber Institute for

Brain Development; Leonardo Collado-Torres, Lieber Institute for Brain Development; Thomas M Hyde, Lieber Institute for Brain Development; Emily E Burke, Lieber Institute for Brain Development; Joel E Kleiman, Lieber Institute for Brain Development; Ran Tao, Lieber Institute for Brain Development; Joo Heon Shin, Lieber Institute for Brain Development; Schahram Akbarian, Icahn School of Medicine at Mount Sinai; Kiran Girdhar, Icahn School of Medicine at Mount Sinai; Yan Jiang, Icahn School of Medicine at Mount Sinai; Marija Kundakovic, Icahn School of Medicine at Mount Sinai; Leanne Brown, Icahn School of Medicine at Mount Sinai; Bibi S Kassim, Icahn School of Medicine at Mount Sinai; Royce B Park, Icahn School of Medicine at Mount Sinai; Jennifer R Wiseman, Icahn School of Medicine at Mount Sinai; Elizabeth Zharovsky, Icahn School of Medicine at Mount Sinai; Rivka Jacobov, Icahn School of Medicine at Mount Sinai; Olivia Devillers, Icahn School of Medicine at Mount Sinai; Elie Flatow, Icahn School of Medicine at Mount Sinai; Gabriel E Hoffman, Icahn School of Medicine at Mount Sinai; Barbara K Lipska, Human Brain Collection Core, National Institutes of Health, Bethesda, MD; David A Lewis, University of Pittsburgh; Vahram Haroutunian, Icahn School of Medicine at Mount Sinai and James J Peters VA Medical Center; Chang-Gyu Hahn, University of Pennsylvania; Alexander W Charney, Mount Sinai; Stella Dracheva, Mount Sinai; Alexey Kozlenkov, Mount Sinai; Judson Belmont, Icahn School of Medicine at Mount Sinai; Diane DelValle, Icahn School of Medicine at Mount Sinai; Nancy Francoeur, Icahn School of Medicine at Mount Sinai; Evi Hadjimichael, Icahn School of Medicine at Mount Sinai; Dalila Pinto, Icahn School of Medicine at Mount Sinai; Harm van Bakel, Icahn School of Medicine at Mount Sinai; Panos Roussos, Mount Sinai; John F Fullard, Mount Sinai; Jaroslav Bendl, Mount Sinai; Mads E Hauberg, Mount Sinai; Lara M Mangravite, Sage Bionetworks; Mette A Peters, Sage Bionetworks; Yooree Chae, Sage Bionetworks; Junmin Peng, St. Jude Children's Hospital; Mingming Niu, St. Jude Children's Hospital; Xusheng Wang, St. Jude Children's Hospital; Maree J Webster, Stanley Medical Research Institute; Thomas G Beach, Banner Sun Health Research Institute; Chao Chen, Central South University; Yi Jiang, Central South University; Rujia Dai, Central South University; Annie W Shieh, SUNY Upstate Medical University; Chunyu Liu, SUNY Upstate Medical University; Kay S. Grennan, SUNY Upstate Medical University; Yan Xia, SUNY Upstate Medical University/Central South University; Ramu Vadukapuram, SUNY Upstate Medical University; Yongjun Wang, Central South University; Dominic Fitzgerald, The University of Chicago; Lijun Cheng, The University of Chicago; Miguel Brown, The University of Chicago; Mimi Brown, The University of Chicago; Tonya Brunetti, The University of Chicago; Thomas Goodman, The University of Chicago; Majd Alsayed, The University of Chicago; Michael J Gandal, University of California, Los Angeles; Daniel H Geschwind, University of California, Los Angeles; Hyejung Won, University of California, Los Angeles; Damon Polioudakis, University of California, Los Angeles; Brie Wamsley, University of California, Los Angeles; Jiani Yin, University of California, Los Angeles; Tarik Hadzic, University of California, Los Angeles; Luis De La Torre Ubieta, UCLA; Vivek Swarup, University of California, Los Angeles; Stephan J Sanders, University of California, San Francisco; Matthew W State, University of California, San Francisco; Donna M Werling, University of California, San Francisco; Joon-Yong An, University of California, San Francisco; Brooke Sheppard, University of California, San Francisco; A Jeremy Willsey, University of California, San Francisco; Kevin P White, The University of Chicago; Mohana Ray, The University of Chicago; Gina Giase, SUNY Upstate Medical University; Amira Kefi, University of Illinois at Chicago; Eugenio Mattei, University of Massachusetts Medical School; Michael Purcaro, University of Massachusetts Medical School; Zhiping Weng, University of Massachusetts Medical School; Jill Moore, University of Massachusetts Medical School; Henry Pratt, University of Massachusetts Medical School; Jack Huey, University of Massachusetts Medical School; Tyler Borrman, University of Massachusetts Medical School; Patrick F Sullivan, University of North Carolina - Chapel Hill; Paola Giusti-Rodriguez, University of North Carolina - Chapel Hill; Yunjung Kim, University of North Carolina - Chapel Hill; Patrick Sullivan, University of North Carolina - Chapel Hill; Jin Szatkiewicz,

University of North Carolina - Chapel Hill; Suhn Kyong Rhie, University of Southern California; Christoper Armoskus, University of Southern California; Adrian Camarena, University of Southern California; Peggy J Farnham, University of Southern California; Valeria N Spitsyna, University of Southern California; Heather Witt, University of Southern California; Shannon Schreiner, University of Southern California; Oleg V Evgrafov, SUNY Downstate Medical Center; James A Knowles, SUNY Downstate Medical Center; Mark Gerstein, Yale University; Shuang Liu, Yale University; Daifeng Wang, Stony Brook University; Fabio C. P. Navarro, Yale University; Jonathan Warrell, Yale University; Declan Clarke, Yale University; Prashant S. Emani, Yale University; Mengting Gu, Yale University; Xu Shi, Yale University; Min Xu, Yale University; Yucheng T. Yang, Yale University; Robert R. Kitchen, Yale University; Gamze Gürsoy, Yale University; Jing Zhang, Yale University; Becky C Carlyle, Yale University; Angus C Nairn, Yale University; Mingfeng Li, Yale University; Sirisha Pochareddy, Yale University; Nenad Sestan, Yale University; Mario Skarica, Yale University; Zhen Li, Yale University; Andre M.M. Sousa, Yale University; Gabriel Santpere, Yale University; Jinmyung Choi, Yale University; Ying Zhu, Yale University; Tianliuyun Gao, Yale University; Daniel J Miller, Yale University; Adriana Cherskov, Yale University; Mo Yang, Yale University; Anahita Amiri, Yale University; Gianfilippo Coppola, Yale University; Jessica Mariani, Yale University; Soraya Scuderi, Yale University; Anna Szekely, Yale University; Flora M Vaccarino, Yale University; Feinan Wu, Yale University; Sherman Weissman, Yale University; Tanmoy Roychowdhury, Mayo Clinic Rochester; Alexej Abyzov, Mayo Clinic Rochester;

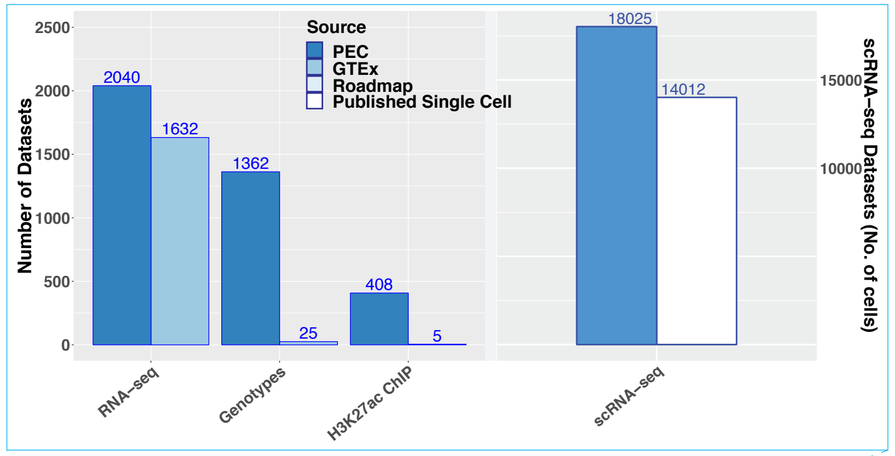
SUPPLEMENTARY MATERIALS

Supplementary Text

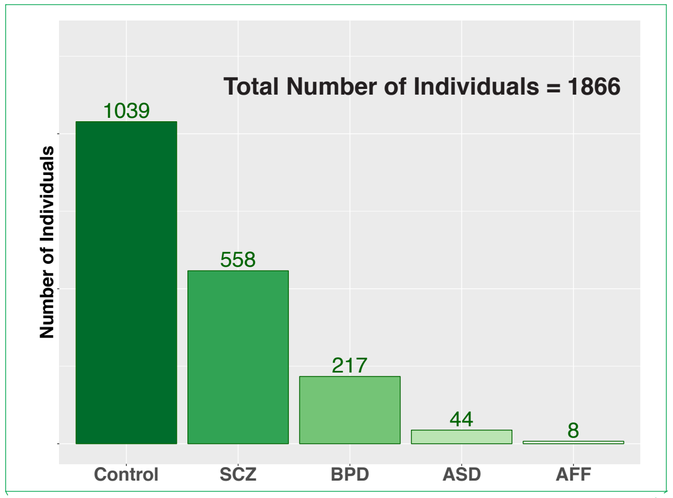
Figs. S1 to S52

Tables S1 to S13

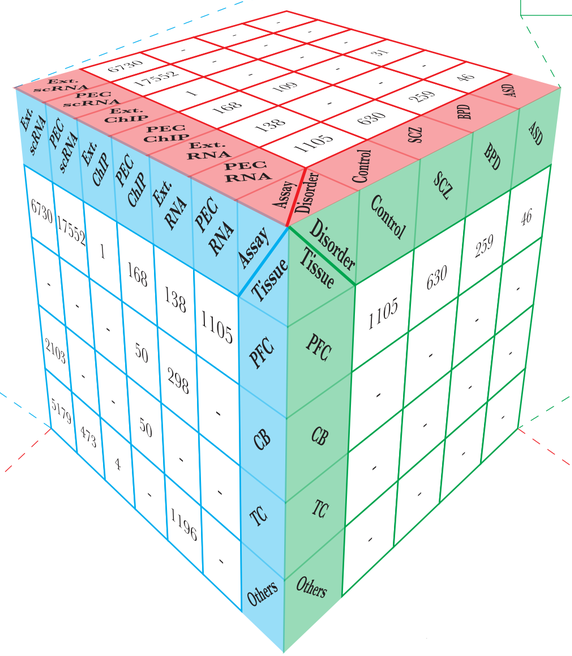
References (62–129)



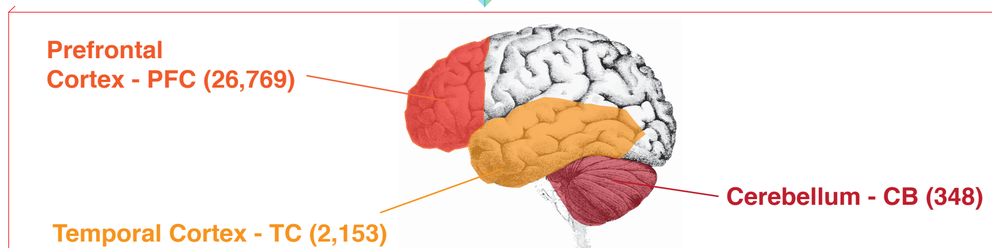
Assay

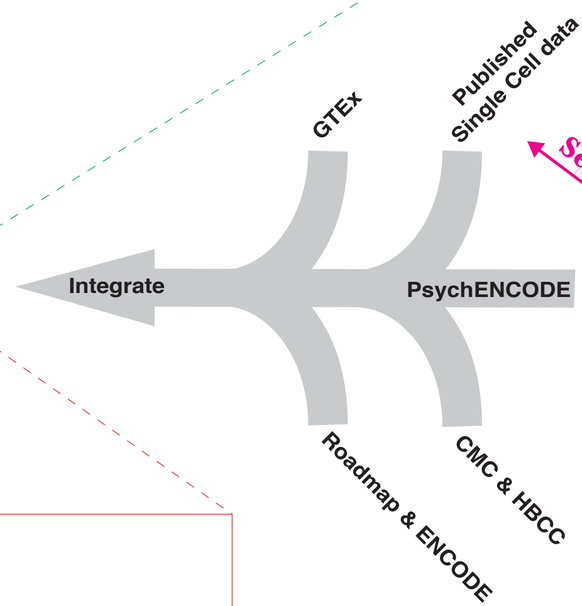


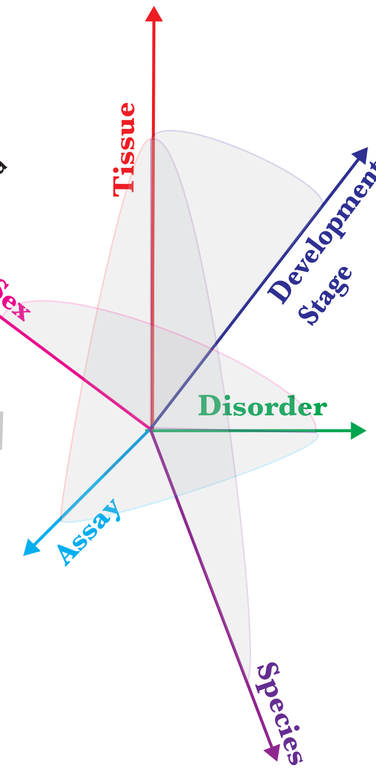
Disorder

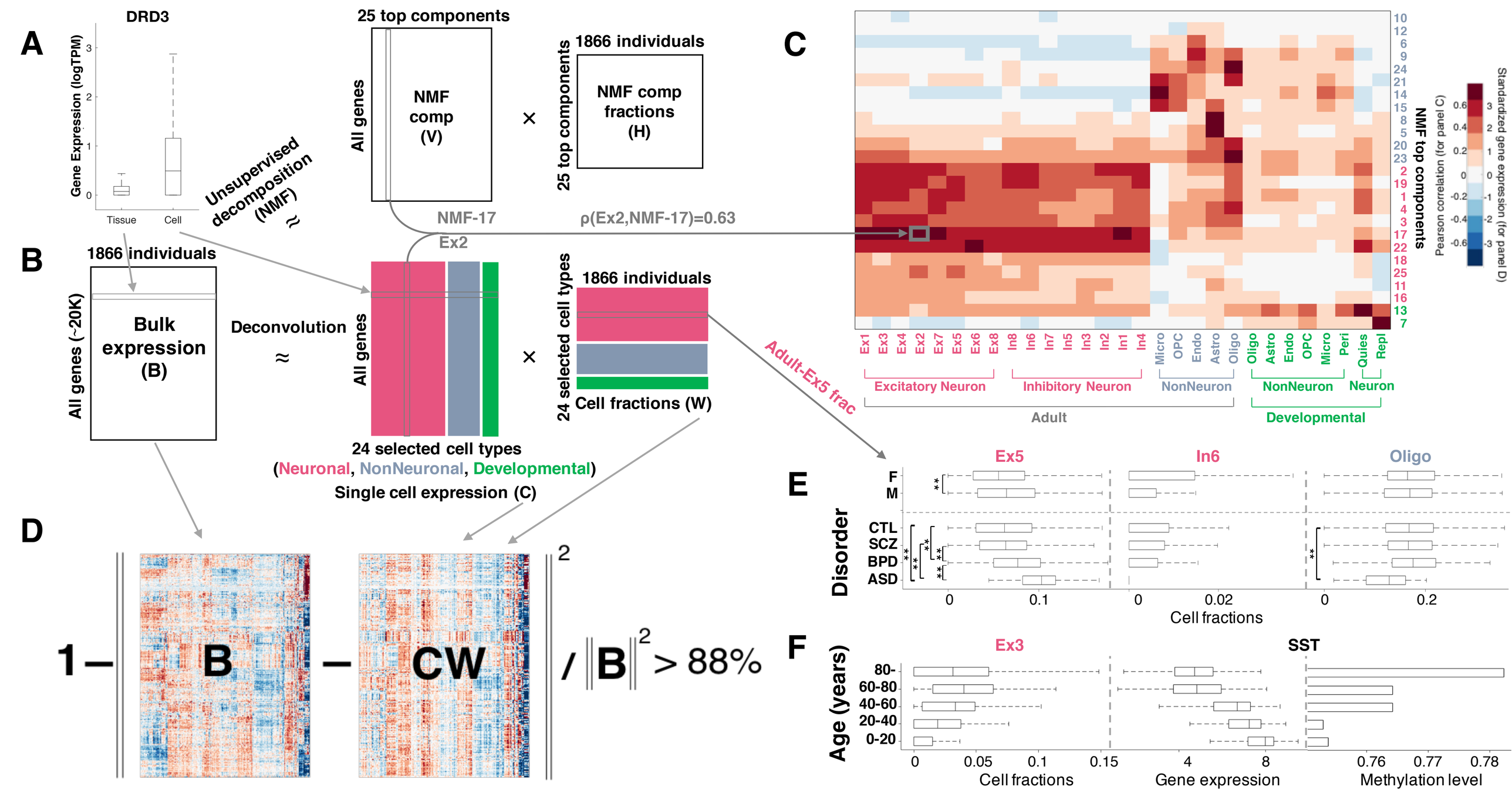


Tissue

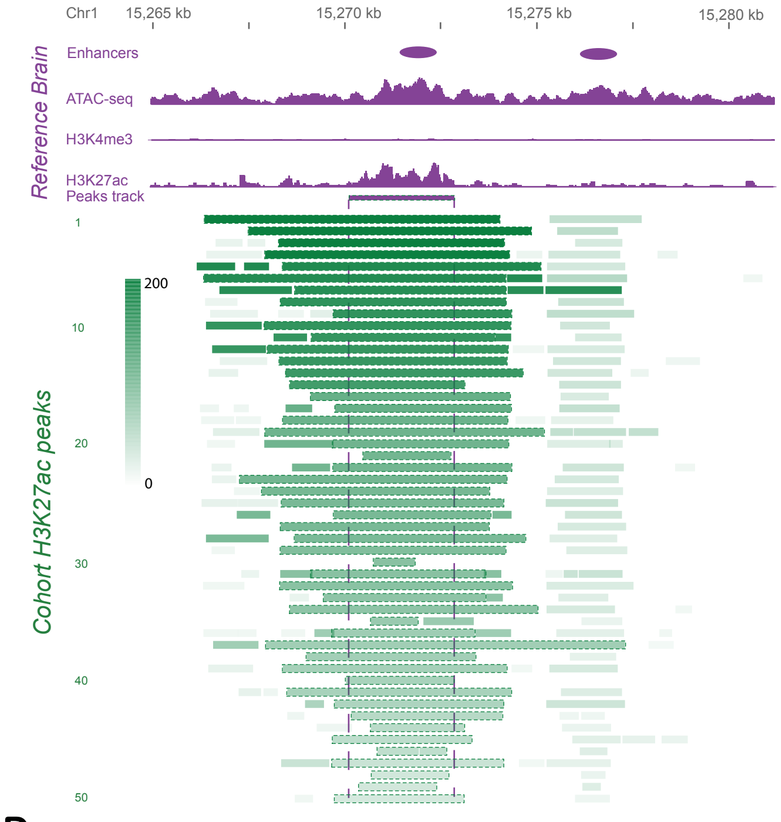




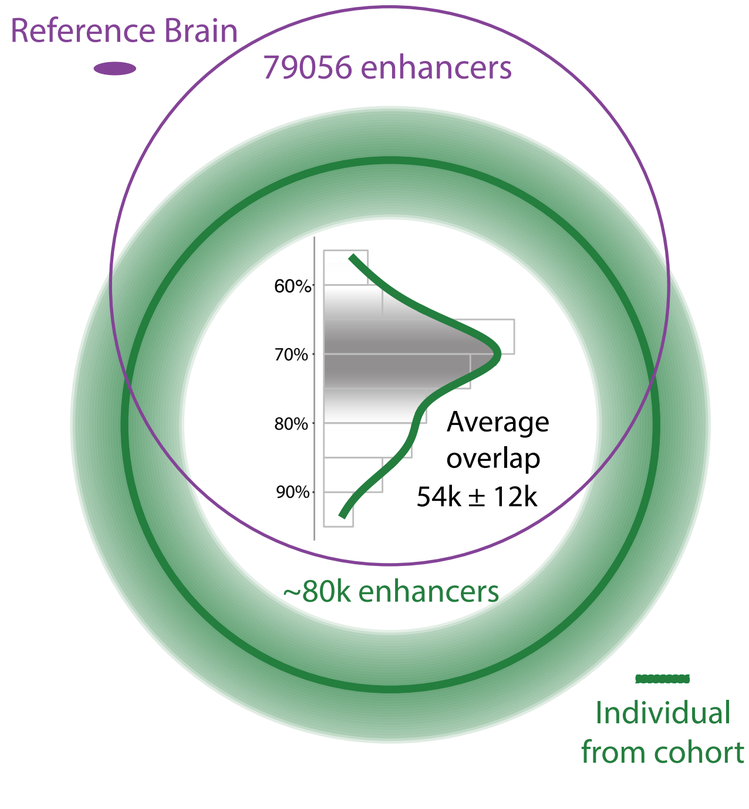



A
DRD3
Gene Expression (logTPM)
Tissue
Cell
Unsupervised decomposition (NMF)
≈
25 top components
All genes
NMF comp (V)
×
1866 individuals
25 top components
NMF comp fractions (H)
NMF-17
Ex2
ρ(Ex2,NMF-17)=0.63
B
1866 individuals
All genes (~20K)
Bulk expression (B)
Deconvolution
≈
All genes
24 selected cell types
(Neuronal, NonNeuronal, Developmental)
×
1866 individuals
24 selected cell types
Cell fractions (W)
Single cell expression (C)
Adult-Ex5 frac
C
NMF top components
Pearson correlation (for panel C)
Standardized gene expression (for panel D)
Ex1 Ex3 Ex4 Ex2 Ex7 Ex5 Ex6 Ex8 In8 In6 In7 In5 In3 In2 In1 In4 Micro OPC Endo Astro Oligo Oligo Astro Endo OPC Micro Peri Quies Repl
Excitatory Neuron
Inhibitory Neuron
NonNeuron
NonNeuron
Neuron
Adult
Developmental
D
1 − ‖B − CW‖² / ‖B‖² > 88%
E
Ex5
In6
Oligo
Disorder
F
M
CTL
SCZ
BPD
ASD
Cell fractions
F
Ex3
SST
Age (years)
80-
60-80
40-60
20-40
0-20
Cell fractions
Gene expression
Methylation level

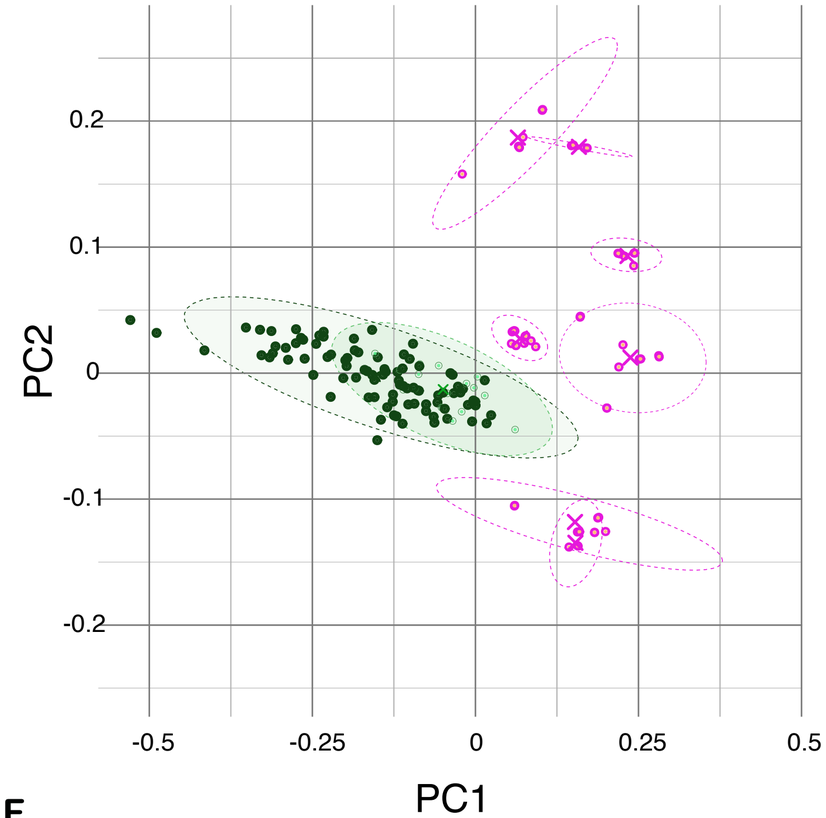
A


Chr1
15,265 kb
15,270 kb
15,275 kb
15,280 kb
Reference Brain
Enhancers
ATAC-seq
H3K4me3
H3K27ac
Peaks track
Cohort H3K27ac peaks
200
0
1
10
20
30
40
50

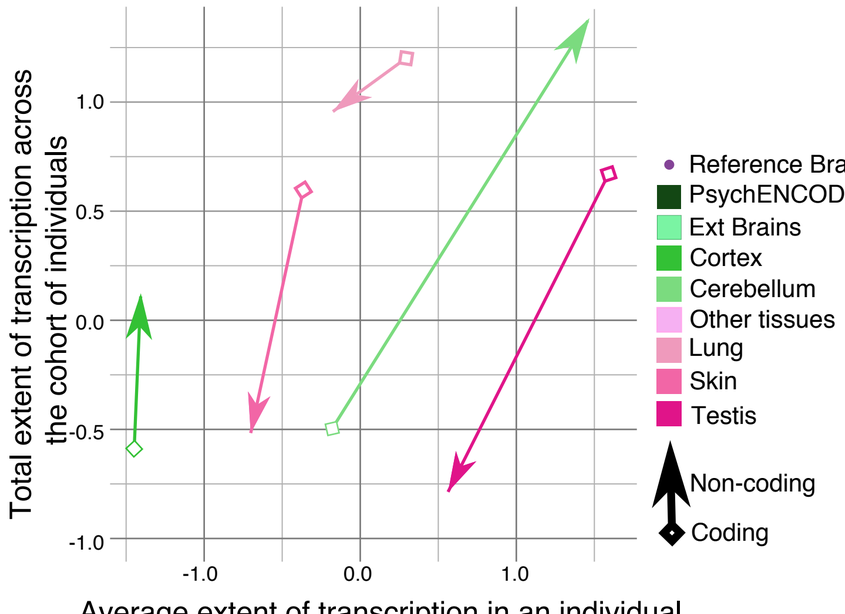
B


Reference Brain
79056 enhancers
60%
70%
80%
90%
Average
overlap
54k ± 12k
~80k enhancers
Individual
from cohort

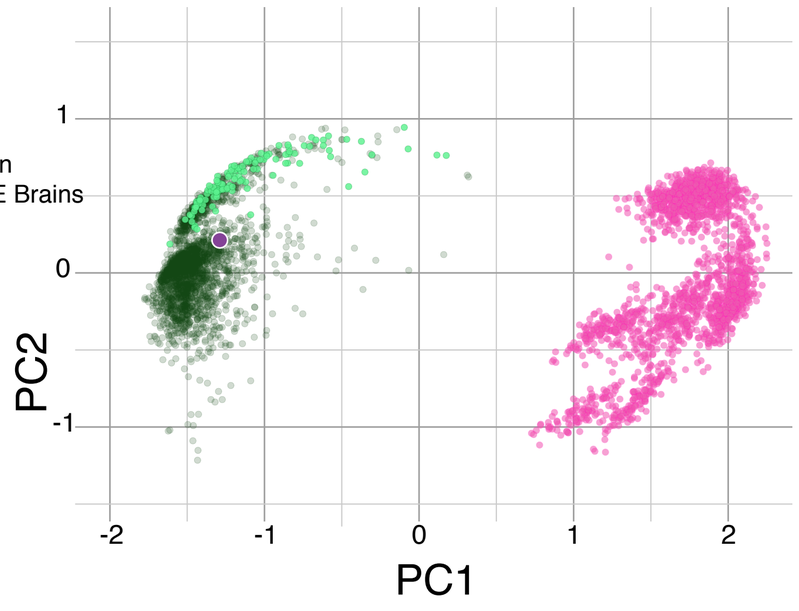
C


PC2
0.2
0.1
0
-0.1
-0.2
-0.5
-0.25
0
0.25
0.5
PC1

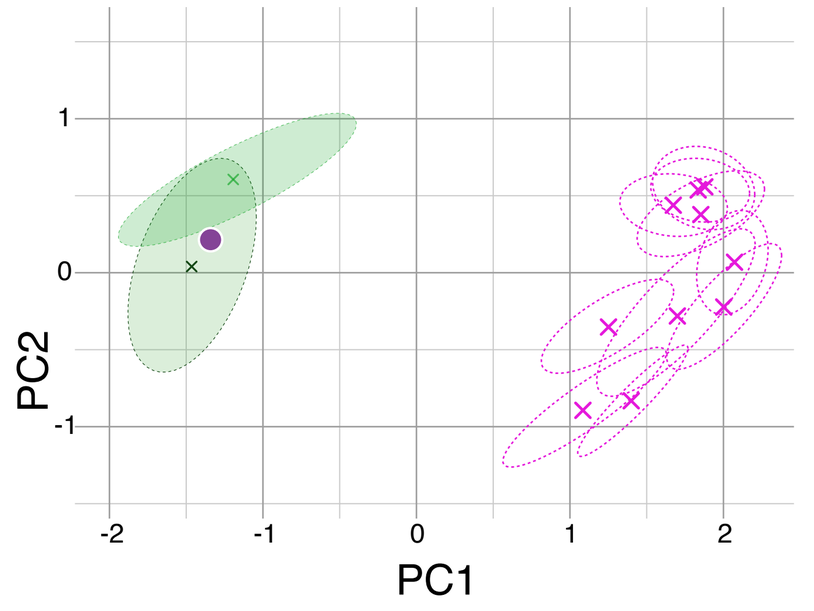
D

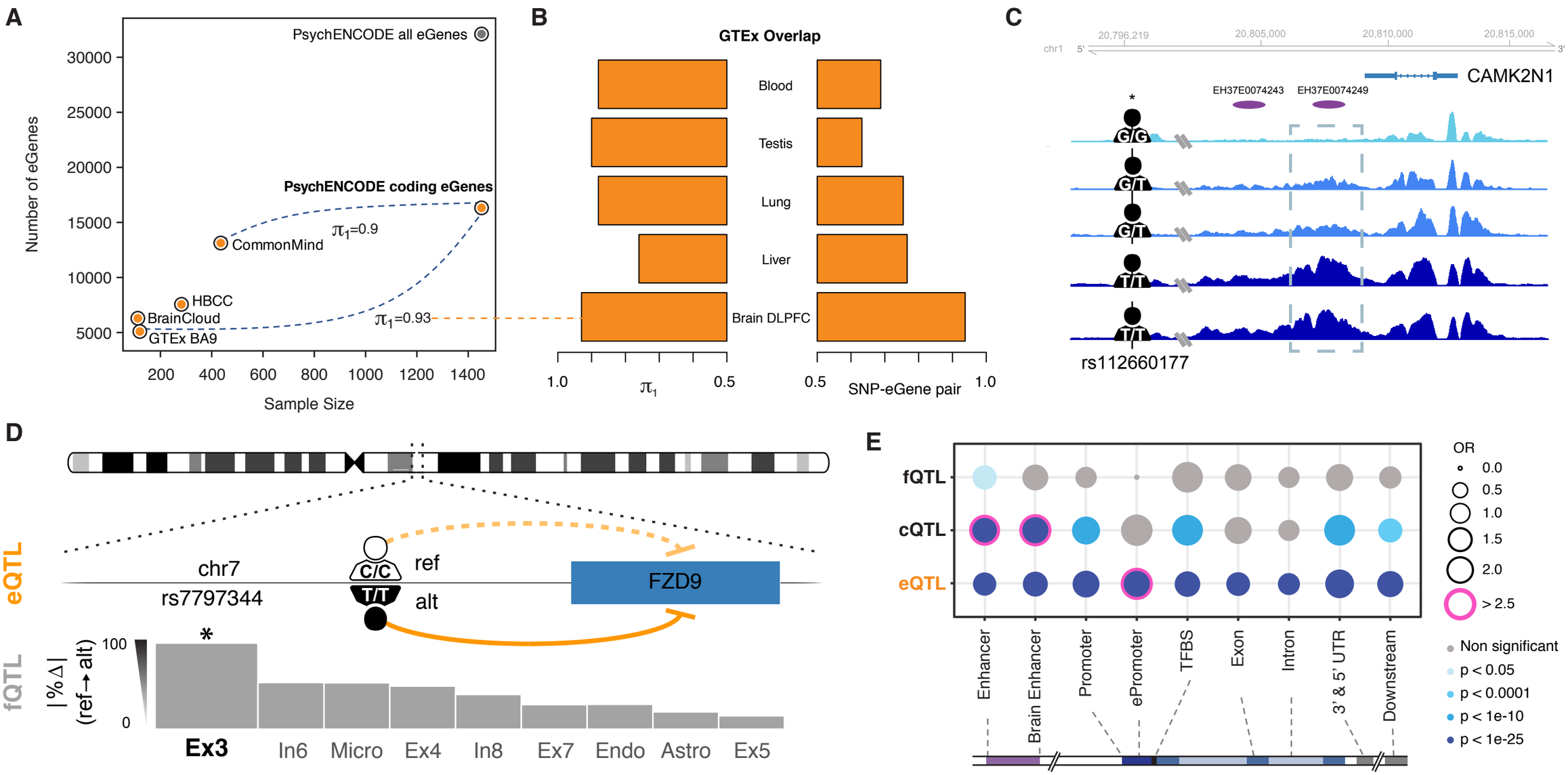

Total extent of transcription across the cohort of individuals
1.0
0.5
0.0
-0.5
-1.0
-1.0
0.0
1.0
Average extent of transcription in an individual
Reference Brain
PsychENCODE Brains
Ext Brains
Cortex
Cerebellum
Other tissues
Lung
Skin
Testis
Non-coding
Coding

E

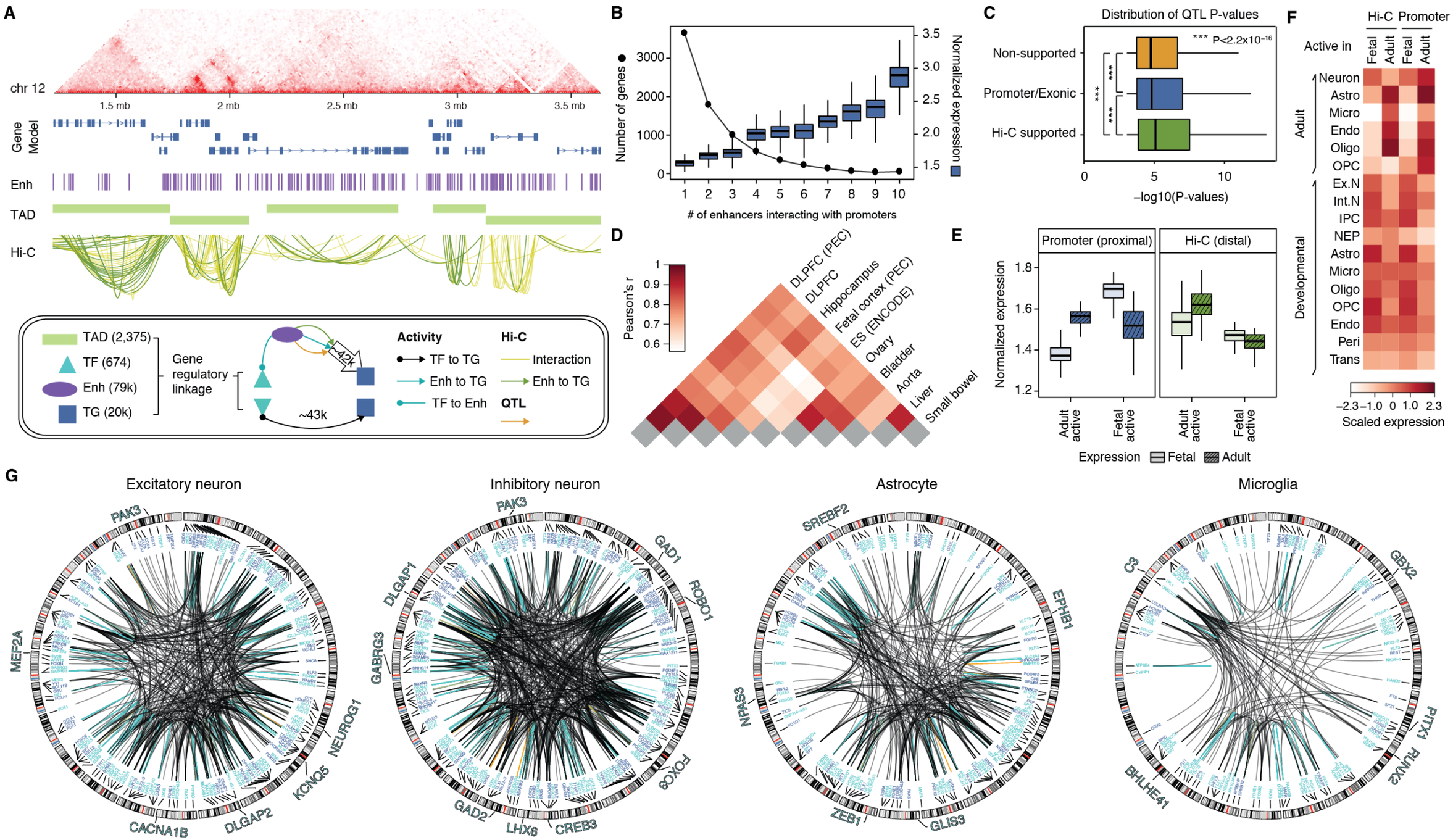

PC2
1
0
-1
-2
-1
0
1
2
PC1

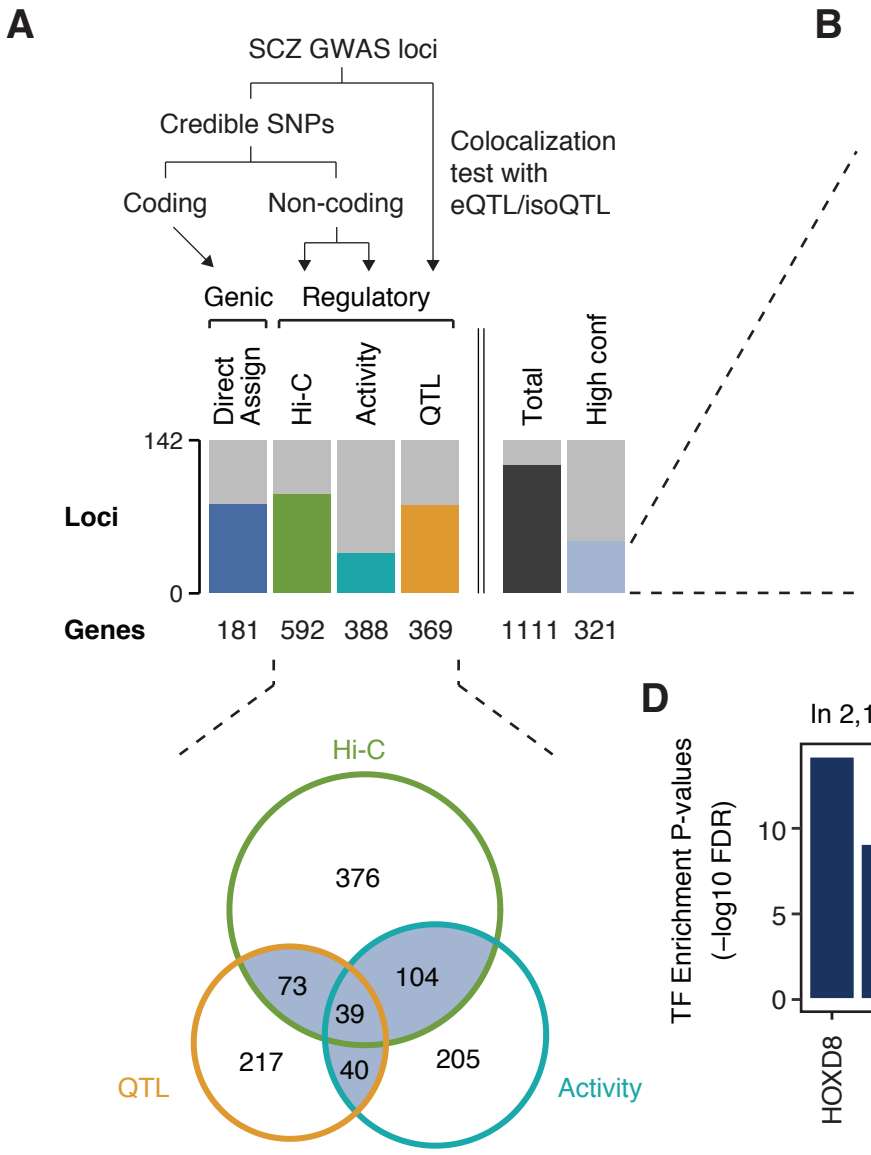
F

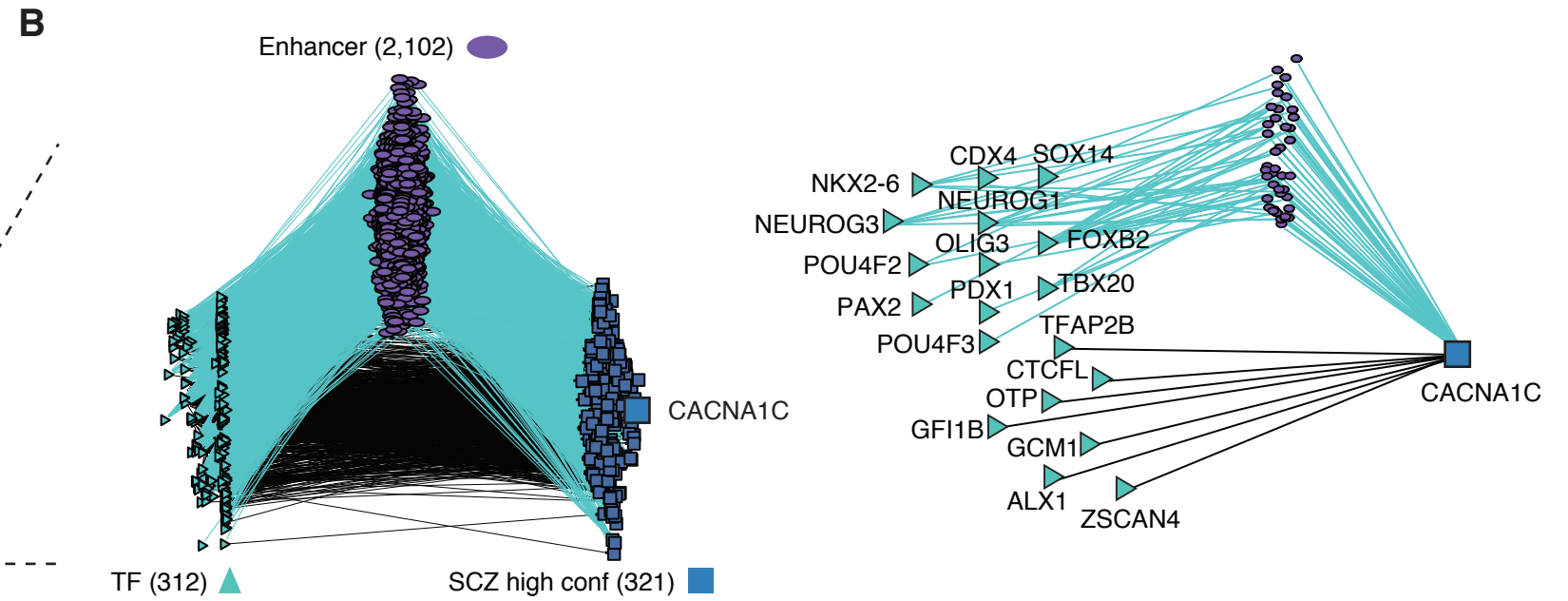

PC1
PC2
1
0
-1
-2
-1
0
1
2
PC1

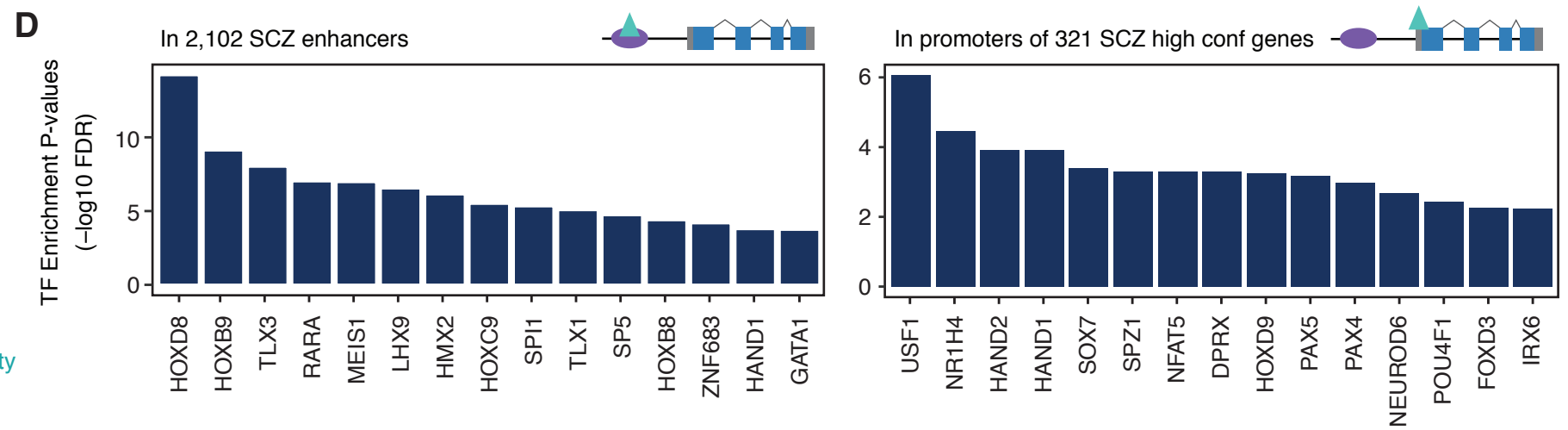


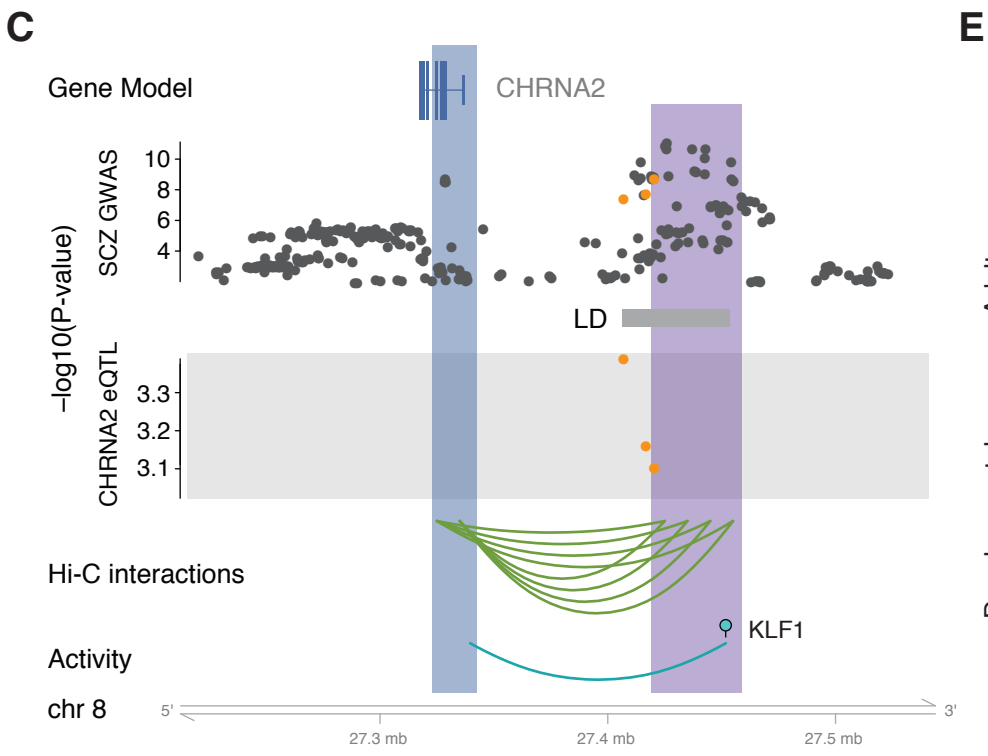
	Number of intermediate phenotype	SNPs
eQTL	32,944 eGenes	1,341,182
cQTL*	8,484 Enhancers	7,983
fQTL	9 Cell types	1,672

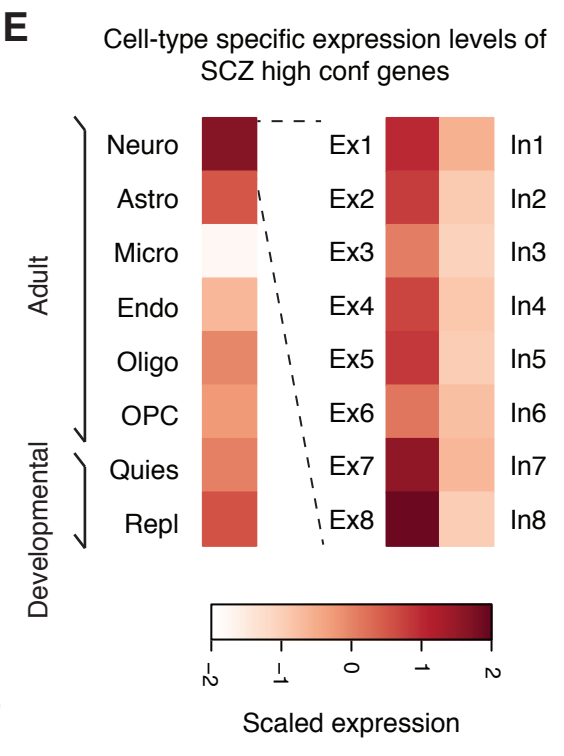

A
chr 12
1.5 mb
2 mb
2.5 mb
3 mb
3.5 mb
Gene Model
Enh
TAD
Hi-C
TAD (2,375)
TF (674)
Enh (79k)
TG (20k)
Gene regulatory linkage
~42k
~43k
Activity
TF to TG
Enh to TG
TF to Enh
Hi-C
Interaction
Enh to TG
QTL
B
Number of genes
Normalized expression
# of enhancers interacting with promoters
C
Distribution of QTL P-values
Non-supported
Promoter/Exonic
Hi-C supported
*** P<2.2x10⁻¹⁶
–log10(P-values)
D
Pearson's r
DLPFC (PEC)
DLPFC
Hippocampus
Fetal cortex (PEC)
ES (ENCODE)
Ovary
Bladder
Aorta
Liver
Small bowel
E
Promoter (proximal)
Hi-C (distal)
Normalized expression
Adult active
Fetal active
F
Hi-C
Promoter
Active in
Fetal
Adult
Adult
Developmental
Neuron
Astro
Micro
Endo
Oligo
OPC
Ex.N
Int.N
IPC
NEP
Astro
Micro
Oligo
OPC
Endo
Peri
Trans
Scaled expression
Expression
Fetal
Adult
G
Excitatory neuron
Inhibitory neuron
Astrocyte
Microglia
PAK3
MEF2A
NEUROG1
KCNQ5
DLGAP2
CACNA1B
DLGAP1
GABRG3
GAD1
ROBO1
FOXO6
GAD2
LHX6
CREB3
SREBF2
EPHB1
NPAS3
ZEB1
GLIS3
C3
GBX2
PITX1
RUNX2
BHLHE41

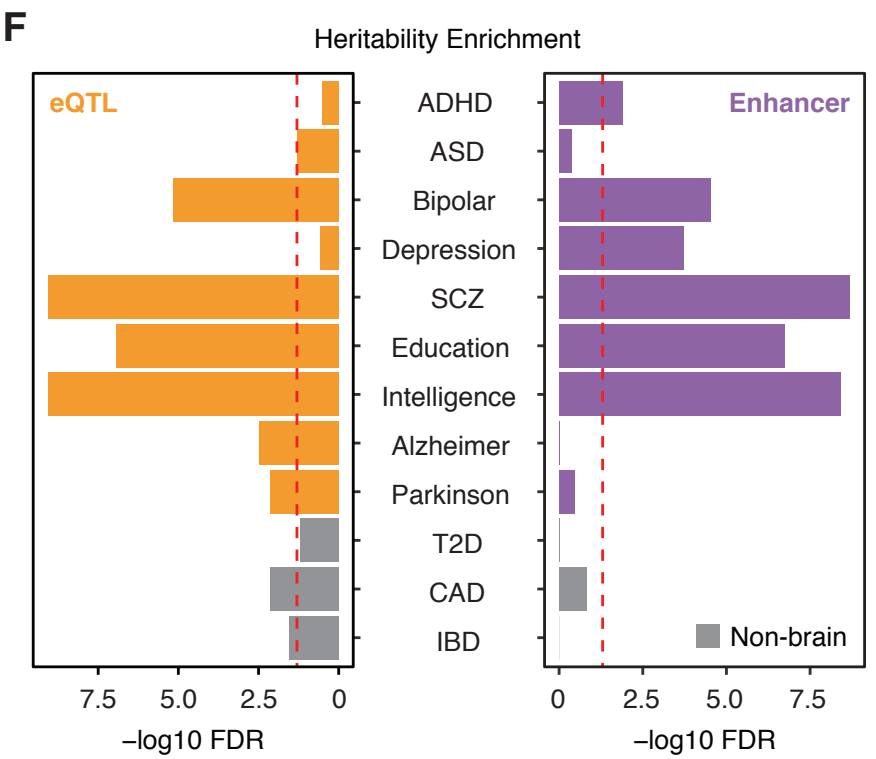

A
SCZ GWAS loci
Credible SNPs
Coding
Non-coding
Colocalization test with eQTL/isoQTL
Genic
Regulatory
Direct Assign
Hi-C
Activity
QTL
Total
High conf
142
Loci
0
Genes
181
592
388
369
1111
321
Hi-C
376
73
104
39
217
40
205
QTL
Activity


B
Enhancer (2,102)
CACNA1C
TF (312)
SCZ high conf (321)
CDX4
SOX14
NKX2-6
NEUROG1
NEUROG3
OLIG3
FOXB2
POU4F2
PDX1
TBX20
PAX2
TFAP2B
POU4F3
CTCFL
OTP
GFI1B
GCM1
ALX1
ZSCAN4
CACNA1C

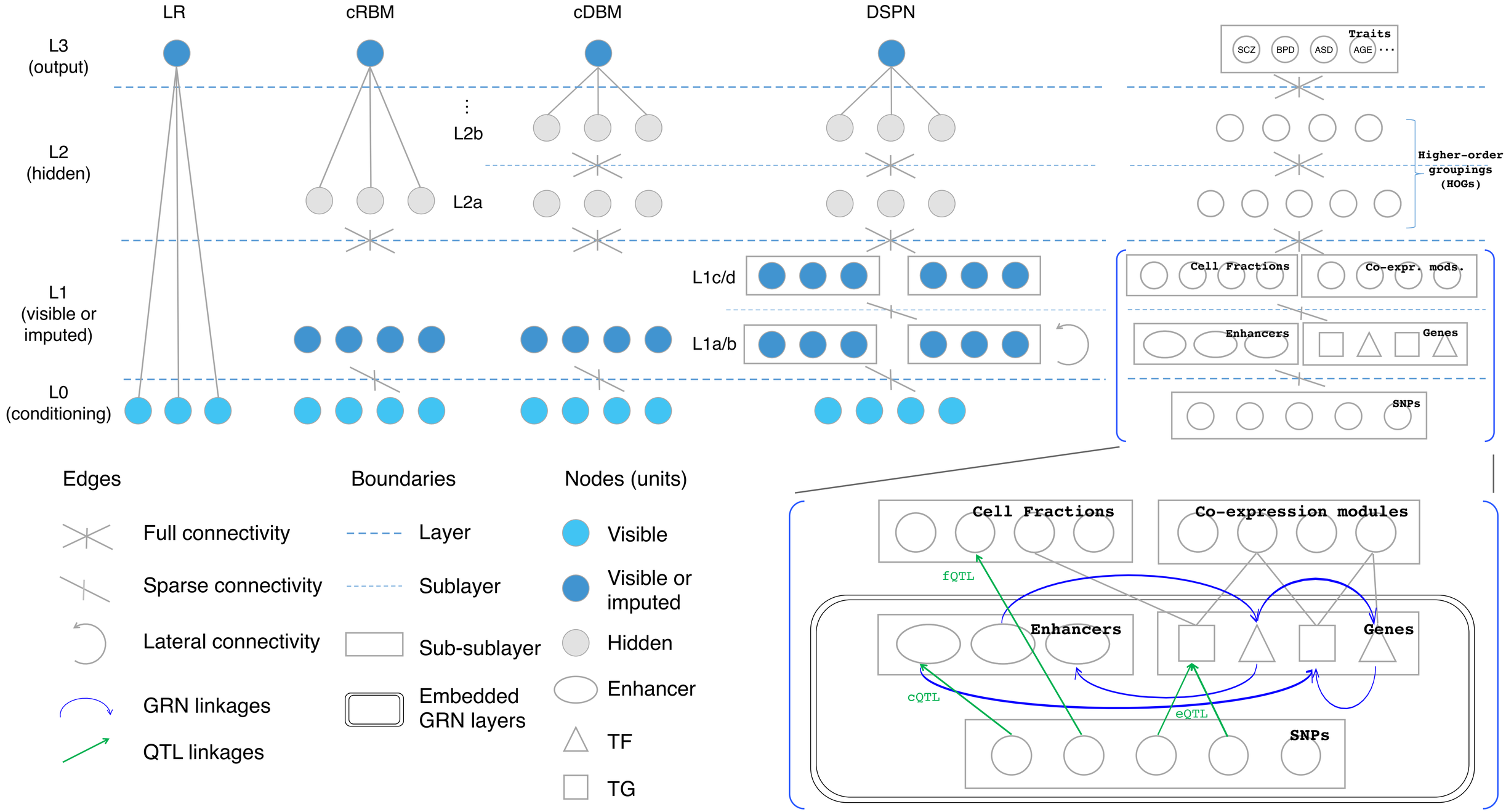

D
In 2,102 SCZ enhancers
TF Enrichment P-values (–log10 FDR)
10
5
0
HOXD8
HOXB9
TLX3
RARA
MEIS1
LHX9
HMX2
HOXC9
SPI1
TLX1
SP5
HOXB8
ZNF683
HAND1
GATA1
In promoters of 321 SCZ high conf genes
6
4
2
0
USF1
NR1H4
HAND2
HAND1
SOX7
SPZ1
NFAT5
DPRX
HOXD9
PAX5
PAX4
NEUROD6
POU4F1
FOXD3
IRX6


C
Gene Model
CHRNA2
SCZ GWAS
–log10(P-value)
10
8
6
4
LD
CHRNA2 eQTL
3.3
3.2
3.1
Hi-C interactions
KLF1
Activity
chr 8
5'
3'
27.3 mb
27.4 mb
27.5 mb


E
Cell-type specific expression levels of SCZ high conf genes
Adult
Developmental
Neuro
Astro
Micro
Endo
Oligo
OPC
Quies
Repl
Ex1
Ex2
Ex3
Ex4
Ex5
Ex6
Ex7
Ex8
In1
In2
In3
In4
In5
In6
In7
In8
–2
–1
0
1
2
Scaled expression


F
Heritability Enrichment
eQTL
ADHD
ASD
Bipolar
Depression
SCZ
Education
Intelligence
Alzheimer
Parkinson
T2D
CAD
IBD
7.5
5.0
2.5
0
–log10 FDR
Enhancer
Non-brain
0
2.5
5.0
7.5
–log10 FDR

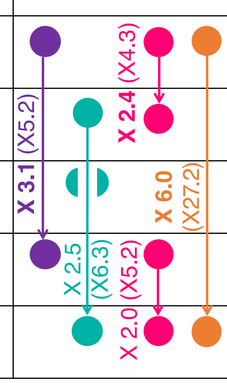
A



B

C

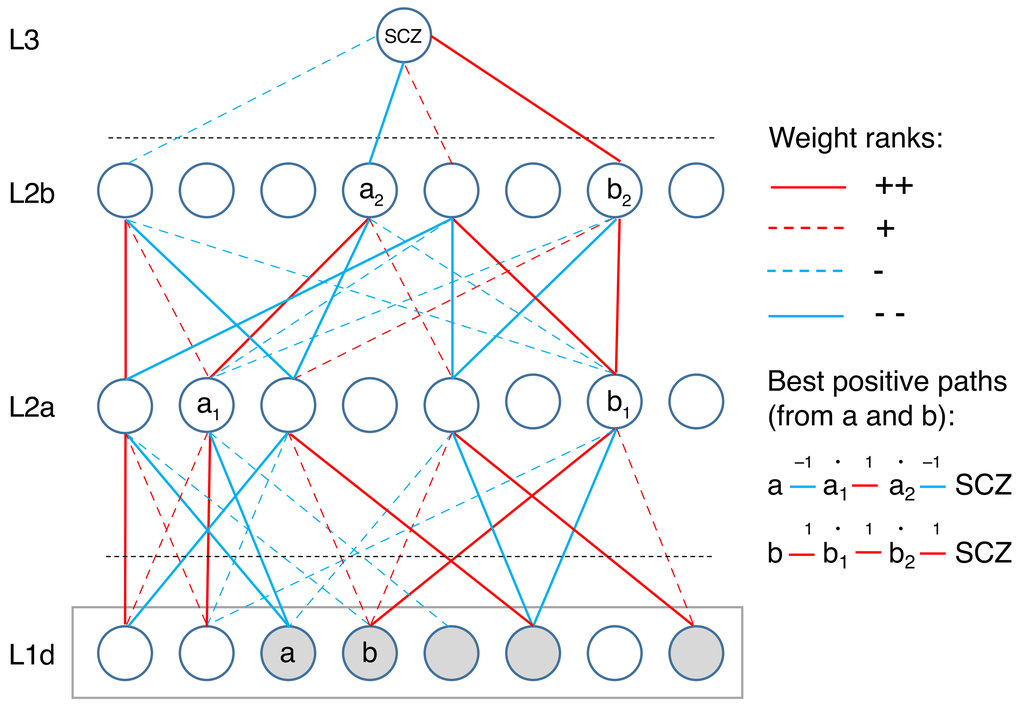
Method	SCZ	BPD	ASD	AVG (SCZ+BPD+ASD)	GEN	ETH	AGE
LR-gene	54.6% (0.5%)	56.7% (2.5%)	50.0% (0.0%)	53.8% (1.0%)	50.0%	99.0%	61.9%(AOD)
LR-trans	63.0% (4.8%)	63.3% (6.3%)	51.7% (1.8%)	59.3% (4.3%)	69.7%	86.0%	81.2%
cRBM	70.0% (31.0%)	71.1% (22.6%)	63.3% (10.8%)	68.1% (21.5%)	71.5%	89.0%	83.1%
DSPN-impute	59.0% (1.8%)	67.2% (10.7%)	58.8% (3.2%)	61.7% (5.2%)			
DSPN-full	73.6% (32.8%)	76.7% (37.4%)	68.3% (11.3%)	72.9% (27.2%)	71.5%	94.3%	86.9%



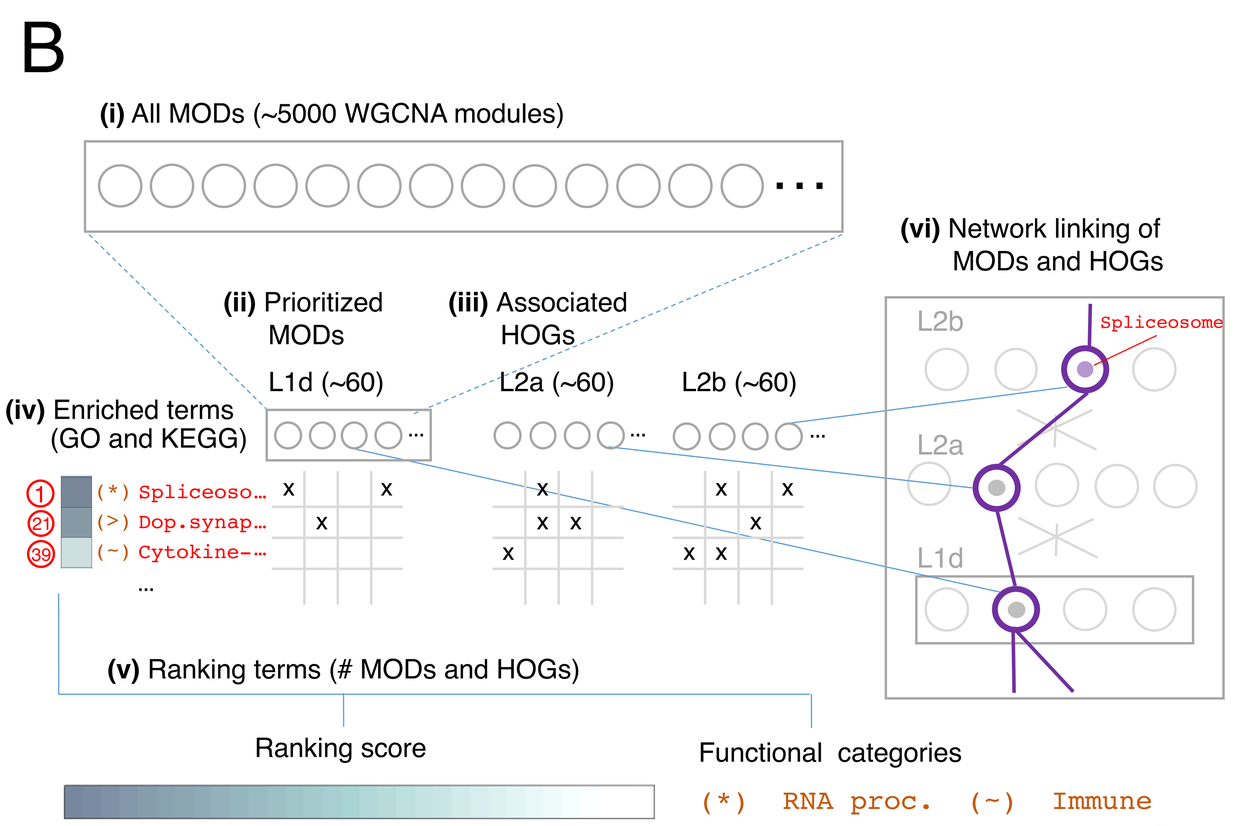
Model complexity	increasing	increasing	constant	increasing
Predictors	genotype	transcriptome	genotype->transcriptome	genotype->transcriptome

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

A



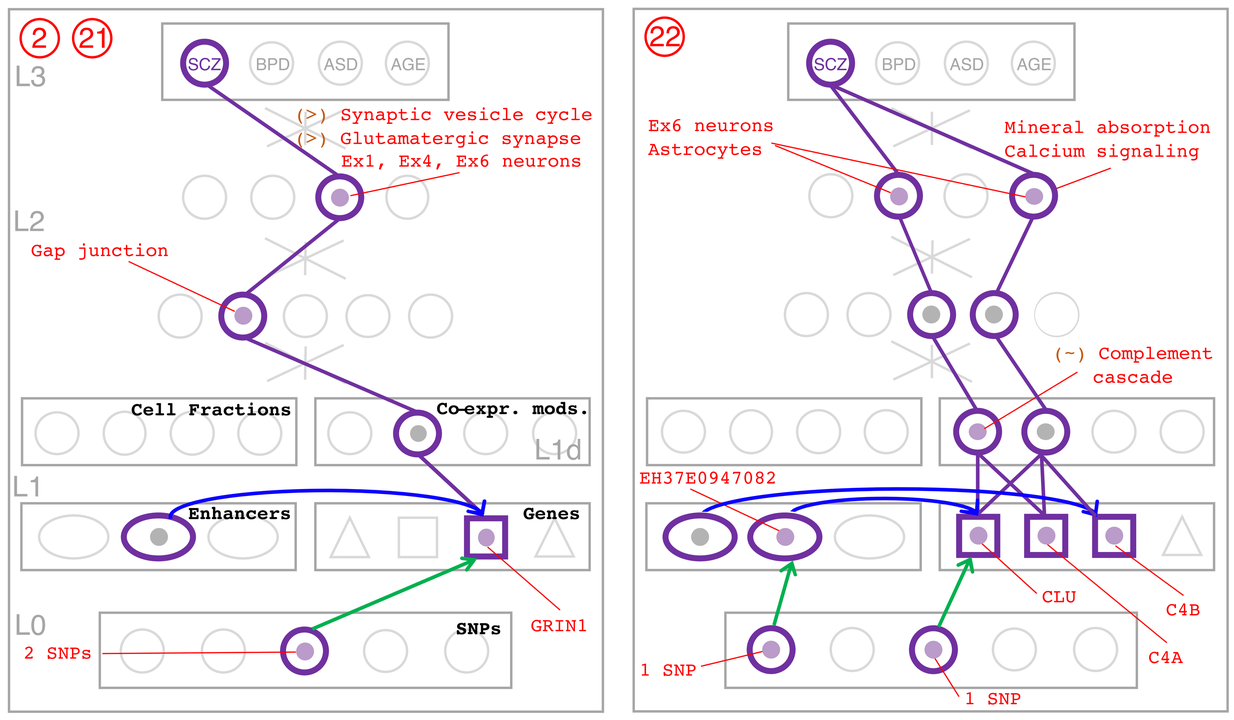
B



C

		SCZ	BPD	ASD		
1					(*)	Spliceosome / RNA splicing
2					(>)	Synaptic vesicle cycle
3					(~)	Antigen proc. and presentation
4					(>)	Vesicle localization
5					(#)	Proteasome
6					(*)	mRNA processing
7						Chromatin modification
8						Oxidative phosphorylation
9						Retrograde endocannabinoid sig.
10					(>)	Chemical synaptic transmission
11						Peptidyl-lysine modification
12						Endocytosis
13					(#)	Ubiquitin mediated proteolysis
14					(>)	Anterograde trans-synaptic sig.
15					(*)	mRNA transport
16						Phosphatidylinositol signaling
17						Hippo signaling pathway
18					(~)	Staph./ Epstein-Barr virus inf.
19					(>)	Synaptic signaling
20					(#)	Autophagy
21					(>)	Dop./GABA/Glutamatergic synapse
22					(>)	Calcium signaling
23						Endocrine calcium reabsorption
24					(*)	RNA degradation / transport
25						Ribosome
(31)						Neuron projection morphogenesis
(33)					(~)	Fc receptor signaling pathway
34						cGMP-PKG signaling pathway
35						mTOR signaling pathway
(39)					(~)	Cytokine-cytokine receptor int.

D



Supplementary Materials for

Comprehensive functional genomic resource and integrative model for the human brain

Daifeng Wang*, Shuang Liu*, Jonathan Warrell*, Hyejung Won*, Xu Shi*, Fabio C.P. Navarro*, Declan Clarke*, Mengting Gu*, Prashant Emani*, Yucheng T. Yang, Min Xu, Michael J. Gandal, Shaoke Lou, Jing Zhang, Jonathan J. Park, Chengfei Yan, Suhn Kyong Rhie, Kasidet Manakongtreecheep, Holly Zhou, Aparna Nathan, Mette Peters, Eugenio Mattei, Dominic Fitzgerald, Tonya Brunetti, Jill Moore, Yan Jiang, Kiran Girdhar, Gabriel E. Hoffman, Selim Kalayci, Zeynep H. Gümüş, Gregory E. Crawford, PsychENCODE Consortium‡, Panos Roussos, Schahram Akbarian, Andrew E. Jaffe, Kevin P. White, Zhiping Weng, Nenad Sestan, Daniel H. Geschwind†, James A. Knowles†, Mark B. Gerstein†

*These authors contributed equally to this work.

†Corresponding author. Email: dhg@mednet.ucla.edu (D.H.G.); James.knowles@downstate.edu (J.A.K.); pi@gersteinlab.org (M.G.)

This PDF file includes:

Supplement to
Comprehensive functional genomic resource and integrative model for the human 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*:

All Supplementary figures, tables, and references are made available at the end of this document. The Supplementary figures and tables are numbered according to the order in which they are mentioned in the main text. Note also that many associated data files are available with unique file IDs on the website http://resource.psychencode.org.

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. These datasets also include processed data such as expression QTL (eQTL) and chromatin QTL (cQTL) 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 resource is organized in 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 (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. Links to the raw data for the RNA-seq, ChIP-seq, and genotype datasets, as well as metadata and phenotypic information are provided under the “Raw Data” section of the website (resource.psychencode.org). Additionally, we identify each of the files provided on the website with a unique file ID. The file IDs follow a convention where the first three letters represent the particular section of the data hierarchy: “RAW” for the raw data, “PIP” for the uniform pipeline processed datasets, “DER” for the derived structures further up the hierarchy, and “INT” for the integrative analysis files at the apex of the pyramid. Next, the file IDs contain two digits corresponding to the sequential order of the file as it occurs on the website. Thus, for example, the enhancers for the prefrontal cortex determined in our analyses are identified as “DER-03” for the hg19-aligned list and “DER-04” for the hg38-aligned list, and are found in the “Derived Data Types” section of the website. Throughout the supplementary materials, we provide these identifiers for the relevant files. Finally, for some data types, we have provided datasets aligned to both hg19 and hg38, with the understanding that the analyses were primarily carried out with alignment to hg19, but UCSC *liftOver* was used to convert the coordinates of some of the files to hg38. These ‘lifted-over’ files include the enhancers (DER-03,

DER-04) and the QTLs (DER-08a-d, DER-09, DER-10a-d, DER-11, DER-12). The coordinates of the target gene and isoforms were also lifted over for completeness where appropriate.

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

The PsychENCODE data covers a number of phenotypes of mental health. These include normal controls (n=1,104), as well as schizophrenia (SCZ) (n=558), bipolar disorder (BPD) (n=217), autism spectrum disorder (ASD), (n=44), and affective disorder (n=8) (Fig. 1). The data include 1,246 males and 685 females. We integrated standard pipelines to uniformly process raw sequencing and genotyping data (Fig. S31). Details are provided in Sections S2.1, S2.2, S3.1, S5.1, S6.1, and 6.2.

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

S2.1 Data processing

Note that the data protocols for this section are provided in detail in Section S9.

S2.1.1 GTEx brain and other tissues

We used several types of data from the GTEx version 7 dataset (*62*). GTEx version 7 contains RNA-seq and matching genotype data for ten 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 (*62*) 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.

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

The PsychENCODE RNA-seq pipeline (Fig. S2) is mostly based on that of ENCODE, which is compatible with stranded and unstranded mRNA 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. S31) at the University of Chicago on an OpenStack cloud system. GTEx samples were processed at Yale University. Links to the RNA-seq signal track and peak files are provided on the website (resource.psychencode.org) under the section "Pipeline-processing results". The bulk tissue normalized gene expression matrix for the PFC is provided as DER-01 (under the section "Derived Data Types"). Outliers of the RNA-seq data were removed using clustering analysis (Fig. S3).

S2.2 Single-cell RNA-seq analysis

S2.2.1 Datasets of single-cell transcriptomics

We build a comprehensive single-cell resource by integrating PsychENCODE-generated adult and developmental single-cell data with published resources. Due to the different sequencing technologies and quantification methods used, we build two single-cell datasets: a read count-based (i.e., TPM) dataset and a UMI count-based dataset (*63-65*). For the read count-based dataset, 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 eight excitatory and eight inhibitory types (*63*), and ~400 cells including two developmental types, one adult neuronal type, and five adult non-neuronal types (i.e., astrocytes, endothelial cells, microglia, oligodendrocytes, and oligodendrocyte progenitor cells (OPCs) (*65*) were downloaded from corresponding publications. For the UMI count-based dataset, we integrated the PsychENCODE adult single-cell profiles for 17,093 cells from dorsolateral prefrontal cortex (DFC) with the published 10,319 adult single-cell data from PFC (*64*). The integrated UMI dataset includes nine excitatory, ten inhibitory and six non-neuronal cell types (astrocytes, endothelial cells, microglia, oligodendrocytes, OPCs, and pericytes), and a newly discovered excitatory neuronal type (Ex9). The details of cell types are shown in Tables S3 and S4. We also compared the gene expression profiles between the read count dataset and the UMI count dataset based on biomarker genes. As shown in Figs. S6 and S7, the expression profiles are very similar between the two single-cell datasets.

The basic cell types have been shared and used by other PsychENCODE Capstone projects focusing on non-coding regulation and development. For PsychENCODE developmental single-cell data, we first applied quality control on ~900 cells using the R 'scater' package (*66*) 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 single cell datasets from PsychENCODE and (*63-65*) by matching the gene names. As the single-cell data suffers from high dropout rates, we used MAGIC (*67*) to impute the missing values in the expression matrix. We compared these single cells based on (biomarker) gene expression similarity using tSNE (*68*), and found that cells of the same type generally could be clustered together. In particular, 99.4% PsychENCODE developmental cells clustered together with known developmental cell types from a previous report (*65*).

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. As the sequencing and quantification methods are different between the read count dataset and the UMI count dataset, we only used the read count dataset for deconvolution analysis to ensure the expression profiles were consistent with bulk data.

S2.2.2 Quantification of gene expression

The gene expression in both bulk and single-cell read count 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). The UMI count data were also normalized to make the total UMI counts per cell as one million and further transformed into log scale. The bulk tissue gene expression matrix for the PFC in units of TPM is provided as DER-02 (under the section "Derived Data Types").

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 (Fig. S52). 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 $||X-V*H||^2$ was minimized and all elements of H were 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. For the decomposition, we used the R package 'NMF' with the 'Brunet' algorithm (*69*) to decompose the gene expression of bulk tissues. In terms of initialization, we computed the decomposition with a random seed 100 times. The decomposition with the best fitting was used for further analysis. To evaluate the robustness of the algorithm, we also calculated the empirical p-value of the residuals for the decomposition, which showed the NMF decomposition is very significant (p-value = 4e-3).

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. 5 and 8 NMF-TCs of the non-neuronal group highly correlated with astrocytes, No. 7 and 13 NMF-TCs correlated with developmental cells, and No. 3, 4, 17, and 22 NMF-TCs of the neuronal group correlated with excitatory neuronal cell types. A similar correlation pattern was

observed by comparing the NMF-TCs with the UMI count dataset profiles (Fig. S9). 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 (*70*). We found here that some of our NMF-TCs correlated with the eigengenes of gene co-expression modules (*71*), especially for the cell type modules, supporting again that they connect the cell type information from the bulk tissue data.

To further evaluate the relationship between the NMF-TCs with the single-cell profiles, we also did the NMF decomposition using the marker genes from UMI dataset. The correlation heatmap between the NMF-TCs and the single-cell profiles is shown in Fig. S5, which shows a similar pattern between the two datasets. In addition, we can still see the NMF-TCs that are correlated with neuronal and non-neuronal cell types.

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. As the bulk data are measured in terms of read counts, we select to use the read count single-cell dataset for this analysis to ensure that the quantification of gene expression is consistent.

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 deconvolution model by calculating the reconstruction accuracy as $1-||B-C*W||^2/||B||^2$, which is related to a cosine similarity measurement that is widely used to quantify the similarity between vectors or matrices. Fig. S11 shows two examples of calculating the reconstruction accuracy. For the convenience of illustration, we just show the example deconvolution of one sample based on two genes. If there is only one cell type, the reconstruction of bulk gene expression will find the coefficient w that minimizes the error vector $\bar{b} - w * \bar{c}$. The optimal solution of the reconstructed vector $w * \bar{c}$ will be orthogonal to the error vector. Then, the reconstruction accuracy will be defined as $1-||\bar{b} - w * \bar{c}||^2/||\bar{b}||^2=\cos^2\theta$ where θ is the angle between the bulk tissue expression and the reconstructed gene expression. The definition of reconstruction accuracy also works for the other example with more than one (two) cell types, as the combination of weighted cell type profiles ($w_1 * \bar{c}_1 + w_2 * \bar{c}_2$) will be another vector similar to the single cell type case. We have calculated the reconstruction accuracy for each individual using our deconvolution analysis, which is shown in Fig. S12. As shown, individual reconstruction accuracies remain at a high level; for example, the median and standard deviation are 0.8875 and 0.0399, respectively. In addition, we deconvolved the tissue expression

data and compared the cell fraction changes for various phenotypes and psychiatric disorders (Figs. S14, S15, and S19 show the cell fractions across different ages. We found that Ex3 and Ex4 had significant increasing trends across age (trend analysis p<6.3e-10 and 1.5e-6, respectively), but some non-neuronal types such as oligodendrocytes were found to decrease (p<2.1e-14). In addition, the astrocytes and microglia cells showed higher cell fractions in ASD samples compared to control (CTR) samples (Fig. S18), which is concordant with the literature (*72*). Furthermore, these age-related cell changes were potentially associated with differentially expressed genes across age groups; for example, a gene involved in the 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 BPD and increased astrocyte fractions for SCZ. The details of the statistical test used for the cell fractions can be found in Table S12.

We have validated our estimated cell fractions on a subset of samples from the EPIMAP study with experimentally measured NeuN+ fractions. Fig. S10 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 a popular deconvolution tool CIBERSORT (*73*). We performed CIBERSORT to deconvolve the tissue expression data with single-cell data of 24 selected cell types and further calculated the reconstruction accuracy; this value (0.8132) was lower than that calculated by our deconvolution method (0.8779). We have also computed root mean squared error (RMSE) for each individual for both our method and CIBERSORT. Fig. S13 shows their RMSE distributions. Our deconvolution has a median RMSE of 1.5029 and a standard deviation of 0.1690. CIBERSORT has a median RMSE of 1.8349 and a standard deviation of 0.2726. Thus, our deconvolution fits the bulk tissue expression better than CIBERSORT with lower RMSEs. To assess the cell fractions difference between different brain regions, we further deconvolved 69 samples with both cerebellum (CB) and PFC data using the CB single-cell data from (*64*). The cell types of the CB single-cell data are shown in Table S2. The cell fractions of CB and PFC from the bulk tissues of the same individuals are shown in Figs. S16 and S17.

Data files associated with both the decomposition (NMF components and fractions) and deconvolution (cell fractions) analyses are available on the website (resource.psychencode.org). The marker genes merged from (*65*) and (*63*) are provided as DER-19, and the marker genes from PsychENCODE (adult) and (*64*) are provided as DER-21. The processed single-cell expression data in TPM merged from the PsychENCODE (developmental), (*65*) and (*63*) datasets are provided as DER-20, and the raw single-cell expression data in UMI merged from the PsychENCODE (adult) and (*64*) are provided as DER-22. The raw and normalized cell fractions from the deconvolution are provided as DER-23 and DER-24, respectively. The NMF components and fractions are provided as DER-25 and DER-26, respectively.

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 SCZ 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 (*71*). 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 (resource.psychencode.org).

S2.6 Gene co-expression network analysis

We used WGCNA to identify modules of co-expressed genes, both within and between tissues (*74*). 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 (*74*) in setting β such that connectivity of the network was as close to scale-free as possible (using the R^2 statistic described in (*74*)). 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 (*75*). 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 (*74*)), 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 different expression levels between brain and non-brain samples (Fig. S48), which suggests that brain-specific regulatory mechanisms drive these brain co-expression modules (*71*). For input to the DSPN-mod model (Sec. S8.2.4), we reran WGCNA multiple times (independently on each region from GTEx, using all GTEx data, independently on SCZ, BPD, ASD and control (CTR) subsets of the PEC data, using SCZ+CTR, BPD+CTR and ASD+CTR PEC subsets, and using all PEC data), generating 5024 modules and submodules in total. Associated data files with the gene and isoform co-expression modules from both the current manuscript and the submitted report are available on the website (resource.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). Fig. S20 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 (*76*) were used. We used the normalized (scaled) proportion of methylated CpGs across individuals' age bins near gene TSS (Fig. S21).

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. S22) for maximal compatibility of our results. 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. Links to the ChIP-seq signal tracks and peak files are provided on the website (resource.psychencode.org) under the section “Pipeline-processing results”.

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 (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 cis-regulatory elements (ccREs) for major cell lines and tissues (Fig. S23). The ccREs 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 ccREs 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 ccREs 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 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 (Roadmap) 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 signals (Z-score > 1.64), which is a characteristic marker for enhancers. To exclude promoters, we excluded proximal regions with enriched H3K4me3 signals. These identified enhancer regions largely overlapped with ChromHMM enhancer annotations of the PFC (>90%). Among these enhancers, we identified a high-confidence subset of 18212 enhancers, which are regions with strong ATAC-seq and DNase signals, as well as strong H3K27ac signals from both the reference PFC and Roadmap PFC ChIP-seq experiments.

We uniformly processed 150 H3K27ac ChIP-seq data from healthy individuals, 50 each from PFC, PC, and CB. We called H3K27ac peaks for each sample using the standard ENCODE ChIP-seq pipeline. In general, the numbers of peaks from cerebellums are smaller, despite that cerebellum samples are of slightly higher average sequencing depth (Fig. S24). The H3K27ac peaks were pooled across the cohort, with each pooled peak present in more than half of the samples in its corresponding brain region, generating a total of 37,761 H3K27ac pooled peaks in PFC, 42,683 in TC, and 26,631 in CB. The difference in the number of enhancers could be due to the cellular heterogeneity in the brain regions: the cortex contains diverse cell types such as neurons and glia, whereas the cerebellum is comprised primarily of neurons, specifically Purkinje and granular cells (*77*). This hypothesis is supported by the cell fraction analysis in Fig. S16, where the cerebellum is shown to contain a high fraction of only a few cell types. Also supporting this hypothesis, previous literature has shown that RNA-seq from cerebellum exhibits a distinct expression pattern compared with other brain regions (*78*). Finally, 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 was 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. S25. 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 PFC. The PsychENCODE enhancer lists are available on the website (resource.psychencode.org). The enhancers aligned to hg19 and hg38 (converted from hg19 using UCSC *liftOver*) are provided as DER-03a and DER-04a, respectively. We further provide the high-confidence subset as DER-03b and DER-04b for the hg19 and hg38 alignments, respectively. The merged H3K27ac from UCLA-ASD studies on the PFC, temporal cortex, and cerebellar cortex are provided as DER-05, DER-06, and DER-07, respectively.

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 ENCODE, PEC, GTEx (*62*), and Roadmap (*79*). 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 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 (*80*). 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 (*68*). 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 (*81*). Moreover, as shown in Figs. 3E and S26, all the brain tissue samples from the different projects tended to group together, which is a consequence of our uniform processing. Here, we also color samples by tissue of origin to highlight other tissue groupings (Fig. S26).

In order to perform 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. S27, Fig. S28, and Table S5).

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. S28; the path is represented by the dark like moving from the brain to other tissues). In total 1,226 genes 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.

Developmental chromatin changes also potentially impact gene expression. Thus, we compared the transcriptome data between fetal and adult samples to see if the gene expression changes across developmental stages. In order to contextualize the development of the human brain with adult brain and other tissues, we projected fetal brain RNA-seq data into the RCA space. We used all PsychENCODE adult brain and GTEx tissues as references. For this analysis we selected brain samples extracted with 8 (1),9 (1),12 (3),13 (3),16 (3),17 (1),19 (1),21 (2),22 (1),35 (1),37 (1) weeks post conception (PCW). RNA-seq quantification was done as described earlier. Next, we selected the genes used in the RCA reference panel to project fetal samples into the RCA space in Figure 3E. The result of this analysis is displayed in Fig. S41. Fetal samples were broadly distributed in the RCA space. However, when samples were colored by PCW, we noticed a trajectory progression. This trajectory suggests that fetal brain becomes more similar to the adult brain as the brain develops but has potential specific regulatory mechanisms involving chromatin changes at early stages. In addition, we noticed that the transcriptome differences between fetal and adult brain were consistently smaller than the differences between the human brain and other tissues (Fig. S41).

Similar to the transcriptome RCA analysis, we built a reference panel using H3K27ac signals overlapping ccREs as previously defined. For reference tissues, we used uniformly processed ENCODE tissue samples and calculated the average H3K27ac signals across ccREs for each reference tissue. We further filtered outlier ccREs to select informative ccREs. Similar to the transcriptome analysis, we selected ccREs with average signal across the ccREs higher than 0.1 from 40 tissues. That filter yielded 5,506 reference ccREs. 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. In order to consistently compare the transcriptome and epigenome, we selected tissues on ENCODE that were also represented in the transcriptome RCA analysis. We also performed a PCA analysis for these samples (Fig. S29).

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 non-annotated) 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. S30).

We fit the curves on Fig. S30 to cumulative exponential curves to estimate a per tissue upper bound of the proportion of coding and non-coding transcribed windows. We observed that most tissues were transcriptomically saturated at approximately 100 individuals. Moreover, although a large percent (65-75%) of the coding transcriptome was active, only 3-10% of the non-coding transcriptome was found to be active. Hence, we scaled the estimated cumulative sample transcriptional activity (i.e., estimated the maximum number of transcribed windows in each tissue). 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 number of 100bp windows (with average RPM higher than 0.1) across samples in a tissue-based fashion. Values displayed in Fig. 3 were normalized by average diversity in coding and non-coding regions. The inter-sample variability was estimated by calculating the mean difference. Absolute values for coding and non-coding transcriptome diversity were also estimated.

The genome-wide TARs and those TARs found with at least 70% of the samples are provided on the website (resource.psychencode.org) as PIP-03 and PIP-04, respectively.

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

S5.1 Genotype data processing

The raw genotype data were called and converted to PLINK files, and we ran an initial quality sample level and marker level using PLINK (see Fig. S33 and its associated caption for further details on genotype data processing).

S5.2 eQTL and isoform QTL

We used a conservative approach for eQTL identification, as well as for measuring splicing activity as it relates to QTL 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), as well as with published GTEx results (Fig. S31).

With respect to splicing-related QTLs, we also used the GTEx pipeline (specifically, to calculate isoQTLs and tQTLs). The lists of significant eQTLs and isoQTLs are available at resource.psychencode.org as the files DER-08a-d and DER-10a-d, respectively.

S5.3 cQTLs

To calculate cQTLs, we used the uniformly processed ChIP-seq data from PsychENCODE (three different brain regions) and Roadmap ChIP-seq data for different tissues. cQTLs were calculated using candidate cis-regulatory regions (ccREs). 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. The list of significant cQTLs is provided on the website (resource.psychencode.org) as DER-09.

S5.4 Cell fraction & residual QTL

We used the QTLtools package (*82*) 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. To best deal with population structure as a potential confounding factor, we restricted our analysis to European adult samples, which comprise a substantial subset of all available genotyped data (Fig. S34a).

We computed the fQTLs by treating cell types (i.e., the phenotypes) in almost the same way that distinct genes are used in the context of calculating eQTLs, with the very important distinction being that the cell type phenotype naturally does not have genomic coordinates. Thus, the cell types (treated like gene labels) were used in a trans-like QTL search, in which the window of all eligible SNVs for each cell type was the entire genome. In this way, each cell type fraction was tested against all SNVs in all chromosomes. Gender and disease status were used as covariates in finding fQTLs. We took the conservative approach for eQTLs and isoQTLs to

define significant fQTLs to be those associated with Bonferroni-corrected p-values of no more than 0.05. The search for fQTLs was carried out using the cell fractions for nine cell types (see below) as phenotypes, and the Bonferroni correction applied considers the nine phenotypes that are tested against all ~5.2 million SNVs. We note that the machinery used in computing these fQTLs was the same as that used for the other QTLs investigated (that is, we used QTLtools, in keeping a protocol that is as consistent as possible with that used by GTEx).

We also note that expression patterns for the 458 biomarker genes were used in order to infer the relative cell fractions for 24 distinct cell types. Of these 24 cell types, we identified fQTLs for nine distinct cell types (Ex3, Ex4, Ex5, In6, In8, astrocytes, microglia, and endothelial cells). Again, these nine cell types are those that exhibit fQTLs when using gender and disease status as input covariates. We found that different cell types exhibit considerable heterogeneity in terms of their abundance within the set of high-confidence fQTLs (Fig. S34b). The SNVs associated with these fQTLs coincided with 106 distinct SNVs associated with cis-eQTLs. The list of significant fQTLs is provided on the website (resource.psychencode.org) as DER-11.

Implicit model with fQTLs, for ***one gene*** in ***one individual*** with ***many (L) SNVs***: We can re-express the model in panel B by integrating the matrix of fQTL effects $\mathbf{\Gamma}$, whose (i, j)'th entry represents the effect of SNP i on cell type j:

$$\mathbf{c^T}(\mathbf{g^T\Gamma} + \boldsymbol{\eta}^\mathbf{T})^\mathbf{T} + \mathbf{g^T}\widetilde{\boldsymbol{\beta}} + \mathbf{g^T\beta}(\mathbf{g^T\Gamma} + \boldsymbol{\eta}^\mathbf{T})^\mathbf{T} + \epsilon = b$$

$$\mathbf{g^T}\widetilde{\boldsymbol{\beta}} + (\mathbf{c^T} + \mathbf{g^T\beta})(\mathbf{g^T\Gamma} + \boldsymbol{\eta}^\mathbf{T})^\mathbf{T} + \epsilon = b$$

where $\boldsymbol{\eta}$ is the vector of the random errors (noise terms) for each cell type from the fQTL analysis. As shown by the rearrangement in line 2 above, the second term of the fully expanded model is quadratic in $\mathbf{g}$, and hence contains implicit SNP-SNP interaction terms. For instance, the coefficient of $g_a g_b$, for $a \neq b$, is $\sum_i(\Gamma_{a,i}\beta_{b,i} + \Gamma_{b,i}\beta_{a,i})$, where i ranges across cell types.

We note that when the fQTLs and resQTLs (Fig. S35) are estimated for each SNP independently, the joint models in panels B and D only hold if there are not strong dependencies between the SNPs themselves (i.e., in the absence of LD). Alternatively, we can consider $\mathbf{\Gamma}$ and $\boldsymbol{\beta}$ to be sparse matrices, with only a single non-zero entry per row (i.e., the effect of a chosen SNP when estimated independently). Each possible combination of individual SNP predictors will give rise to a subset of SNP-SNP interactions of the same form as above.

S5.5 QTL replication and sharing

We evaluated the replication of GTEx and CommonMind PFC eQTLs in our study using the π_1 statistic (*83, 84*), estimating 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. In addition to the similarity with GTEx and CMC brain eQTLs, we evaluated the similarity between our eQTLs and GTEx eQTLs of other tissues using $\pi 1$ statistics and SNP-eGene overlap rates. The SNP-eGene overlap rates were calculated based on the percentage of shared SNPs associated with the same eGene using LD-independent eQTL SNPs. The π_1 values of liver, lung, testis, and blood eQTLs were 0.76, 0.88,0.9, and 0.88, respectively, which are each lower than the $\pi 1$ value 0.93 of GTEx brain eQTLs. We also

evaluated the SNP-eGene overlap rate of all LD-independent eQTL SNPs of different tissues. The SNP-eGene overlap rate was the highest in brain DLPFC among all the tissues tested (Figure 4B). We then used the π_1 statistic to investigate the sharing of SNPs between different types of QTLs in our study. We merged our cQTL list with the cQTL list from CMC cohort for this analysis. 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.

A list of the identified multiQTLs are available on the website (resource.psychencode.org) as DER-12.

Note on the QTL naming convention on the website: We have provided a number of QTL files on the website, including several variations for the eQTLs and isoQTLs. We have followed the following convention for naming the files (with {a,b,c,d} indicating the sequential label of the parameterization in the case of multiple parameterizations):
DER-##{a,b,c,d}_hg##_{e,c,f,iso,t}QTL.significant

S5.6: Univariate vs. multivariate-based QTL calculations

Although multivariate-based methods for identifying QTLs have been used elsewhere in the literature (*85, 86*), we consistently employed univariate-based approaches for all QTLs identified in our study (as an alternative to using multivariate methods for isoQTLs and fQTLs, and univariate methods for other QTLs). A number of factors have motivated this decision, including the need to more directly and reliably compare different QTL types, our desire to be as consistent as possible with the protocols used by GTEx, and the fact that the results from our univariate-based methods result in QTLs that contribute significantly to the predictive power of our integrative models. In addition, we note that the dependencies between phenotypes (for which multivariate methods are designed to resolve) are already taken into consideration by the DSPN models.

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

S6.1 Generation of Hi-C libraries

Hi-C libraries were generated as previously described (*87*). Briefly, adult DLPFC from three individuals (sample information provided below) were acquired through the 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 labeled by biotin-14-dCTP (ThermoFisher, 19518-018). Proximity-based ligation was performed within nuclei to prevent random collision-based ligation (*88*). 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 (*87*) 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 (*87*) (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 at a 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.

The Hi-C contact matrices at 10kb and 40kb resolution are provided on the website (resource.psychencode.org) as PIP-01 and PIP-02, respectively. The TAD regions are provided as DER-18.

S6.3 Detection of promoter-based interactions

Promoter-based interactions were identified as previously described (*87*). 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 a 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 (*89*) to evaluate the epigenetic state enrichment of regions that interact with promoters, using a 15-state chromatin model in adult PFC from Roadmap (*79*). 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), where 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 (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.

S6.4 Hi-C data comparison

Hi-C libraries for 14 tissues from Roadmap (*90*) and embryonic stem cells from ENCODE (*91*) were uniformly processed as described in S6.2, which included independently mapping the paired-end reads to the hg19 human reference genome, filtering the mapped reads at the restriction fragment level, and binning and normalizing the contacts maps. As Hi-C datasets from (*90*). have much lower coverage than Hi-C data for fetal and adult brain (see S6.2), the contacts were binned in 1Mb, and then corrected with iterative correction and eigenvector decomposition (ICE) at the whole genome level. We used two metrics to measure the similarities of chromatin structures in different tissue/cell types. First, we calculated PC1 values in the same manner as we analyzed compartments, but at 1Mb resolution, and calculated Pearson's correlations between PC1 values in different tissue/cell types. Second, we used HiC-spector (*92*) to quantify the Hi-C data similarity. For a pair of samples, the median value of the 23 HiC-spector similarity scores for the 23 chromosomes (chr1 to chr22 and chrX) was used to quantify the overall data similarity.

As this cross-tissue comparison involves four independent studies (PEC) (*87, 90, 91*), we note that the comparison can be affected by a range of confounders, including sequencing depths, laboratories that performed library preps, and batches. However, to our surprise, the comparison did not show signs of batch effects, as (1) chromatin structures of adult brain tissues clustered together even though they were generated from different labs, and (2) fetal brain (PEC) did not cluster with adult DLPFC (PEC) even though they were generated from the same lab (Fig. S38).

S6.5 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 their compartment categories, and centered expression values for each group were calculated. As our RNA-seq data are mainly focused on the adult brain transcriptome, we processed expression values from (*93*) to generate centered expression values(*93*). 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 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.

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 (*79*) 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, and Enh, and repressive elements were defined as Het, ReprPC, ReprPCWk (weak Polycomb repressive sites), and Quies (quiescent states). These elements are referred as developmental regulatory elements. Because 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 six types of developmental regulatory elements: both active promoters, both active enhancers, fetal active promoters, fetal active enhancers, both active promoters, and both active enhancers.

We next assigned genes to developmental regulatory elements: elements that overlapped with promoter coordinates were directly assigned to their genes based on linear genome, whereas those that did 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, and both active enhancers were assigned based on both adult and fetal brain Hi-C data. In total, this analysis led to seven 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, and 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 resulted 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 enhancer number and gene expression. To measure the relationship between enhancer numbers and gene expression levels, we integrated promoter-based interactions, brain active enhancers, and expression data. As enhancers and Hi-C interactions were defined in different resolutions (Hi-C was defined at the 10kb bin level, whereas enhancers were defined at a much higher resolution), we clumped enhancers within 10kb bins so that they matched 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 centered 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 relationships is not 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).

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 the 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 were 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 of chromatin marks, as chromatin interactions within this range are undetectable.

An e-Gene/chromatin often has multiple QTLs due to 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 interact 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 rather than isoform switching. We also found that 6.5% of cQTLs have Hi-C evidence. Although this overlap is lower than what we found for eQTLs and isoQTLs, we think this reflects the low power of cQTLs (292 samples for cQTL vs. 1,387 samples for eQTL). In detail, 27.4% of eQTLs were supported by promoter-based interactions and 30.9% were supported by exon-based interactions, suggesting that exon-level interactions also have the potential 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 and 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 and 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 were more than two groups to compare, multiple testing correction was performed using FDR.

Imputed gene regulatory networks. We integrated and imputed all possible regulatory relationships in the PFC including the enhancers, transcription factors (TFs), miRNAs, and target genes in this resource. As shown in Fig. S42, 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 within 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) (*94*), with TFBS locations on the hg19 genome build found using the program FIMO from the MEME suite (version 4.11.4) (*95*) with a threshold of 0.00001. Two iterations of this genome-wide TFBS map were constructed: one with the repetitive regions of the genome masked out using RepeatMasker (version 3.2.7) (*96*), and the second without such masking applied. Downstream analyses were used to construct two parallel gene-regulatory networks (GRNs) with slightly different choices in parameters and procedures (described in the following paragraph). We designate these networks as GRN1 and GRN2. The repeat-masked TFBS map was used in GRN1, whereas the complete map was used in GRN2.

Next, we defined the promoter regions by a window of $\pm$1.25 kb (=2.5 kb in total) relative to the transcription start site (TSS), and used the PEC enhancer regions of uniform length 1 kb. 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. At this stage, the full DHS peaks were used, even if they overlapped with the promoter windows by just 1 base pair. Because 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. The DHS overlap with promoters was carried out in the same manner for GRN1 and GRN2 up to this point. However, we carried out a further processing step for GRN2: we first truncated the DHS overlap peak to a region directly overlapping with the promoter window defined above (i.e., the direct intersection of the DHS peak with the promoter window), and then expanded these peaks to a uniform size of 1kb, so as to be consistent with the PEC enhancer length. The goal was simply to assess the importance of peak homogenization on TF linkage. Subsequently, the TFs with TFBSs within these open chromatin regions of the regulatory elements were linked to the corresponding elements. The details of the choices made for GRN1 and GRN2 are summarized in Figure S42.

Finally, we tentatively linked 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 was a set of preliminary linkages in the form of [Enhancer TFs]=>Enhancers=>Promoters<=[Promoter TFs].

There are some noteworthy points on this analysis. First, when the PEC enhancers were expanded to a uniform size of 1 kb, there was some overlap 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. Second, 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 linkages 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 (resource.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 assume that its gene expression is determined by a linear combination of the expression levels of TFs. In particular, 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 binding sites on the gene's enhancers and promoter. We then identified the activated TF-target regulatory relationships if TFs had 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, $B= \text{argmin}_B ||Y-XB||^2+\text{alpha}*||B||_{L2}+(1-\text{alpha})*||B||_{L1}$, where alpha is a parameter to adjust the contributions from L_2 and L_1 regularizations of B. The mean square error of Elastic net regression is equal to $||Y-XB||^2/N$ 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 a mean square error < 0.034 (R^2=0.77). In summary, we found a number of TF-target gene pairs with high absolute Elastic net coefficients (302, 214 pairs > 0.2 and 644, 292 pairs > 0.1 in GRN1; 310, 905 pairs >0.2 and 664, 371 pairs > 0.1 in GRN2).

Finally, we compared the HiC enhancer-promoter interactions and the interactions between eGenes and associated e/isoQTLs on enhancers with TF-target gene activity to determine highly confident, overlapped enhancer-target-promoter linkages from GRN1 or GRN2. In particular, we found that there were 43,181 TF-to-target-promoter and 42,681 enhancer-to-target-promoter linkages among the top 5% Elastic net regression coefficients (absolute value >0.2 in GRN1), covering 11,573 protein-coding target genes that are supported by at least two of the three data sets we examined (Hi-C, QTLs, or activity relationships): (i) activity relationships (447,919 enhancer-to-target-promoter linkages), (ii) physical chromatin interactions (73936 Hi-C enhancer-promoter interactions), and (iii) 52,449 QTLs (e/isoQTL-SNP on brain enhancers to eGene). We hypothesized that cell type-specific expression profiles are caused by different regulatory linkages in different cell types. Therefore, we employed this network to find the linkages for cell type biomarker genes defined by single-cell profiles (*64*). Potential cell type-specific regulatory linkages (e.g., various excitatory and inhibitory neuronal cell types, astrocytes, microglia) are shown in Fig. S44.

Associated data files with the final, Elastic-Net-Based TF network and HiC-derived enhancer-promoter linkages are on the website (resource.psychencode.org). The pre-ElasticNet-filtered reference networks are provided as INT-10 (GRN1) and INT-13 (GRN2), the ElasticNet-filtered networks as INT-11 (GRN1) and INT-14 (GRN2), and the ElasticNet mean square errors as INT-12 (GRN1) and INT-15 (GRN2). The HiC-derived enhancer-promoter linkages are provided as INT-16.

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

S7.1 Identification of GWAS associated genes for schizophrenia

We used 5,996 SCZ-associated autosomal putative causal (credible) SNPs reported in the original study (*97*) 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, whereas 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 (absolute Elastic net coefficient > 0.1 in GRN1, see Section S6.5).

Credible SNPs colocalized with 2,064 eQTLs associated with 282 eGenes, 91 of which overlapped with those identified by the Hi-C-driven approach. To confirm that this overlap was mediated by the shared causal variants in GWAS and eQTLs, we performed a colocalization test (*98*), from which we identified 293 genes across 79 loci in which GWAS and eQTLs share causal variants.

Collectively, we identified 181 genes across 83 loci from the direct assignment, 592 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 145 genes across 26 loci from isoQTL associations. In total, this includes 1,111 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 genes (321 genes).

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), 32,986 genes have QTLs, and Hi-C has the genome-wide search space. Therefore, we defined a background gene list by taking the intersect of eGenes and E-T genes. For each metric, we took the intersect of SCZ risk genes and the background gene set and used them for Fisher's exact test.

To assess what fraction of SCZ genes has 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 (*99*), assigning 137 genes to 68 loci. Notably, our newly generated eQTLs identified twice as many genes as the CMC eQTLs.

Associated data files with the full list of 1,111 SCZ risk genes and the filtered list of 321 high-confidence SCZ genes are available on the website (resource.psychencode.org) as INT-18 and INT-17, respectively.

S7.2 Functional enrichment analysis

To assess whether SCZ genes and SCZ high-confidence 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, SCZ, and BPD) identified by (*71*), (2) genes affected by rare loss-of-function variants in SCZ (TADA<0.3; (*100*)), and (3) genes located in recurrent SCZ copy number variation (CNVs) (*101*). 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 clusters (i.e., neurons, astrocytes, OPC, oligodendrocytes, microglia, endothelial cells, fetal neurons, and the neuronal subcluster (excitatory and inhibitory neuronal cell types)) and we 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), where p = fraction of promoters/enhancers associated with credible SNPs, n = the number of total binding sites of a TF A (TFBSA) in promoters/enhancers, and s = the number of total promoter/enhancer TFBSA associated with credible SNPs (Fig. S45).

For promoter enrichment, p = the number of SCZ genes / the number of genes that have TF-target-promoter links from the elastic net, and 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, and 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. The gene regulatory network (GRN1 in this case) edges for the SCZ high-confidence genes are provided on the website (resource.psychencode.org) as INT-19.

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) (*102*) (https://github.com/bulik/ldsc/wiki/Partitioned-Heritability). We included nine brain disorder GWAS and three non-brain disorder GWAS (GWAS sets and sources in Table S9) 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, as LD scores count for LD. We also used top SNPs (pruned for LD r2>0.5) to ensure that the enrichment signal didn't come from the spurious LD structures, where we got similar enrichment results.

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

S8.1 Data

We integrated data described above into a single model connecting genotype, functional genomics, and phenotype data from PsychENCODE in the prefrontal cortex (PFC). 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 ten balanced train / test splits as described below, and we report the performance of all models averaged across these ten splits of the data. For the disease conditions, these splits contained equal numbers of cases and controls, whereas for age, gender, and ethnicity, only control samples were 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. In addition, for the cRBM, cDBM, and DSPN models, all functional genomics data were 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 TF-gene and enhancer-gene linkages estimated in the gene regulatory network analysis (Section S6).

8.1.1 Balanced Datasets

We first describe how the balanced datasets were created for SCZ, and then describe how the balanced datasets were created for the other high-level phenotypes using a similar process with small modifications. For SCZ, we divided 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 (male or female), the same ethnicity (Caucasian or African American, 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, or 90+). For each subset, we found all SCZ and control 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 partitioned 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 added equal numbers (t_1, t_2) of controls to each partition. We repeated the whole process ten times to generate ten data splits. The above process ensures that each training and test partition contains a 50/50 split of SCZ/CTR samples, and that the distribution of covariates (assay, gender, ethnicity, and age) is approximately the same for cases and controls in the training and testing partitions.

The same method was used to create balanced data splits for BPD. For ASD, due to the limited number of cases, we set $\tau = 0.8$, and balanced only for assay and gender (not for ethnicity and age range). For the non-disease phenotypes (age, gender, and ethnicity), a similar method was used but with the following modifications. Here, we used control samples only, and split the PEC data into subsets containing samples from a common assay, which were matched

on all covariates as above except the high-level phenotype being modeled. Then, equal numbers of samples were 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 gender, the binary values were male/female for ethnicity they were Caucasian/African American, and for age we binarized the trait as 0/1 such that 1 indicates that a sample is older than the median age of 51 (the median age binarization was used only when age was the modeled phenotype; for all other phenotypes age was balanced using the decade age bins as above). The above method generates ten data splits each of the following sizes (training/testing): SCZ (640/70); BPD (170/18); ASD (50/12); age (244/26); ethnicity (284/30); gender (312/34).

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

8.2.1 Logistic regression (LR)

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

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

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, $\sigma(a) = 1/(1 + e^{-a})$. As training and test sets are both balanced, for a test sample we used 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 performed feature selection by calculating the correlation of each predictor with the high-level phenotype:

$$s_j = \text{corr}([y_1, y_2, \ldots y_N], [x_{1j}, x_{2j} \ldots x_{Nj}]), \quad (2)$$

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, $\text{corr}(\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b}/(||\mathbf{a}|| \cdot ||\mathbf{b}||)$. To perform feature selection, we ranked the predictors by the absolute value of s_j in descending order for a given training split, and included only predictors $1 \ldots \lceil \pi N \rceil$ in the model for that data split. We learned 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 used the Matlab Statistics and Machine Learning toolbox (glmfit). Model parameters for the genotypic (LR-gen) and transcriptomic (LR-trans) LRs are provided on the website (resource.psychencode.org) as the files INT-01 and INT-02, respectively.

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 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 (or multivariate binary in the case of $\mathbf{x}$ and $\mathbf{h}$). An RBM has the form $p(\mathbf{x}, \mathrm{y}, \mathbf{h}) = \exp(-E_{RBM}(\mathbf{x}, \mathrm{y}, \mathbf{h}))/Z$, 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 (*103*), which we denote $\mathbf{z}$, and which are assumed to be discrete:

$$p(\mathbf{x}, \mathrm{y}, \mathbf{h}|\mathbf{z}) = \exp\left(-E_{cRBM}(\mathbf{x}, \mathrm{y}, \mathbf{h}|\mathbf{z})\right)/Z(\mathbf{z}),$$
$$E_{cRBM}(\mathbf{x}, \mathrm{y}, \mathbf{h}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{x} - [\mathbf{x}^{\mathrm{T}}, y]\mathbf{W}\mathbf{h} - [\mathbf{x}^{\mathrm{T}}, y]\mathbf{b_1} - \mathbf{h}^{\mathrm{T}}\mathbf{b_2},$$
$$Z(\mathbf{z}) = \sum_{\mathbf{x},\mathrm{y},\mathbf{h}} \exp\left(-E_{cRBM}(\mathbf{x}, \mathrm{y}, \mathbf{h}|\mathbf{z})\right), \quad (3)$$

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 can 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:

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

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 (*104*) for the RBM case). Approximate gradients for interactions involving y and $\mathbf{z}$ and the bias terms can 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, $\Delta_t\left(w_{ij}\right)$, can then be calculated as:

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

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

Given a test sample, we aimed 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 (*104*) 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 ten balanced data split above to train a series of models for each phenotype. We initially performed 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 included all eSNPs associated with these genes in $\mathbf{z}$. We also enforced sparsity on the matrix $\mathbf{V}$ during training, so that only connections supported by eQTLs were allowed to be non-zero. Further, we set $N_h = 400$ (the number of hidden nodes), $\alpha = 0.1$, and $\epsilon = 1e-4$, and used mini-batches of size 61, 10, 17, 71, 39, and 64 for age, ASD, BPD, ethnicity, gender, and SCZ models, respectively. For τ_{stop}, we used either a variable setting that was set independently for each model trained, or a fixed setting that 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, whereas for the fixed setting we chose the τ_{stop} that had the minimum mean test error across data-splits. Results are shown using both variable (Fig. 7C) and fixed (Table S10) settings for all phenotypes except ASD; for ASD

we used only the fixed setting to control for the smaller number of samples in the ASD cohort. Performance for each phenotype was calculated as an average across data splits for the accuracy of a model on its corresponding test partition. Model parameters for the cRBM are provided on the website (resource.psychencode.org) as the file INT-03.

8.2.3 Conditional Deep Boltzmann Machine (cDBM)

A DBM may be defined as in (*104*) 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 $E_{DBM}(\mathbf{x}, \mathbf{y}, \mathbf{h_1}, \mathbf{h_2}) = -\mathbf{x}^{\mathrm{T}}\mathbf{W_1}\mathbf{h_1} - \mathbf{h_1}^{\mathrm{T}}\mathbf{W_2}\mathbf{h_2} - \mathbf{h_2}^{\mathrm{T}}\mathbf{W}_{lab}\mathbf{y} - \left[\mathbf{x}^{\mathrm{T}}, \mathbf{h_1}^{\mathrm{T}}, \mathbf{h_2}^{\mathrm{T}}, \mathbf{y}\right]\mathbf{b}$. 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, as for convenience we assumed 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, **z.** This is equivalent to converting the DBM from a Markov random field into a conditional random field (*105*). We can thus define a conditional DBM analogously to the cRBM:

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

The cDBM can be trained by adapting the Persistent Markov Chain Monte Carlo algorithm used in (*104*). In this approach, following a pre-training phase that 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:

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

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 were then made to the weights as in Eq. 5 (substituting pMCMC(.) for CD(.)). Finally, as in (*104*) back-propagation can be applied for fine-tuning, and we used 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 DSPN as a conditional DBM, with extra structure added to the visible units to reflect regulatory relationships between various intermediate phenotypes. The general form of the model is:

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

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 required that $\mathbf{x}$, $\mathbf{U}$, and $\mathbf{V}$ have specific forms, such that:

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

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) were 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 did 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 used 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 were 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 could not be made in parallel for the visible layer unlike the cDBM; for this reason, we chose a random update schedule of the nodes within the visible layer on each iteration, and updated 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 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

adopted a slightly different training process when the goal was 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 performed 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), $N_{fantasy} = 5$, $\alpha = 0.1$, and $\epsilon = 1e-4$, and used variable/fixed settings of τ_{stop} and mini-batch sizes as described above for the cRBM. Model parameters for the modular (DSPN-mod) and full (DSPN-full) DSPN are provided on the website (resource.psychencode.org) as INT-05 and INT-06, respectively. We make the DSPN code available at github.gersteinlab.org/PsychENCODE-DSPN . Summary data files are at resource.psychencode.org. Data formatted for use with the code are at the github site.

S8.3 Imputation of intermediate phenotypes

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

To optimize performance for prediction of high-level phenotypes from genotype data with imputation of intermediate phenotypes, we adopted a specialized training process. We assume that during training we would 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 trained 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 fixed the $\mathbf{V}$ matrices of the DSPN directly to the coefficients of the logistic regression models, and trained $\mathbf{U}'$ 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; as 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 used the Contrastive Divergence (Eq. 5) for optimization. Additionally, we performed 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 ordered by their predictive accuracy using the initial logistic predictor. We then used 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 labeled $\mathbf{x_{impute}}$ (we inferred $\mathbf{x_{impute}}$ 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 trained a full DSPN (with $\mathbf{V}$ still fixed) on the validation data but optimized using the imputed rather than the original intermediate phenotype data, that is using $(\mathbf{z}, \mathbf{x_{impute}}, \boldsymbol{y})$ as training samples.

At test time, we did not make use of the intermediate phenotype data. Instead, we followed 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 treated the imputed phenotype data as fixed, and predicted the associated high-level phenotype data from the full DSPN model using a forward pass as described for prediction in the cDBM model. We trained the imputation-based DSPN model using the same hyper-parameters as for the DSPN above, setting $\pi_{\text{gene}} = 0.01$, $\pi_{\text{enh}} = 0.01$, $\pi_{\text{frac}} = 0.5$, $\pi_{\text{mod}} = 0.05$. Model parameters for the imputed DSPN (DSPN-impute) are provided on the website (resource.psychencode.org) as INT-04.

S8.4 Variance explained on liability scale

To covert predictive performance of all models onto the liability scale, we used the following conversion due to Falconer (*106, 107*):

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

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 $a, b, c,$ and d be the true negatives, false negatives, false positives, and true positives, respectively, for a given model on test data, we estimated $p_{pos} = (c + \left(\frac{K}{1-K}\right) d)/(a + c + \left(\frac{K}{1-K}\right)(b + d))$, and $\mathrm{GRR} = (\frac{d}{c+d})/(\frac{b}{a+b})$. We set $K = 0.011, 0.01,$ and 0.015 for SCZ, BPD, and ASD, respectively.

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

To provide interpretation of the DSPN model, we developed 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 that are most 'important' in influencing the value of the node of interest. In our analysis, we took 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 used 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 had 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, to 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 defined $S_{(m,+)} = \{n^*\}$ and $S_{(m,-)} = \{\}$. Then, for all $l < m$:

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

where we define the sets $B_{(n,+)}, B_{(n,-)} \subset L_l$ for $n \in L_{l+1}$ as $B_{(n,+)} = \{n' \mid \mathrm{rank}_n^+(n') \leq B\}$ and $B_{(n,-)} = \{n' \mid \mathrm{rank}_n^-(n') \leq B\}$, where the function $\mathrm{rank}_n^+(n')$ 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 fixed the 'projection layer' l to be the co-expression module sublayer in the DSPN (L_{1d} in Fig. 7A) and found the sets $S_{(l,+)}$ and $S_{(l,-)}$ when n^* was set to the output trait node. We repeated this analysis for models trained on the ten splits of the data for the given trait, generating ten positive and negative projected sets. For module $n_{l,i}$, we then calculated the counts $c_{(i,+)} = \sum_t [n_{l,i} \in S^t_{(l,+)}]$, where $S^t_{(l,+)}$ was the positive projected set from the model trained on data split t, and $c_{(i,-)} = \sum_t [n_{l,i} \in S^t_{(l,-)}]$. For our final list of positive and negative prioritized modules we used $S^+ = \{n_{l,i} \mid c_{(i,+)} > \tau_{\text{prioritize}}\}$ and $S^- = \{n_{l,i} \mid c_{(i,-)} > \tau_{\text{prioritize}}\}$, respectively. The threshold $\tau_{\text{prioritize}}$ was set such that $p\big(c_{(i,+)} > \tau_{\text{prioritize}} \mid B\big) < \alpha$ under a null distribution where the network weights were sampled from a standard normal distribution and the same branching factor B was used. We set $\alpha = 0.001$, and evaluated $\tau_{\text{prioritize}}$ using 10,000 simulations. Setting $B = 4$, we found that this implied an estimate of $\tau_{\text{prioritize}} = 3$, and generated ~ 60 positive and negative prioritized modules per trait (out of $\sim 5{,}000$).

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. 7A), for each data split we found nodes $n_{l+1,j}$ and $n_{l+2,k}$ such that $(n_{l,i}, n_{l+1,j}, n_{l+2,k})$ formed the 'best path' from module $n_{l,i}$ to the trait output node in the sense that it minimize the score:

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

across all 'positive' paths, meaning:

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

and ties were broken arbitrarily (a similar annotation can be made for negative paths by placing -1 on the RHS of Eq. 14). Writing $n^t_{l+1,j},\ n^t_{l+2,k}$ 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, $c_{(i',+)} = \sum_t [n_{l,i'} \in S^t_{(l,+)}(n^t_{l+1,j})]$ and $d_{(i',+)} = \sum_t [n_{l,i'} \in S^t_{(l,+)}(n^t_{l+2,k})]$, where we write $S^t_{(l,+)}(n)$ for the positive projected set at level l for data split t when we set the node of interest $n^* = n$. We then evaluated $S^+_c = \{n_{l,i'} \mid c_{(i',+)} > \tau_{\text{prioritize}}\}$ and $S^+_d = \{n_{l,i'} \mid d_{(i',+)} > \tau_{\text{prioritize}}\}$ where $\tau_{\text{prioritize}}$ was defined as above, and annotated 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' (*108*) using KEGG pathway annotations, and setting the p-value and q-value cutoffs to 0.05 and 0.1, respectively. Further, we performed enrichment analysis for the cell-type marker genes corresponding to the cell-types used in our single-cell deconvolution analysis. Here, we thresholded 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 tested for enrichment of cell type markers using the hypergeometric test with a p-value cutoff of 0.1. Finally, we compared the modules prioritized for our SCZ model using the above approach with those prioritized using a gradient-based approach, following (*109*) 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) was used to prioritize salient input nodes (modules).We found the results 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 (*110*), which attempted 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).

To rank functional terms as in Fig. 8B-C, we count the number of prioritized modules (MODs) and higher-order groupings (HOGs) they associate with. A separate ranking is established for terms using positive and negative prioritized MODs/HOGs (with ties broken using nominal p-values), and the rank-score for a term in a disease is the highest rank it receives in either ordering. For the cross-disorder rankings shown in Figs. 8C and S49, all terms are included receiving a rank between 1-10 in each disorder individually; a modified rank-score of 11, 12, 13, 15 is assigned to represent rank-scores 11-50, 51-100, 101-150, 151-200 and >200 respectively from the individual disease rankings, and the cross-disorder ranking is ordered using the average of the modified scores.

The module enrichment scores are provided on the website resource.psychencode.org as the file INT-07, and the WGCNA modules are provided as INT-08 (in terms of Ensembl IDs) and INT-09.

S9. Resource website and raw data

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

The website contains a large amount of supplementary information related to this project, including the raw and processed data files. For convenience, we reproduce below some sections of the site and from the PsychENCODE Synapse website related to the descriptions of the data files. Additionally, we have created an interactive website psychint.psychencode.org, linked from resource.psychencode.org, where we allow QTL queries, and the exploration of the QTL maps and the PsychENCODE enhancer set aligned to both hg19 and hg38 in UCSC browser tracks. We also provide links to network visualization tools hosted externally on this interactive website.

The following is a list of the files hosted on resource.psychencode.org:

Bulk download of the Integrative and Derived datasets is available. A link to the interactive portal at psychint.psychencode.org is also provided.

Integrative Analysis:

1. INT-01_LR_gen
2. INT-02_LR-trans
3. INT-03_cRBM
4. INT-04_DSPN-Imput
5. INT-05_DSPN-mod
6. INT-06_DSPN-full
7. INT-07_DSPN_prioritized_module_enrichments
8. INT-08_WGCNA_modules_ensembl_ids
9. INT-09_WGCNA_modules_hgnc_ids
10. INT-10_Reference_Network_GRN_1
11. INT-11_ElasticNet_Filtered_Cutoff_0.1_GRN_1
12. INT-12_ElasticNet_Mean_square_error_GRN_1
13. INT-13_Reference_Network_GRN_2
14. INT-14_ElasticNet_Filtered_Cutoff_0.1_GRN_2
15. INT-15_ElasticNet_Mean_square_error_GRN_2
16. INT-16_HiC_EP_linkages
17. INT-17_SCZ_High_Confidence_Gene_List
18. INT-18_SCZ_Risk_Gene_List
19. INT-19_GRN_scz_hc_genes

Derived Data Types:

1. DER-01_PEC_Gene_expression_matrix_normalized
2. DER-02_PEC_Gene_expression_matrix_TPM
3. DER-03a-b_hg19_PEC_enhancers
4. DER-04a-b_hg38lft_PEC_enhancers
5. DER-05_PFC_H3K27ac_peaks
6. DER-06_TC_ H3K27ac_peaks
7. DER-07_CBC_ H3K27ac_peaks
8. DER-08a-d_hg{19,38}_eQTL.significant
9. DER-09_ hg{19,38}_cQTL.significant

10. DER-10a-d_ hg{19,38}_isoQTL.significant
11. DER-11_ hg{19,38}_fQTL.significant
12. DER-12_capstone4.multiQTL.list
13. DER-13_Disorder_DEX_Genes
14. DER-14_Disorder_DEX_Transcripts
15. DER-15_Disorder_DifferentialSplicing
16. DER-16_Disorder_Gene_Modules
17. DER-17_Disorder_Isoform_Modules
18. DER-18_TAD_adultbrain
19. DER-19_Single_cell_markergenes_TPM
20. DER-20_Single_cell_expression_processed_TPM
21. DER-21_ Single_cell_markergenes_UMI
22. DER-22_Single_cell_expression_raw_UMI
23. DER-23_Cell_fractions_Raw
24. DER-24_Cell_fractions_Normalized
25. DER-25_NMF_comp
26. DER-26_NMF_coef

Pipeline-Processing Results:

- Links to "RNA-seq Signal Tracks and Peak Files" and "ChIP-seq Signal Tracks and Peak Files" hosted on Synapse.org

1. PIP-01_DLPFC.10kb (Hi-C contact matrix at 10 kb resolution)
2. PIP-02_DLPFC.40kb (Hi-C contact matrix at 40 kb resolution)
3. PIP-03_all_TARs
4. PIP-04_all_TARs.70pc.active (all TARs active within at least 70% of the individuals)

Raw Data:
1. RAW-01_PEC_Table_of_Datasets
2. RAW-02_Assay_Cross_Reference

- Links to "RNA-seq files for alignment to both genome and transcriptome", "RNA-seq fastq files", "ChIP-seq files for alignment to genome", "Imputed Genotypes", "Metadata for Imputed Genotypes" and "Phenotype Information" hosted on Synapse.org
- Links to "RNA-seq alignment files" hosted on the GTEx portal, and "ChIP-seq alignment files" hosted on the Roadmap Epigenomics Consortium portal.

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 nine 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 of Bipolar cohorts. The detailed descriptions of PsychENCODE related eight studies are listed below and may also be found in Supplemental Table S11, as well as in the PsychENCODE Knowledge Portal (https://www.synapse.org/#!Synapse:syn4921369/wiki/390659). RNA-seq datasets for Study 1-8 were described in the supplement of the Capstone 1 paper (*71*). The processing flowchart from

the Capstone 1 and 4 papers are shown in the following figure. The description of the datasets available in Supplemental Table S11 is also made available on our website (resource.psychencode.org) as RAW-01. In addition, we provide a master file, RAW-02, for easy cross-referencing of genotype, RNA-seq and ChIP-seq samples from the same individuals.

Study 1 - BrainGVEX
Genotyping: DNA genotyping were done using two different platforms. A total of 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 (*111*). The rest of the 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) (*112*).

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 > 1×10^{-6}. A total of 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. Thus, 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 > 1×10^{-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
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 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 260 mm Tissue-Tek Accu-Edge trimming blade. The regions of interest were dissected using a scalpel blade and immediately processed. The sampled area corresponds to DLPFC and 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 a gentleMACS dissociator (Miltenyi Biotec). Cell suspension was then processed on a Fluidigm C1 machine to capture single cells, according to the manufacturer's protocol. RNA extraction from each single cell was carried out on a Fluidigm C1 machine, according to the 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

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 protocol can be found in the published methods paper (*113*). We used H3K27ac from DLPFC of 117 neurotypical controls and 109 SCZ individuals.

Genotyping: DNA was isolated from approximately 10 mg dry homogenized tissue 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 rates 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
Genotype data is not available yet for this study.

Study 5 - UCLA-ASD
Full details of the UCLA-ASD study have been published (*114*).

ChIP-seq: For each ChIP-seq experiment approximately 100 mg of frozen brain tissue per sample was aliquoted and thawed on ice in 1 ml 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 40 uM cell strainer (Falcon) by spinning at 2,000 RPM for 1 minute at 4°C 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 50 µl 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 (*115*). After 15 cycles of PCR using indexing primers, libraries were size selected for 300-500 bp on low melting agarose gel and four libraries were pooled and sequenced in one lane of 2 x 100bp using the same Illumina HiSeq 2,000 with V3 reagents.

Genotyping: Genotyping was performed using Illumina Omni 2.5 arrays.

Study 6 - BipSeq
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.

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

Study 8 - LIBD_szControl + BipSeq
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 originated 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 were processed separately. The same individual ID has been used on all three studies. The CMC data also has a set of ten 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 nine centers were merged into one gene expression matrix, and subsequently normalized using the protocol detailed by GTEx (*62*).

S10. Supplementary Figures

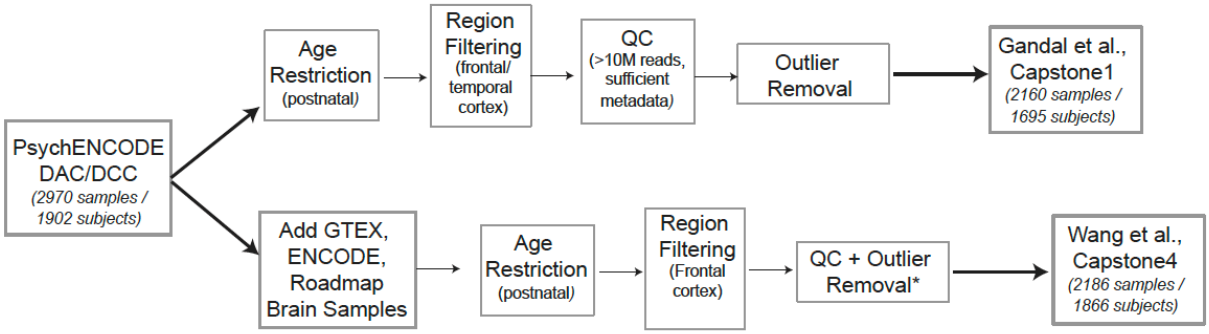


Fig. S1. RNAseq datasets filter flowchart in Capstone 1 and Capstone 4. Figure shows the flowchart of filtering RNAseq datasets in Capstone 1 and Capstone 4 papers. These two papers used different criteria to filter the datasets. The number of samples and subjects used for downstream analysis were not exactly the same in two papers. *Details of QC and outlier removal in Capstone 4 paper could be found in previous section S2.1.2. Note that the protocols for all associated data are provided in detail in Section S9.

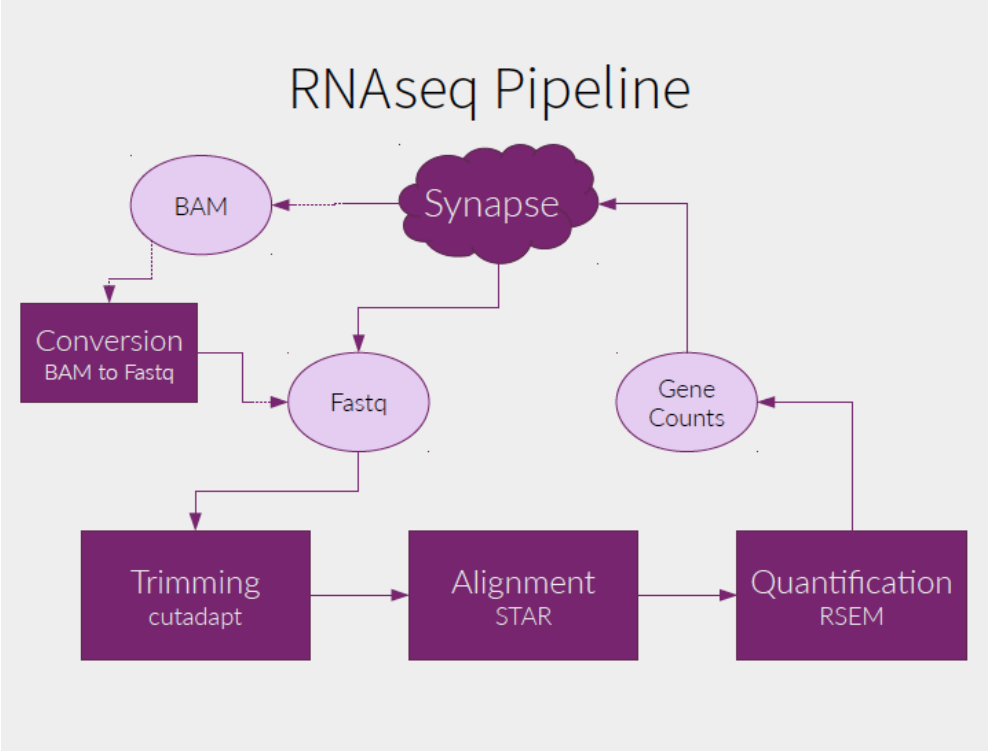


Fig. S2. PsychENCODE RNA-seq pipeline. The flowchart of the uniform RNA-seq pipeline is shown. This pipeline was modified based on the long-RNA-seq-pipeline used by the ENCODE Consortium. For details, see Section S2.1.2.

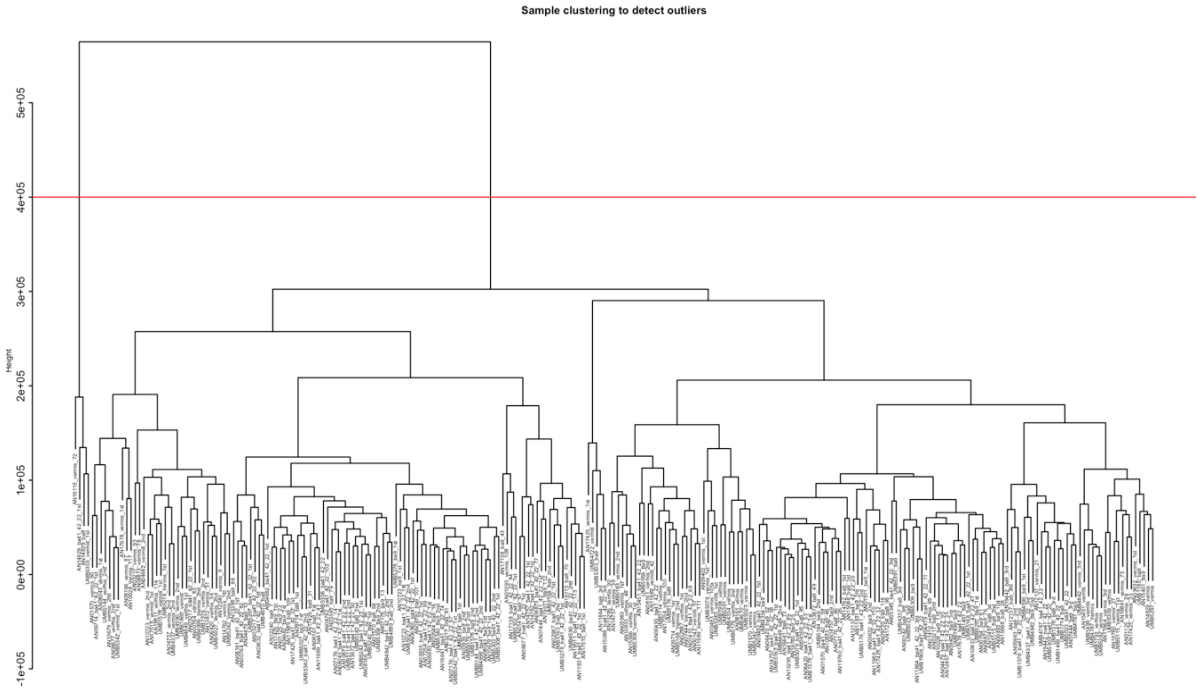


Fig. S3. Dendrogram of clustering analysis for identifying outliers of gene expression. An example of removing 4 outlier samples from a UCLA-ASD study according to hierarchical clustering of the gene expression data. For details, see Section S2.1.2.

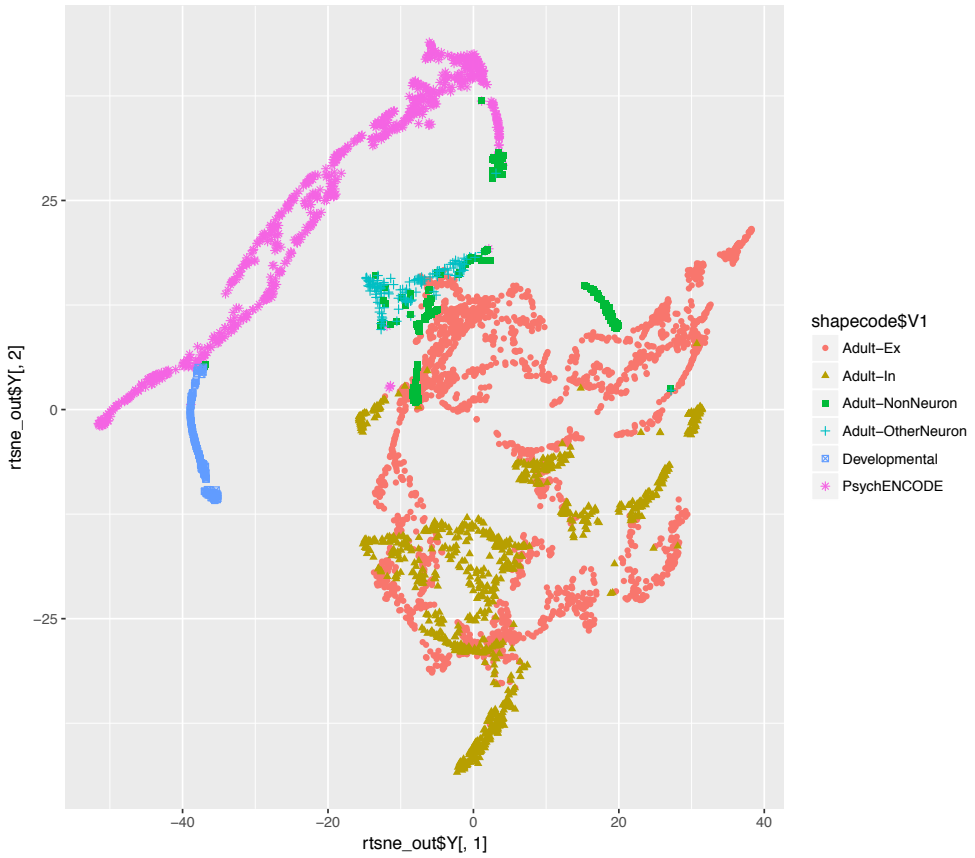


Fig. S4. t-SNE plot of the read count-based dataset. Most of the PsychENCODE data were found to be clustered together with public developmental data in (*65*). For details, see Section S2.2.

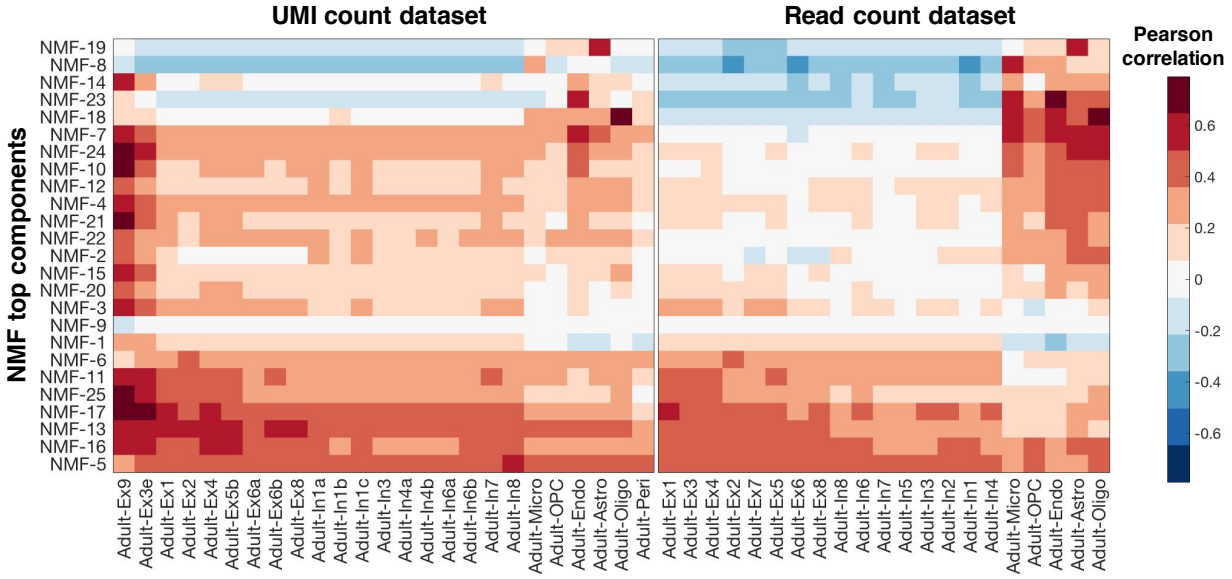


Fig. S5. Correlation between NMF-TCs and a single-cell dataset based on marker genes from the UMI count dataset. The UMI count dataset and read count data showed similar correlations with the NMF-TCs. For details, see Section S2.3.

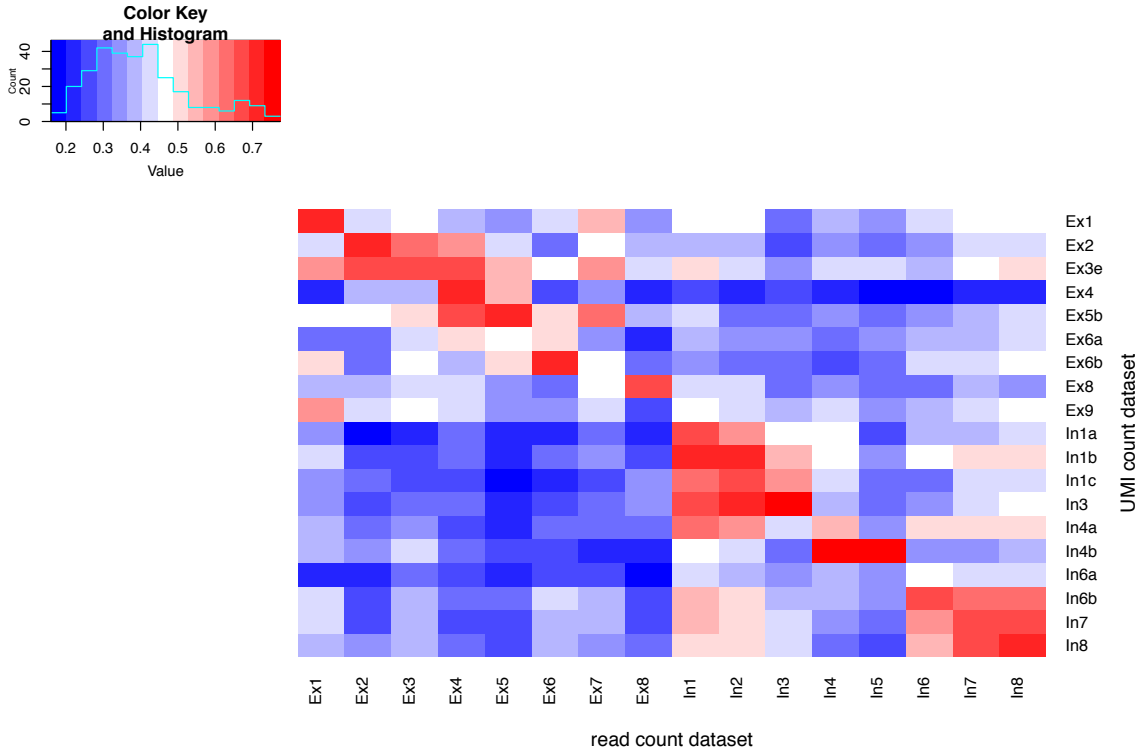


Fig. S6. Expression correlation of neuron cell types between the read count dataset and the UMI count dataset. The correlation pattern shows that the neuron cell types have very similar expression profiles between the two datasets. For details, see Section S2.2.

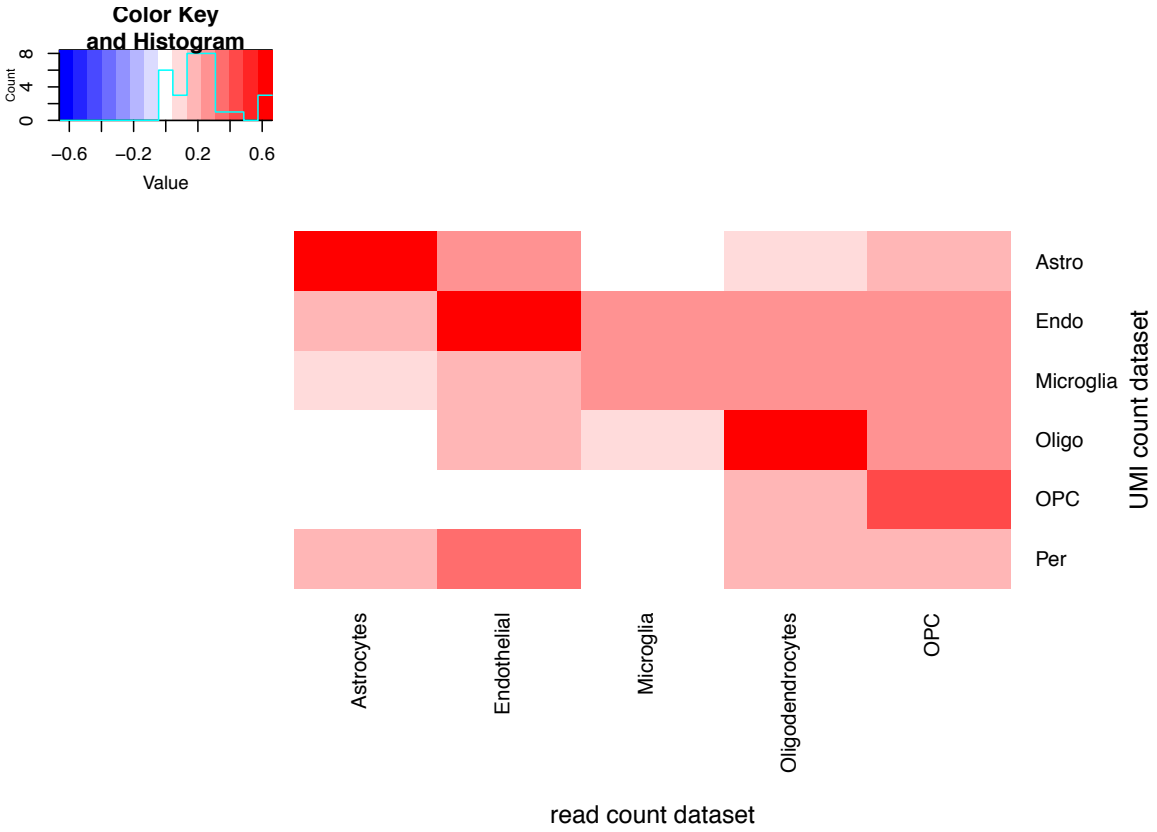


Fig. S7. Expression correlation of non-neuronal cell types between the read count dataset and the UMI count dataset. The correlation pattern shows that the non-neuronal cell types have very similar expression profiles between the two datasets. For details, see Section S2.2.

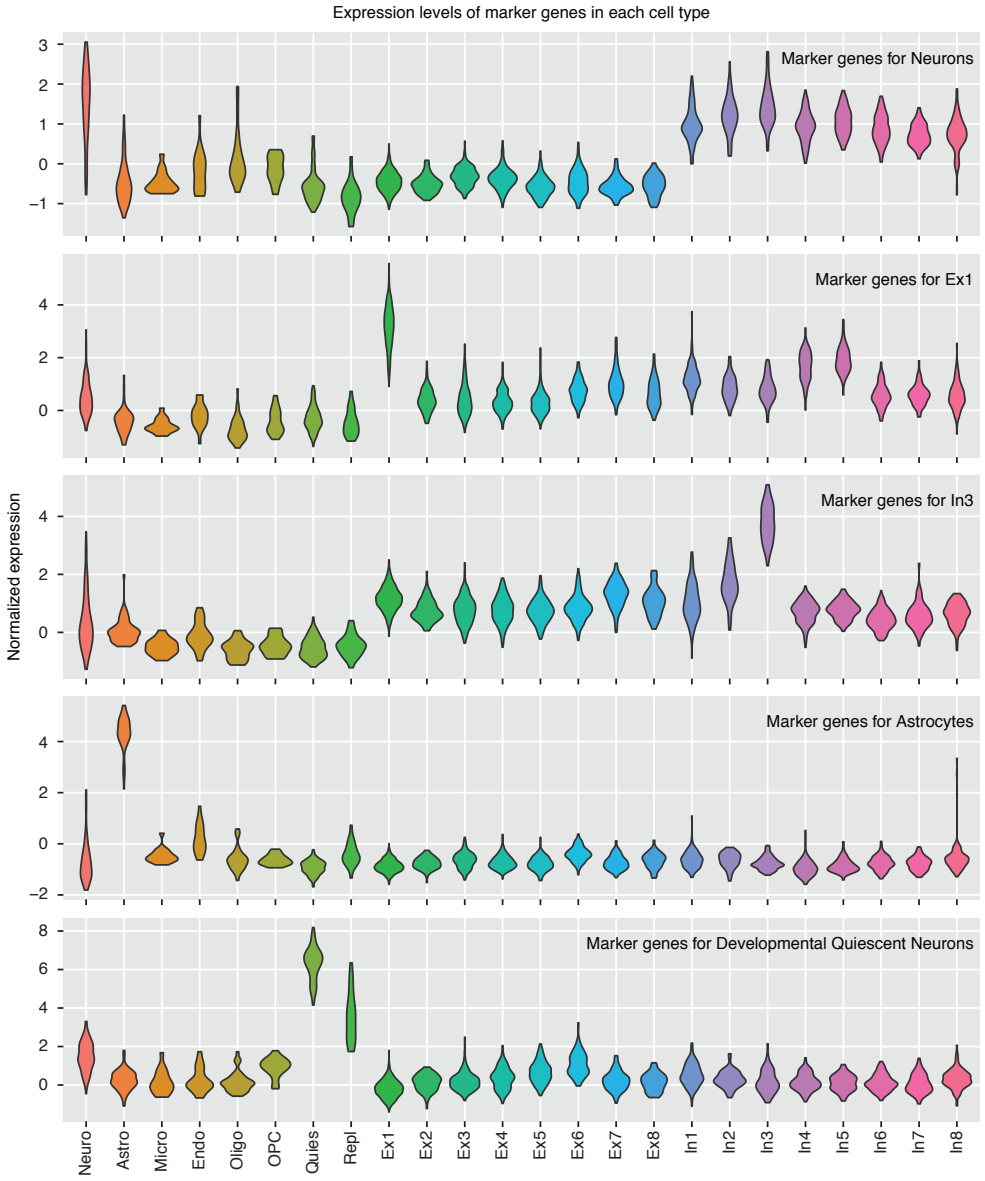


Fig. S8. Biomarkers show higher expression in the cell type from which they were defined compared to other cell types. Expression signatures of biomarkers are conserved in the newly constructed expression matrix, which integrates multiple sources of single-cell expression data. For details, see Section S2.2.

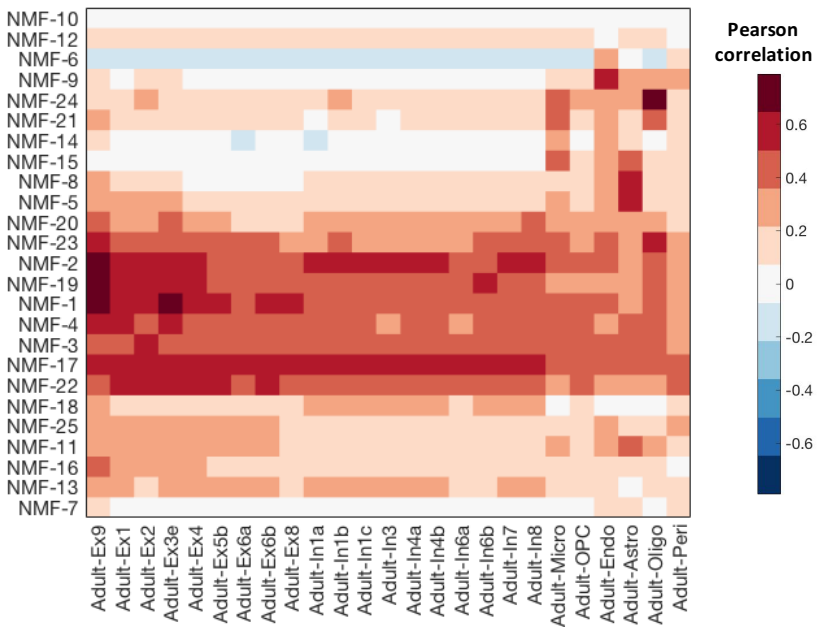


Fig. S9. Correlation between NMF TCs and the UMI count dataset. The correlation pattern is very similar to that of the read count dataset (Fig. 2C). For details, see Section S2.3.

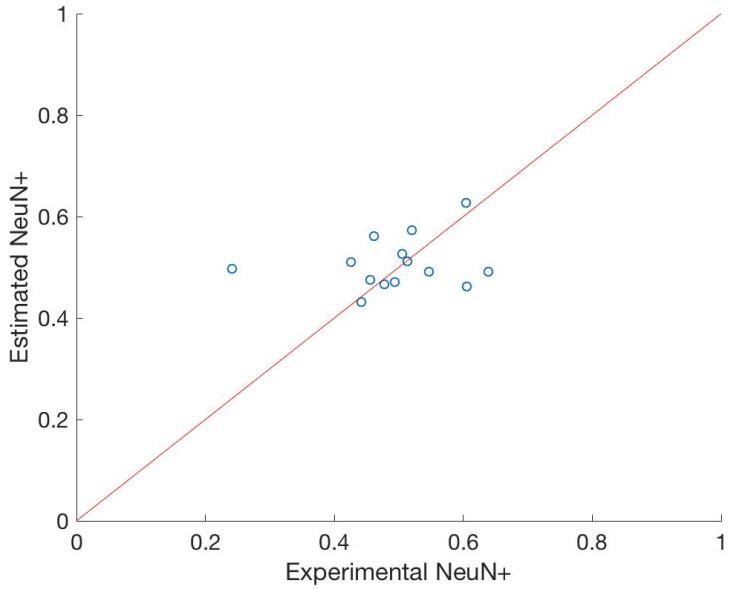


Fig. S10. Validation of estimated cell fractions from deconvolution. The X-axis shows the NeuN+ cell fractions measured in experiments and the y-axis shows the NeuN+ cell fractions estimated from deconvolution. The median error is 0.04. For details, see Section S2.4.

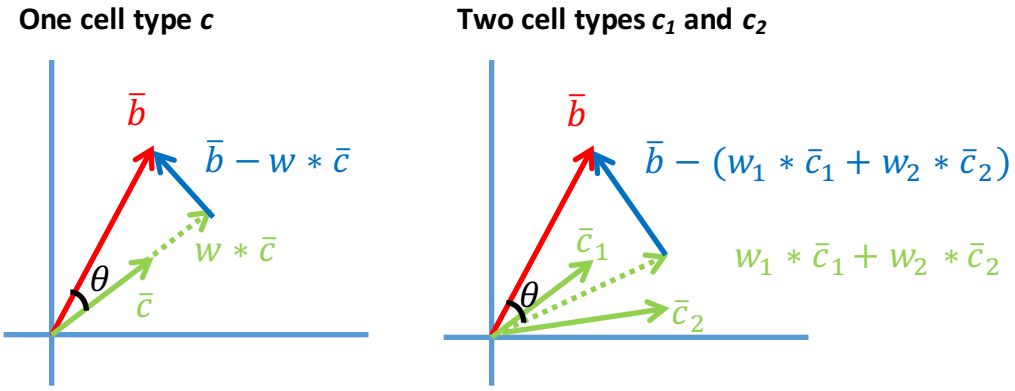


Fig. S11. Illustration of the reconstruction accuracy. The reconstruction accuracy was used to evaluate the reconstruction of bulk tissue expression. For details, see Section S2.4.

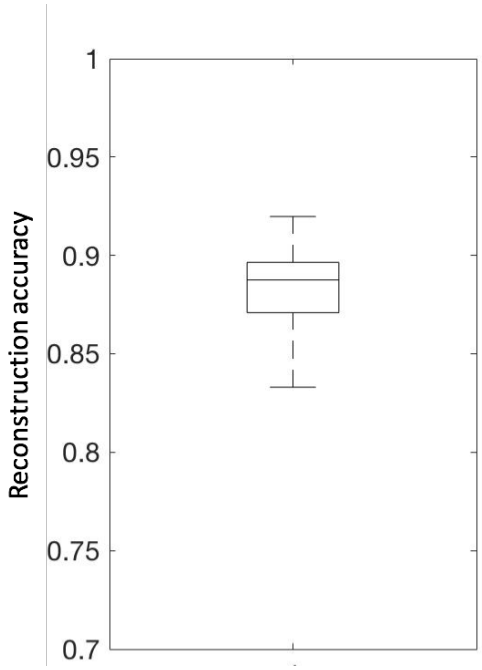


Fig. S12. Reconstruction accuracy of our deconvolution. The distribution shows that weighted combinations of single-cell expression signatures from our deconvolution can explain much of the individual variance of gene expression with high accuracy (mean~=88%). For details, see Section S2.4.

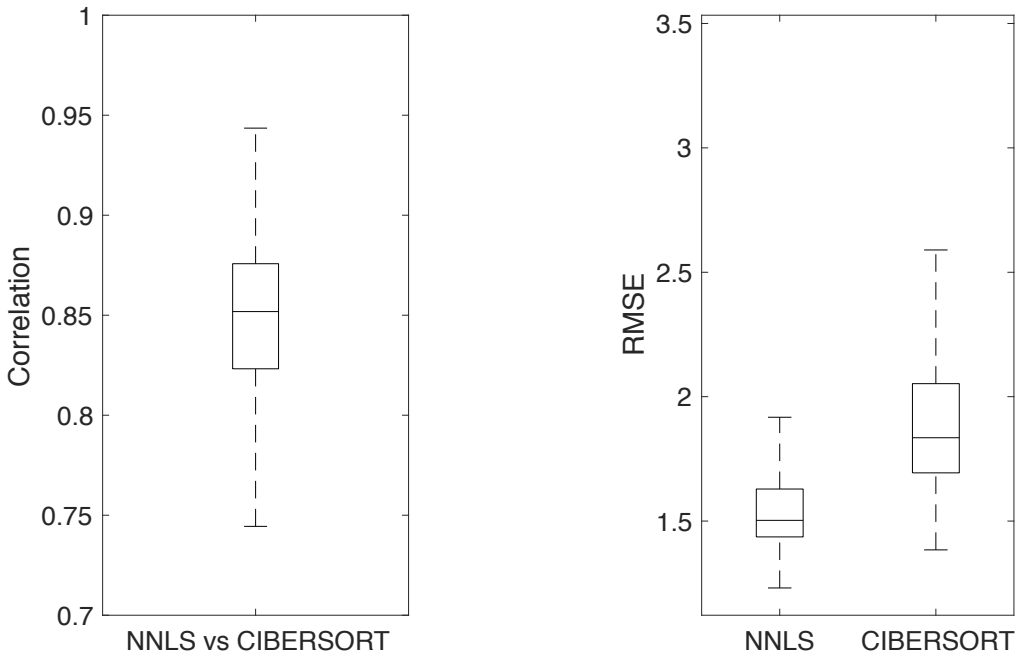


Fig. S13. RMSE and correlation distributions of our deconvolution (NNLS) and CIBERSORT. (Left) The boxplot shows the Pearson correlation coefficients of estimate cell fractions between our deconvolution and CIBERSORT across tissue samples (Median=0.85, Variance=0.0027), suggesting our consistency with CIBERSORT in terms of relative proportions of different cell types of individual tissues. (Right) The boxplot shows that our deconvolution has significantly smaller RMSEs than CIBERSORT (both KS-test and t-test p-values < 2.2e-16), suggesting our improvement over CIBERSORT in terms of cell proportion estimate accuracy. For details, see Section S2.4.

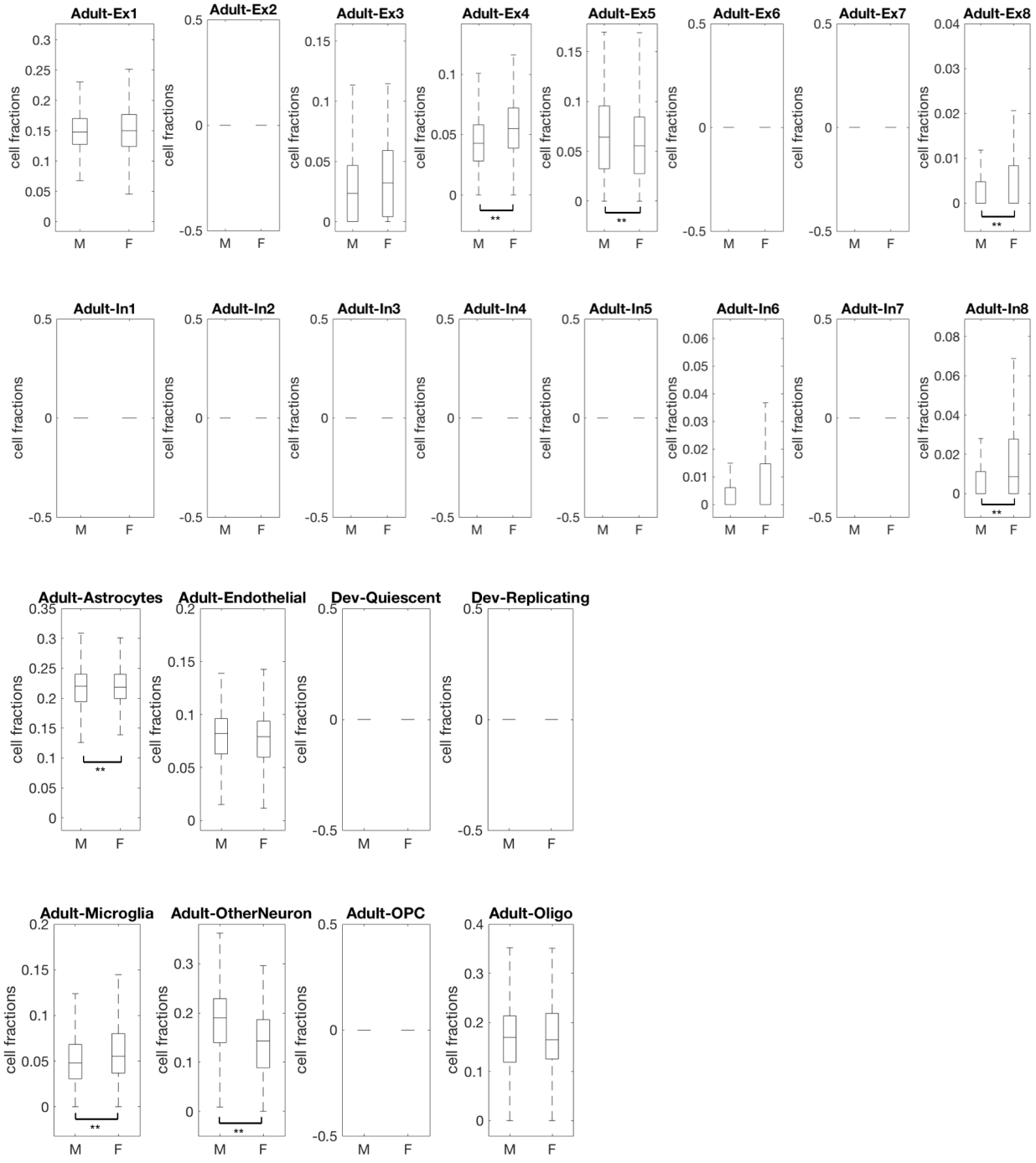


Fig. S14. Estimated cell fractions of 24 selected cell types in control samples. The cell types with significant changes (FDR < 0.05) between genders after balancing age distribution are labeled with double asterisks (**). For details, see Section S2.4.

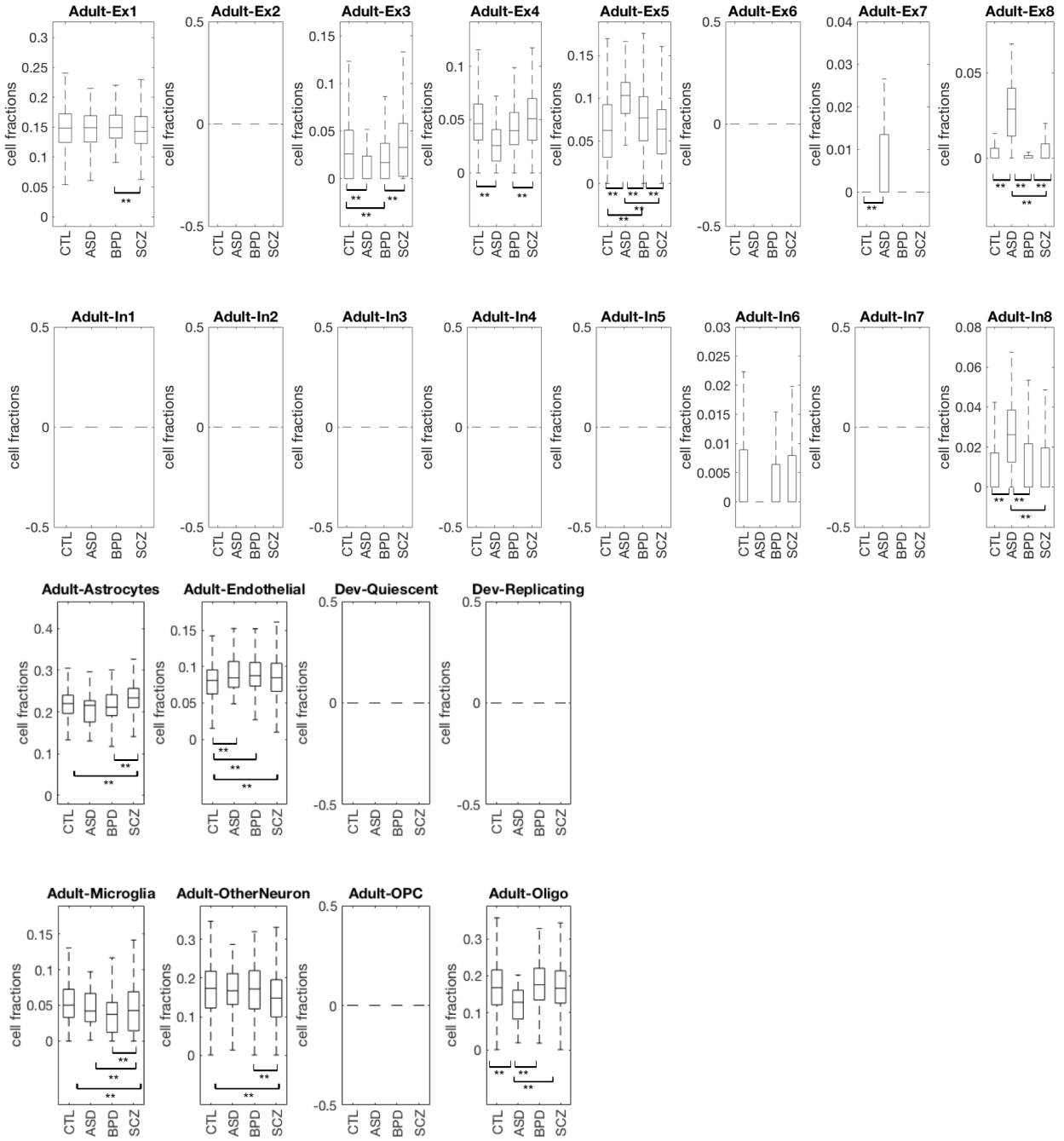


Fig. S15. Estimated cell fractions of 24 selected cell types in samples with different disorders. For each cell type, the boxes with double asterisks (**) indicate the disorder types that show significant differences (FDR < 0.05) after balancing the age distribution. For details, see Section S2.4.

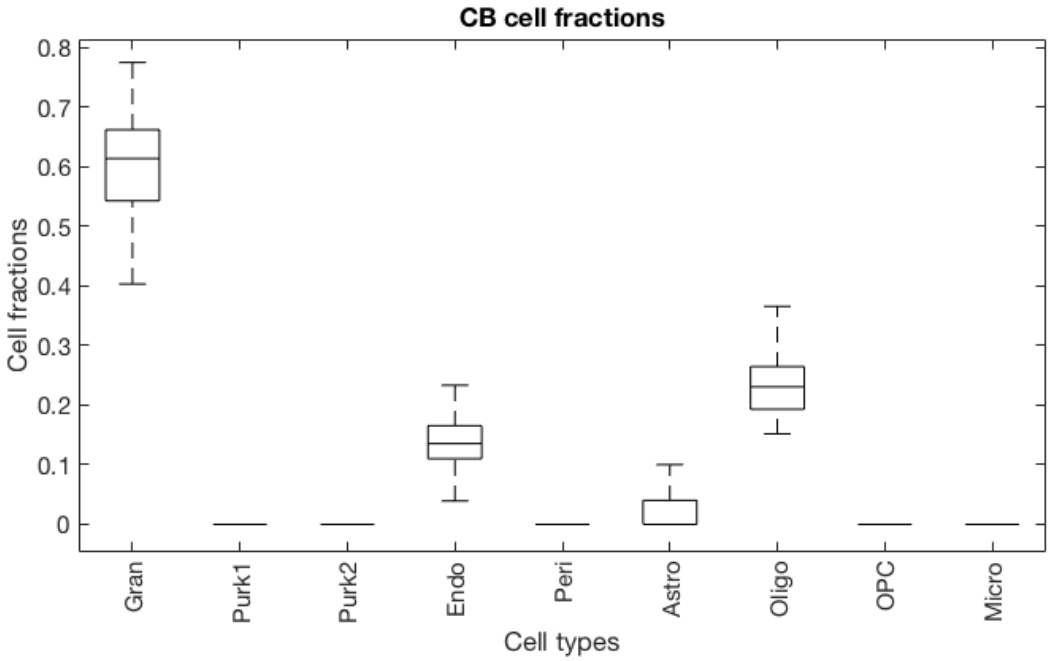


Fig. S16. Cell fractions from deconvolution of CB bulk tissue using a CB single-cell dataset. The figure shows the distribution of neurons and non-neuronal cell types of 69 samples. For details, see Section S2.4.

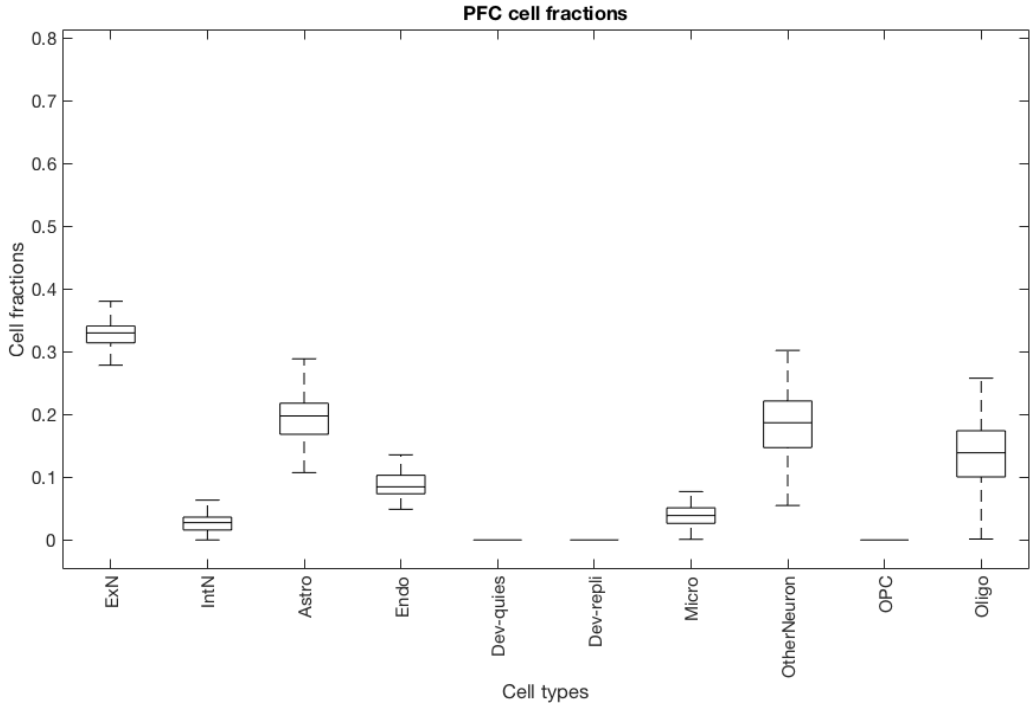


Fig. S17. Cell fractions of PFC bulk tissues from the samples that also have CB bulk tissue. The figure shows the distribution of neurons and non-neuronal cell types of 69 samples that also have CB bulk tissue data. For details, see Section S2.4.

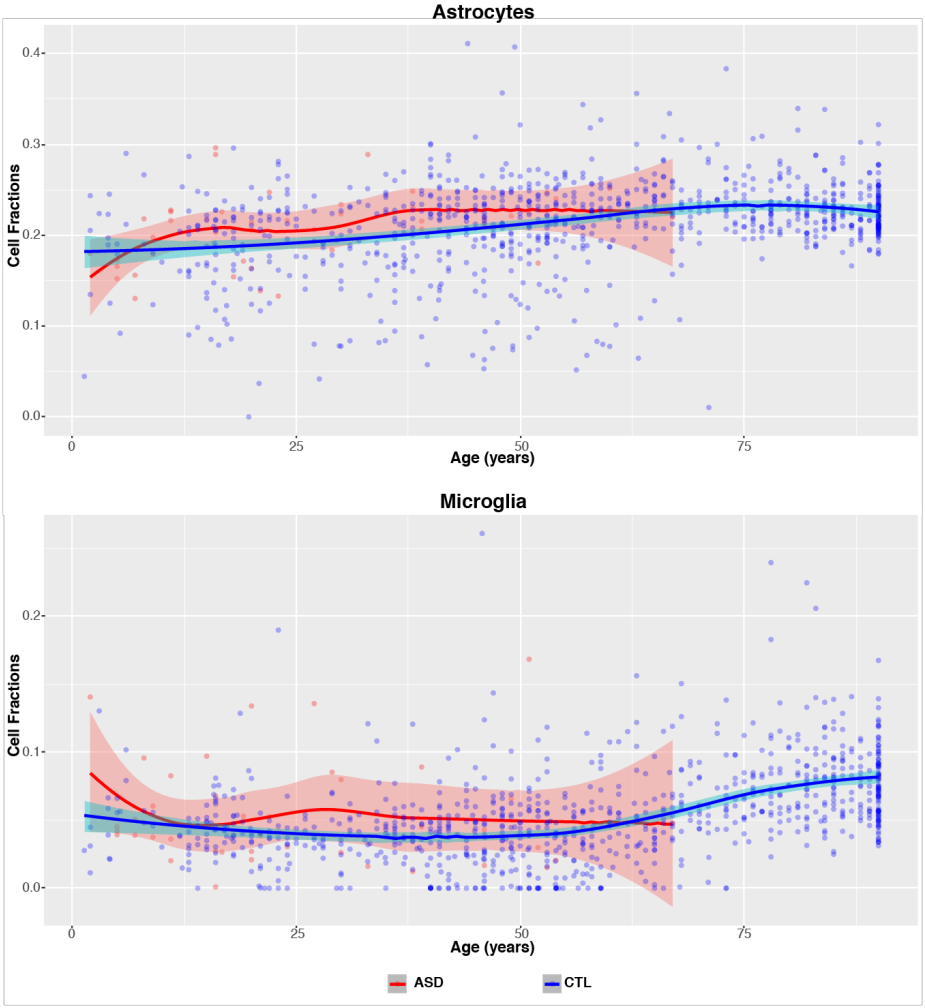


Fig. S18. Cell fractions of astrocytes and microglia of ASD and CTR samples across age. The red and blue points represent ASD and CTR individuals, respectively. The corresponding curves show the fitted trend. For details, see Section S2.4.

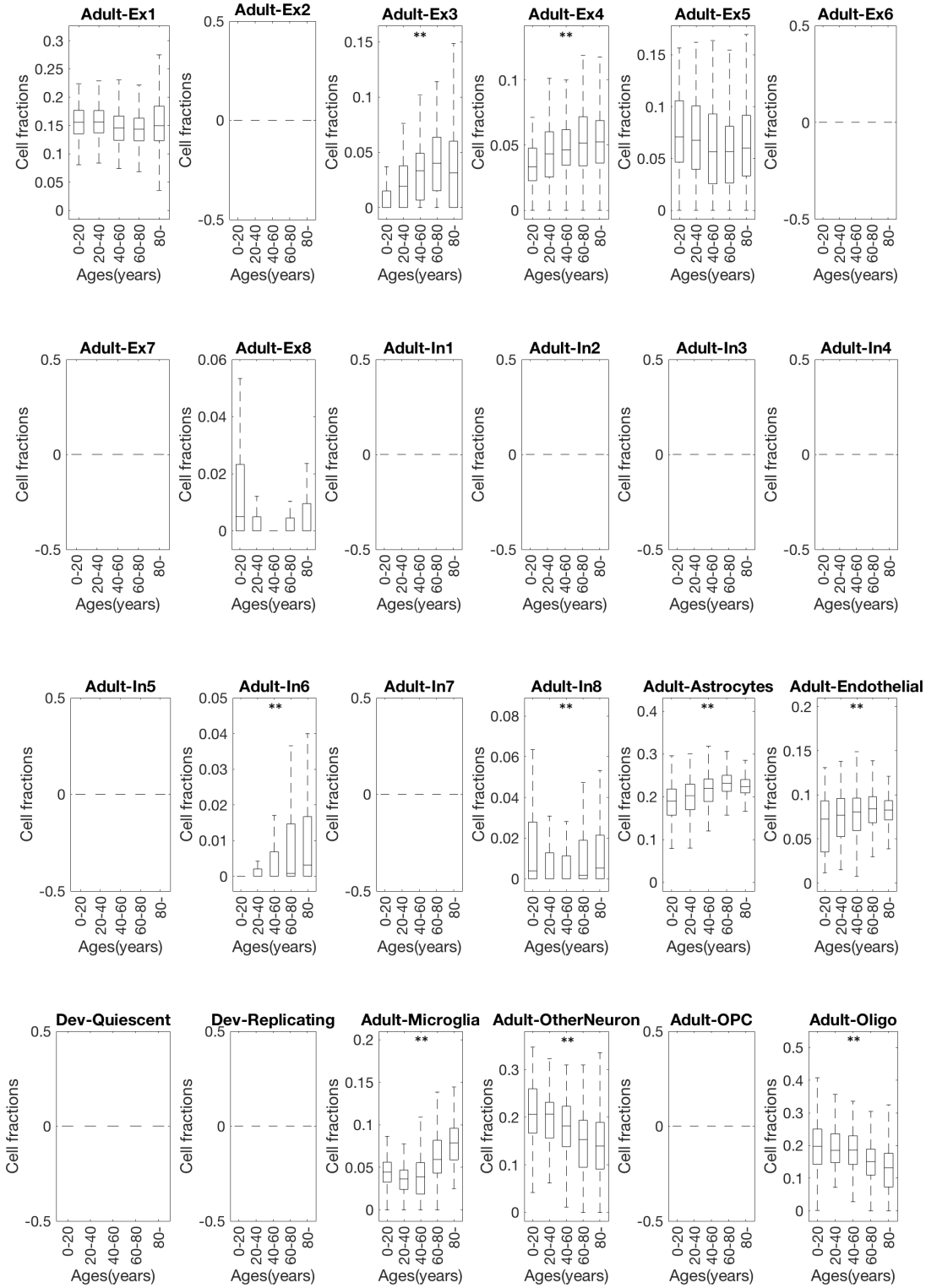


Fig. S19. Estimated cell fractions of 24 selected cell types in control samples with different ages. The cell types showing significant increasing/decreasing trends across ages (trend analysis p-value < 1e-2) are labeled with double asterisks (**). For details, see Section S2.4.

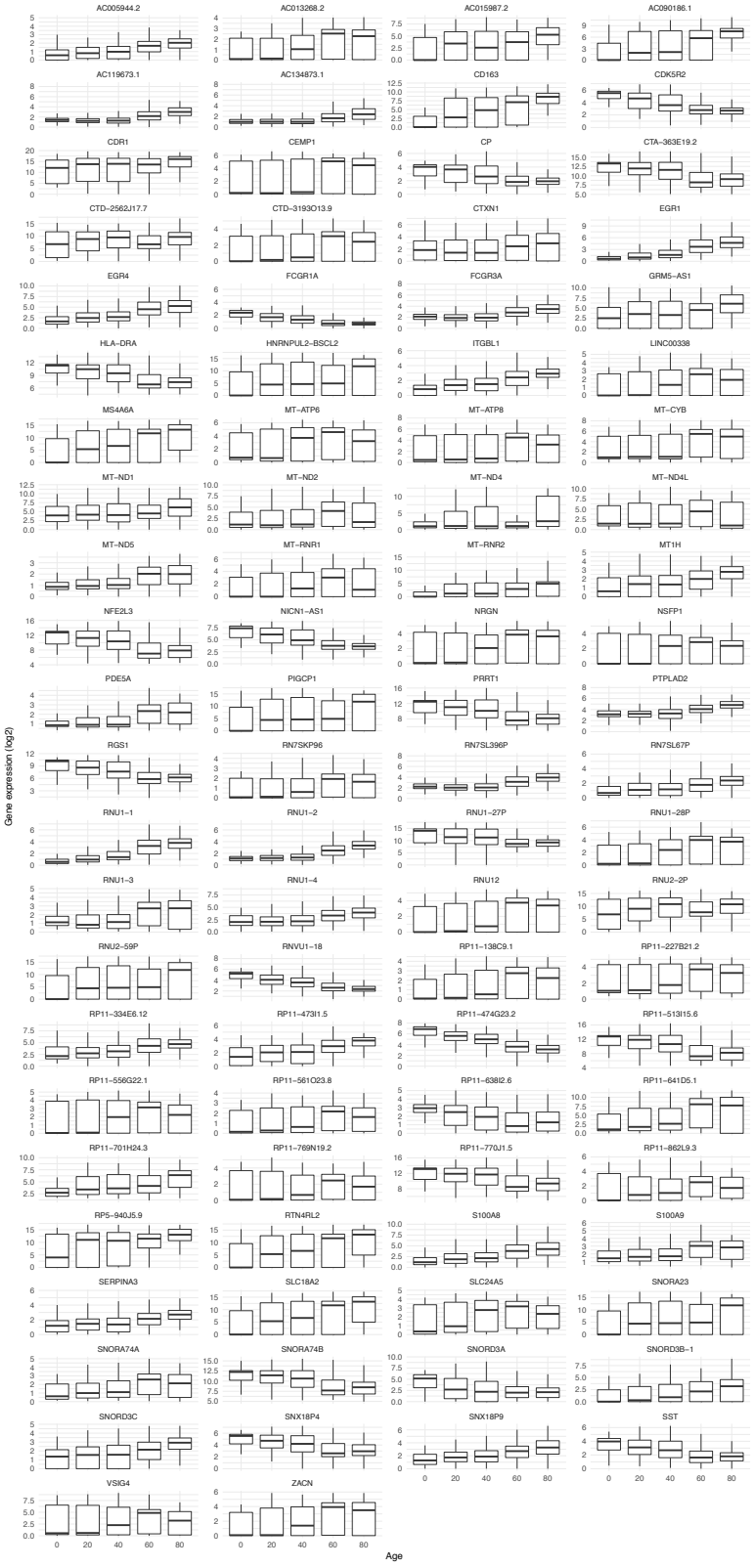


Fig. S20. Gene expression variation in the human brain across ages. The X axis shows five bins of age and the Y axis shows the log2(rpkm) for genes positively or negatively correlated with age. Each panel refers to a gene, where the identification was made by ENSEMBL ID. For details, see Section S2.4 and S2.5.

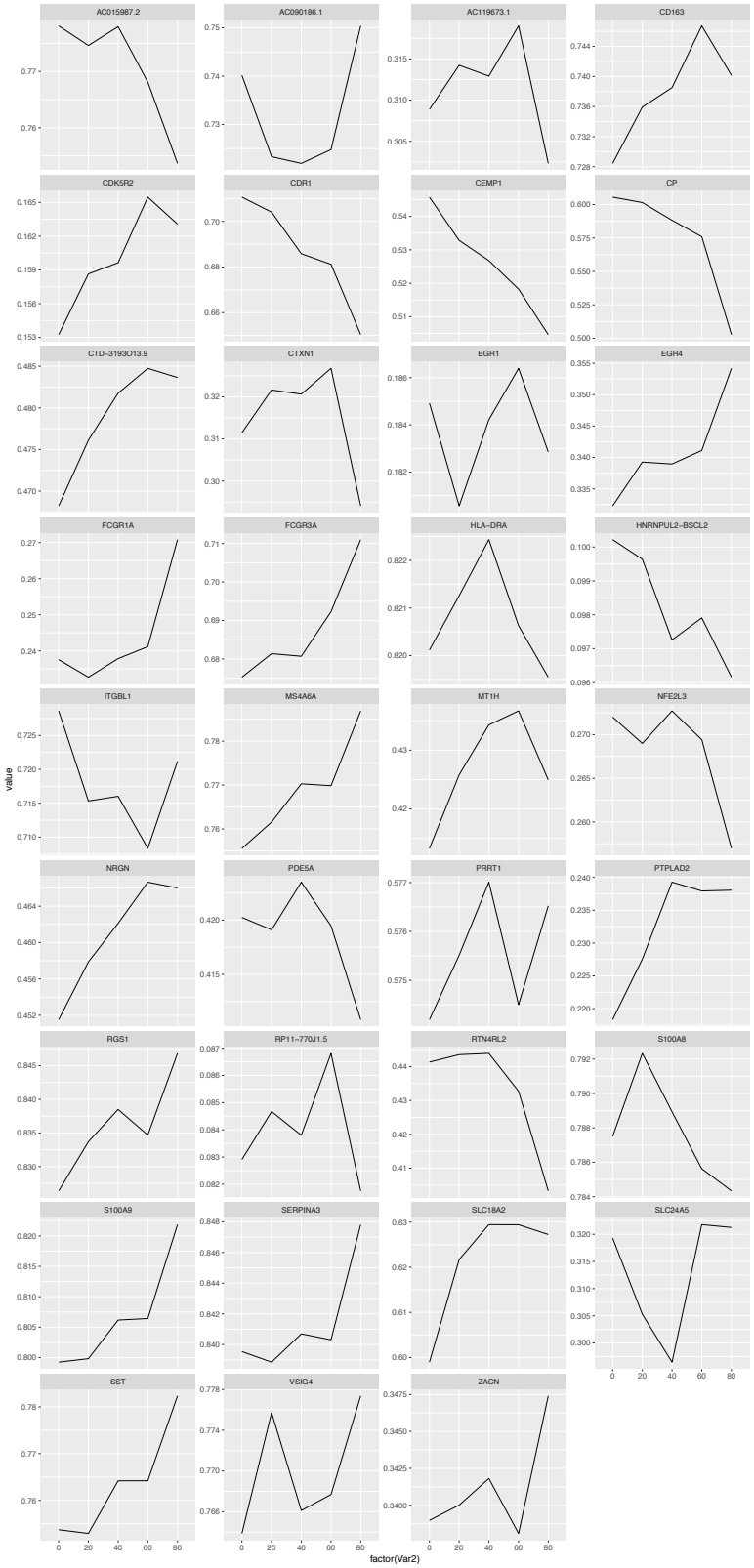


Fig. S21. Promoter and enhancer region methylation of genes correlated to aging. Genes with methylation data available were assessed for their methylation status. The X axis shows five bins of age and the Y axis shows the normalized distribution of methylation near the gene TSS. Each panel refers to a gene, where the identification was made by gene name. For details, see Section S2.7.

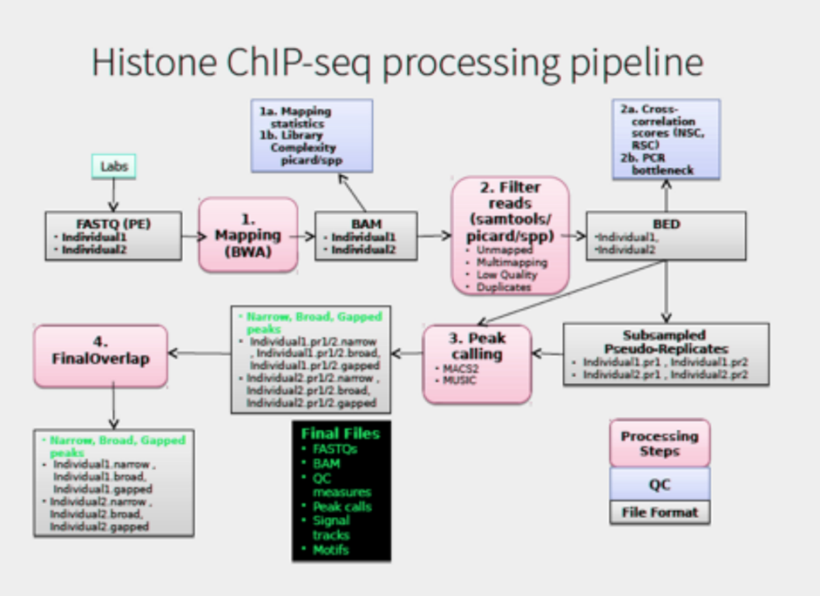


Fig. S22. PsychENCODE ChIP-seq processing pipeline. This pipeline flowchart was adapted and modified from the ENCODE ChIP-seq pipeline (https://goo.gl/KqHjKH). FASTQ files were aligned using BWA and the reads were filtered to get only unique mapped reads for peak calling using MACS2. Pseudo-replicates were generated before peak calling for each individual to find robust peaks. NSC, RSC, and PCR bottlenecks were generated for quality control. For details, see Section S3.1.

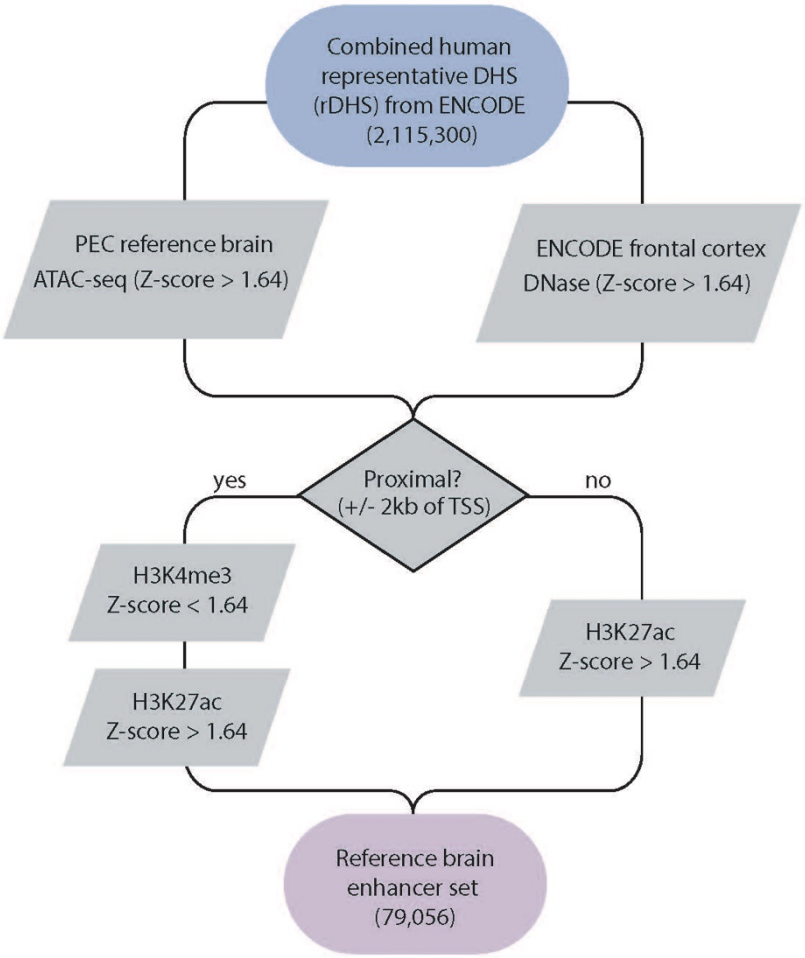


Fig. S23. Enhancer calling pipeline. Consistent with the ENCODE consortium annotation pipeline, we started from the combined representative DHS (rDHS) regions and filter them to retain those in open chromatin status (with ATAC-seq signal or DNase signal) in the frontal cortex. Enhancers are defined as distal open chromatin regions (>2kb away from any TSS) with H3K27ac signal, or proximal regions with H3K27ac but not H3K4me3 signal. Z-scores of the corresponding signals were calculated at each of the rDHS region. Z-score > 1.64 was used as the threshold to call signal significance, which corresponds to the 95th percentile of a one-tailed Z-test. For details, see Section S3.2.

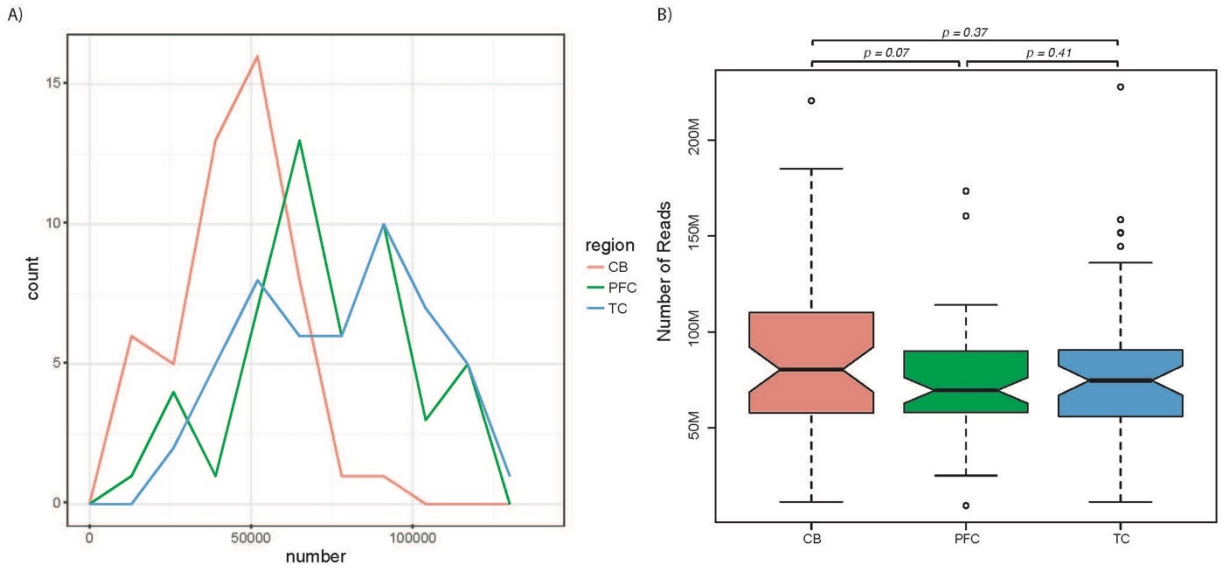


Fig. S24. Comparision of the number of H3K27ac peaks in three brain regions across the cohort. A) Histogram of the number of H3K27ac peaks in CB, PFC, and TC. B) Boxplot of the sequencing depth of the H3K27ac ChIP-seq experiments. For each distribution, the box shows the interquartile range (IQR) of the number of reads from a sample in the corresponding brain region. Whiskers show 1.5 times the IQR, and flier points show individual datasets that are outliers. CB: cerebellum, PFC: prefrontal cortex; TC: temporal cortex**.** For details, see Section S3.3.

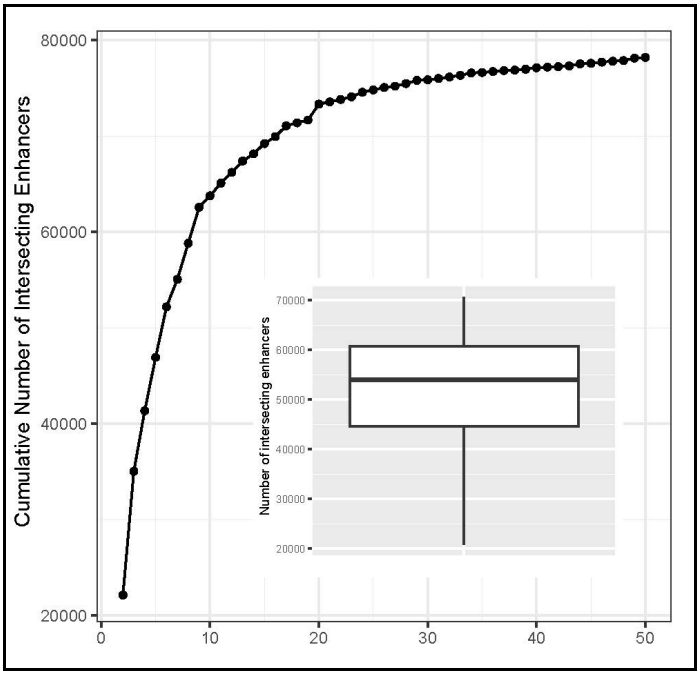


Fig. S25. Active reference brain enhancers in the population. The dotted line shows the cumulative number of identified reference sample enhancers in the cohort, which saturates at the 20th individual from the sorted cohort. The boxplot shows the number of identified reference enhancers found active in each individual, with the lower and the upper boundaries of the box showing the first and the third quantiles. For details, see Section S3.3.

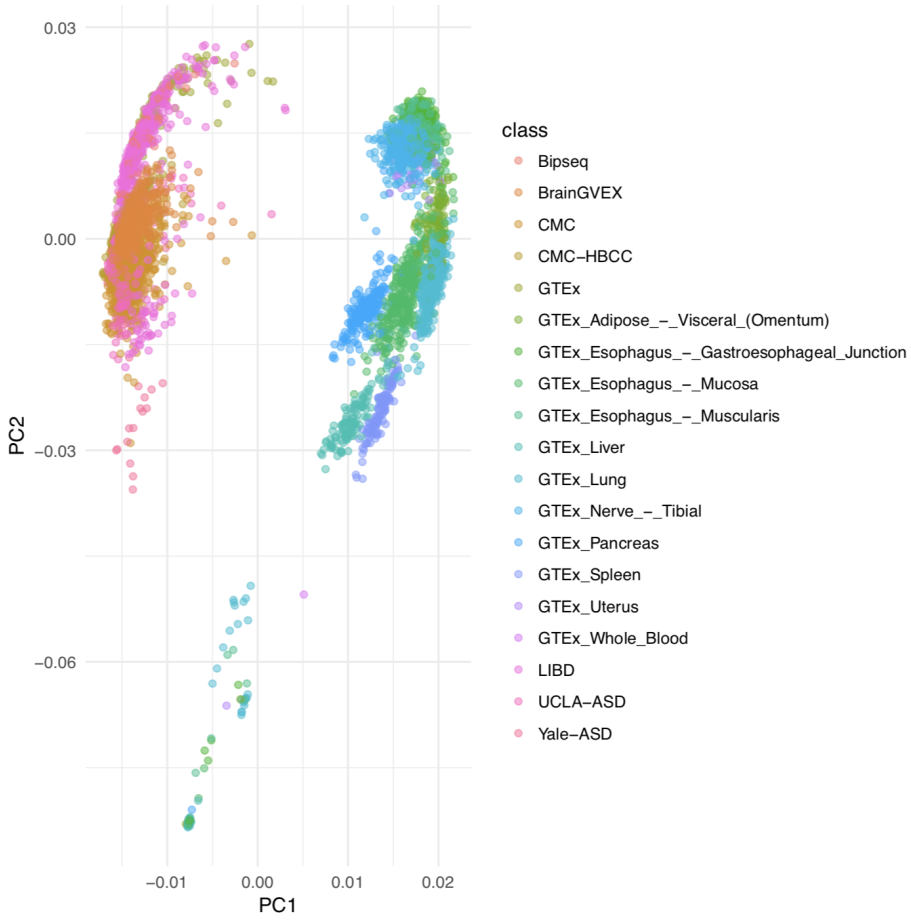


Fig. S26. Analog version of Figure 3E with other tissues colored for comparison. For details, see Section S4.1.

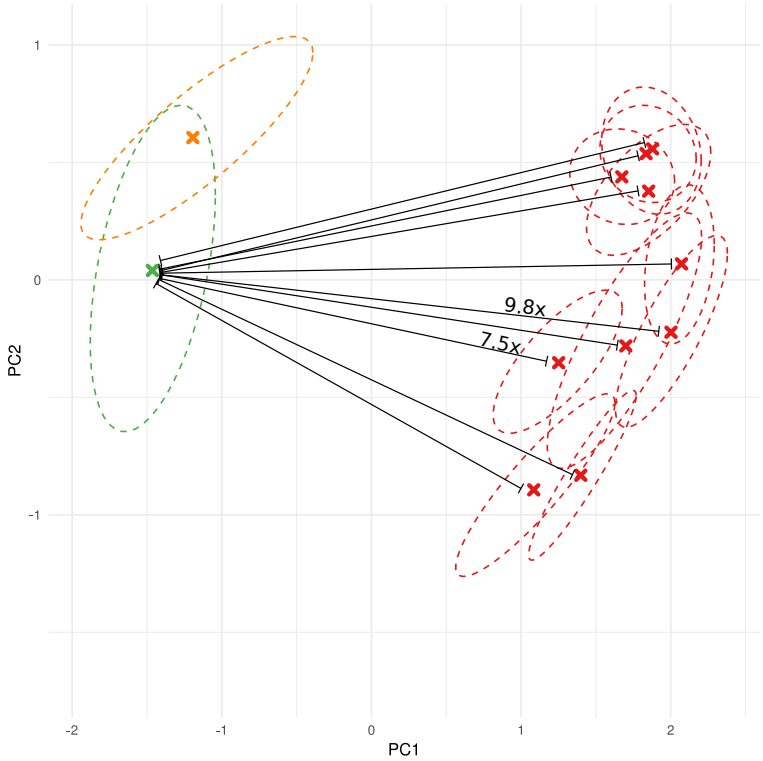


Fig. S27. Distances from the brain to other tissues centroids. Transcriptome RCA plot as in Fig. 3. Brain samples' distributions are displayed in green and orange. Other tissues are shown in red. Euclidean distance was calculated between all centroids (Table S5) and normalized by the median brain distance (Table S6). For details, see Section S4.1.

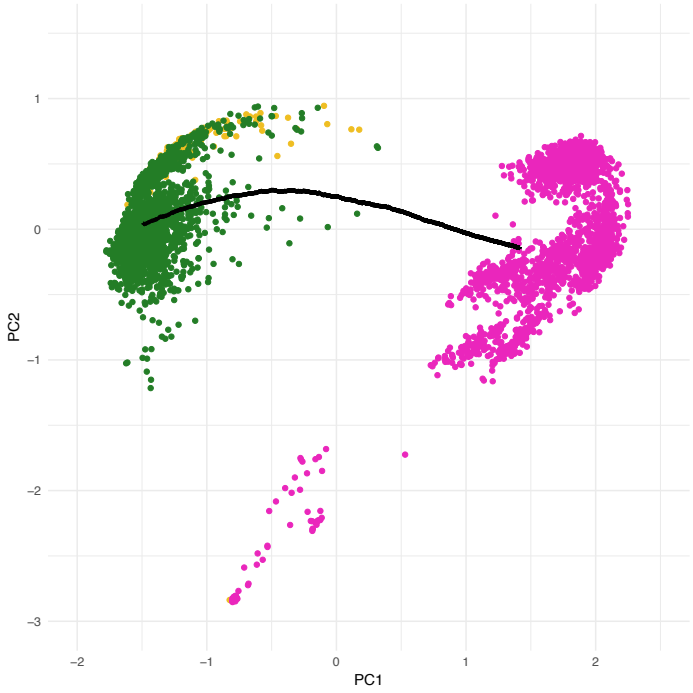


Fig. S28. Assessment of the most impactful genes in the PC1 dimension. All analyzed RNA-seq samples are displayed. Green and yellow samples are brain samples and pink samples were extracted from other tissues. Dark line represents hypothetical samples with gene expression changes. For details, see Section S4.1.

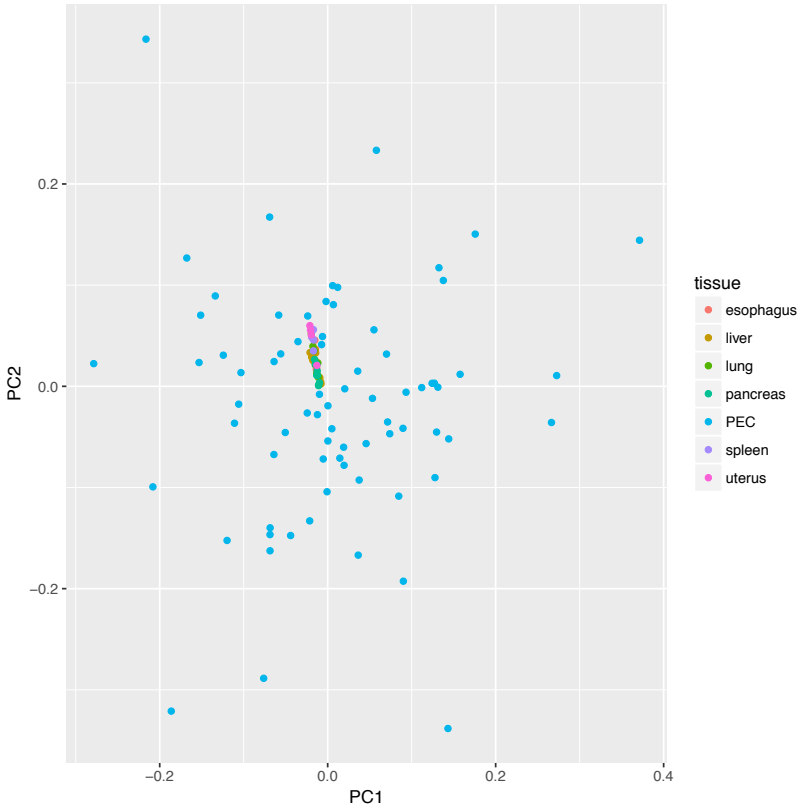


Fig. S29. PCA plot for regulatory data. H3K27Ac signals were used to calculate the PCA after batch correction. Brain samples are scattered in both PC1 and PC2, whereas the roadmap samples are clustered together. For details, see Section S4.1.

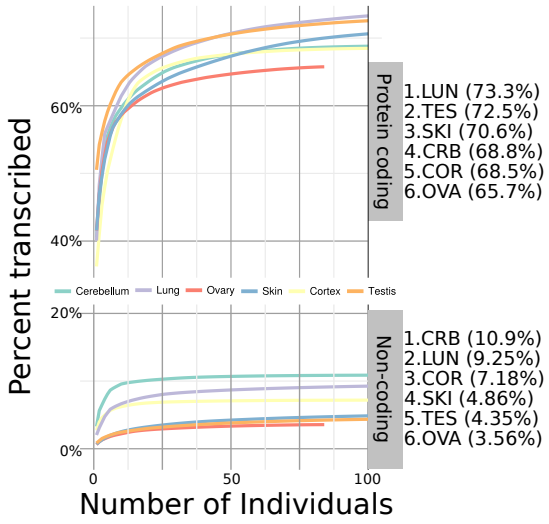


Fig. S30. Cumulative distribution of transcribed regions in the human brain and other tissues. The Y axis shows the cumulative transcribed proportion of annotated and unannotated regions (coding or non-coding). The X axis shows the number of transcriptomes (or individuals) analyzed. Labels on the right-hand side of the figure display the maximum cumulative proportion found. For details, see Section S4.2.

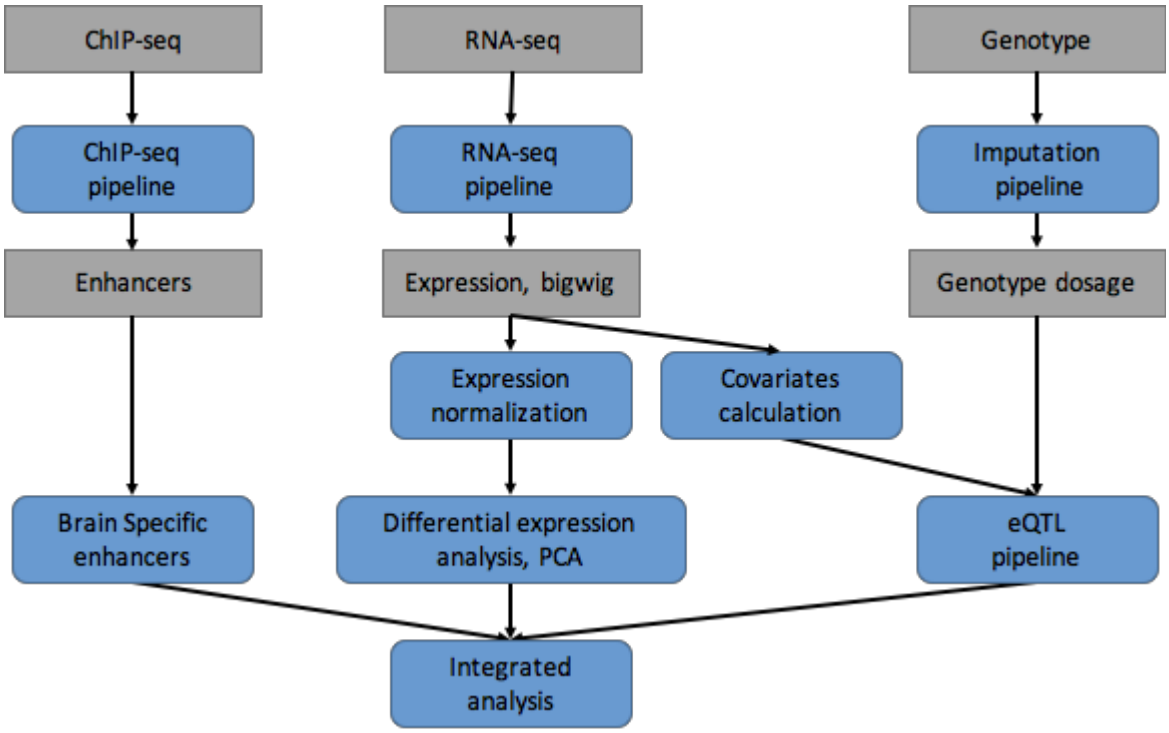


Fig. S31. Integrated analysis pipeline of PsychENCODE. We used the standard pipelines from ENCODE, GTEx and other large consortia to uniformly process the raw sequencing data from PsychENCODE, including RNA-seq, ChIP-seq, and genotype data, as well as to identify functional genomic elements such as brain enhancers, expressed genes, and eQTLs. We also processed other data types, such as Hi-C and single cell data. Details on data processing are provided in the sections below. As shown by this flowchart, we then performed integrative modeling and analysis of functional genomic elements in the adult brain.

We closely followed the GTEx pipeline for our eQTLs. We did this to ensure maximal compatibility between our results and previously published results, and also to optimally enable comparisons between our results and those published previously. In a manner parallel to the approach used by GTEx, we retain those genes for which at least 10 individuals have an FPKM greater than 0.1. We used the QTLtools software package for eQTL identification. Following the normalization scheme used by GTEx, the gene expression matrix was first normalized using quantile normalization, followed by inverse quantile normalization to map to a standard normal distribution (and to remove outliers). 50 PEER factors, genotype PCs, gender, and respective study were used as covariates in our calculations to identify cis-eQTLs. (Given our much larger sample size, we used considerably more PEER factors than GTEx.) For cis-eQTLs, we calculated the associations between gene expression and variants within a 1Mb window of each gene's TSS (both upstream and downstream). 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 tested for potential QTLs).

We performed multiple testing corrections on nominal P-values by limiting FDR values to less than 0.05. We identified 2,542,908 significant cis-eQTLs and 15,626 protein-coding eGenes. 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 (r2) values to exceed 0.5. Enforcing this resulted in 373,686 eQTLs. The number of eQTLs and eGenes is similar to those found by the recent CommonMind study *(19)*, and we further note that our sample size is almost twice as large as that used by the CommonMind study.

These 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 larger sample size. This larger sample size also enabled us to generate more conservative eQTL lists by using different parameters from those used by GTEx. We did multiple calculations by varying the parameters, such as sample expression thresholds and forms of multiple testing corrections.

For instance, we used the Bonferroni method for multiple testing correction and generated cis-eQTLs. As expected, this resulted in fewer eQTLs: 674,815 cis-eQTLs and 5,489 protein-coding eGenes. We also used more conservative gene expression filtering thresholds by retaining genes with an FPKM>0.1 in more than 150 individuals. Separately, we also carried out another calculation including only genes having an FPKM>1 in more than 20% of our samples.

This change does not affect the results very substantially, giving 2,566,206 eQTLs with 15,139 protein-coding eGenes (FPKM>0.1 in more than 150 individuals) and 2,120,751 eQTLs with 12,822 protein-coding eGenes (FPKM>1 in more than 20% of samples). This supports our observation that the large numbers of eQTLs are fairly robust to parameter choice, and that retaining the original GTEx parameters provides sensible results. All of the eQTL lists are available on the website as files DER-08 a-d.

Finally, with respect to all of the network calculations discussed below, we used only our first set of eQTLs (i.e., the eQTLs identified using the thresholds employed by GTEx).

Using a similar pipeline to that used for our eQTLs, we also identified splicing activity-related QTLs based on isoform percentages (isoQTLs) and transcript TPM values (tQTLs). We calculated isoQTLs and tQTLs in 2 different ways: one in which all the genes included in eQTL calculation were used, and another in which we used a filtered gene list. In the filtered list, we retained only the 3,819 genes at an FPKM>5 in all samples. For 1,147 individuals, we identified 2,628,260 isoQTLs and 2,528,818 tQTLs based on all genes at an FDR<0.001. We also identified 604,286 isoQTLs and 726,350 tQTLs on the filtered gene list with an FDR<0.001. All of the isoQTL and tQTL lists are available on the website as files DER-10 a-d.

In addition to the similarity with GTEx and CMC brain eQTLs, we evaluated the similarity between our eQTLs and GTEx eQTLs of other tissues using π1 statistics and SNP-eGene overlap rate. The SNP-eGene overlap rates were calculated based on percentage of shared SNPs associated with the same eGene using LD independent eQTL SNPs. The π_1 values of liver, lung, testis, and blood eQTLs are 0.76, 0.88, 0.9, and 0.88, respectively, which are smaller than the π_1 value 0.93 of GTEx brain eQTLs. The SNP-eGene overlap rates are also the highest in brain DLPFC among all the tissues tested (Figure 4B). For details, see Section S5.

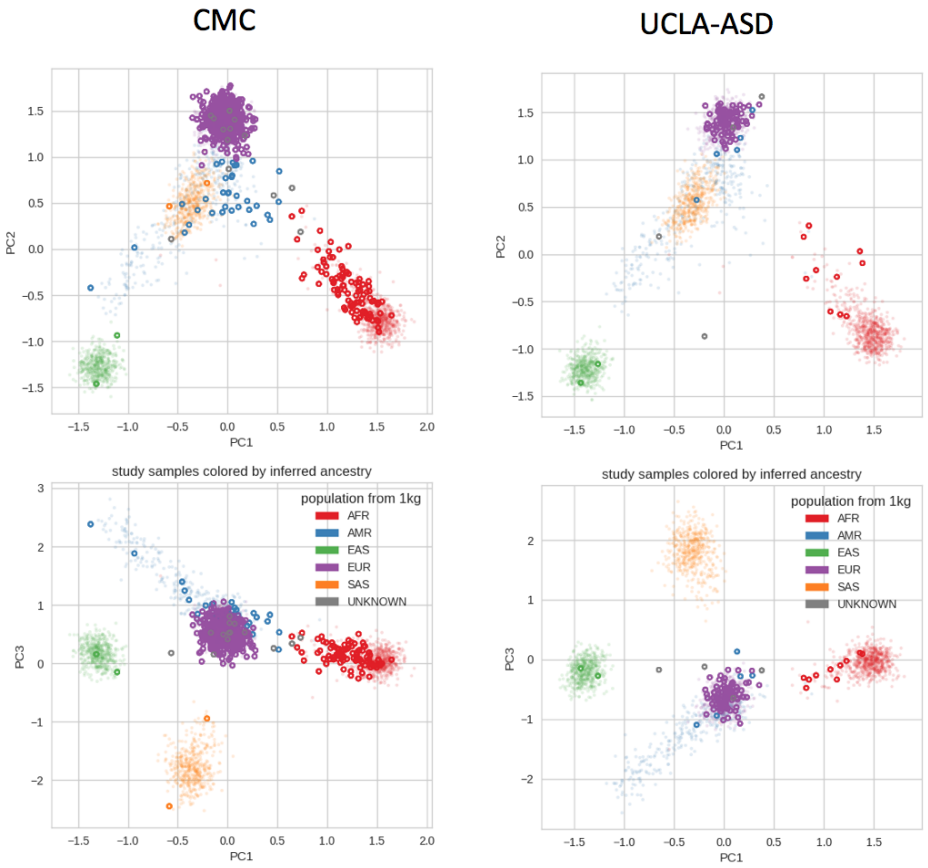


Fig. S32. Genotype PCs showing the population structure in CMC and UCLA-ASD studies. The first three genotype PCs could capture most of the population structures. The top panels show genotype PC1 vs. PC2. The bottom panels show genotype PC1 vs. PC3. A majority of the individuals in these two studies were from EUR populations. For details, see Section S5.1.

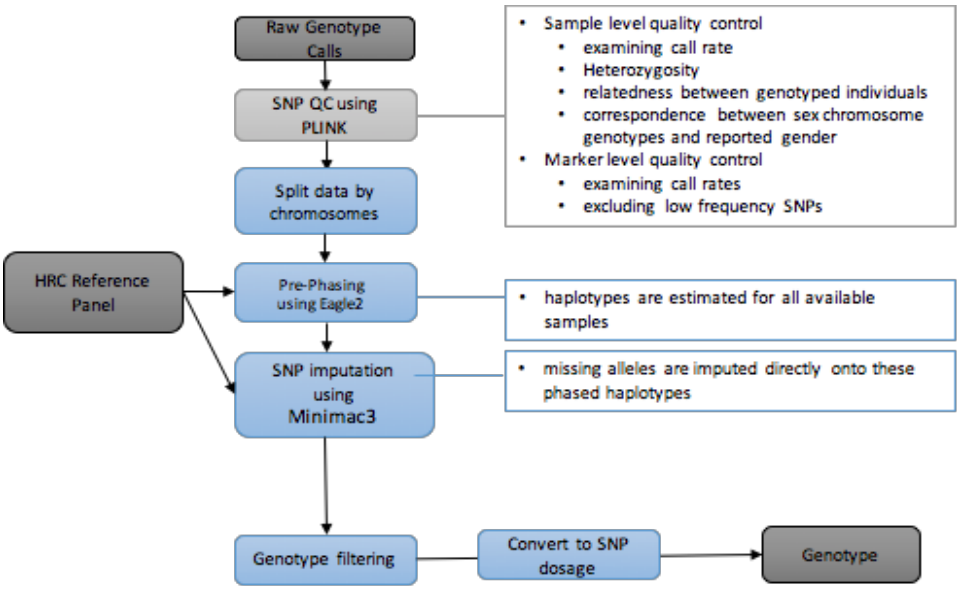


Fig. S33. PsychENCODE genotype data processing pipeline. The raw genotype data were called and converted to PLINK files. We ran an initial quality sample level and marker level using PLINK. The quality-controlled genotype data were then prepared by prephasing using Eagle2. The prephased data were imputed using Minimac3 and HRC. After imputation, we filtered genotype using R2>0.3 to get high-quality imputation data.

Genotyping was done on several different genotyping platforms listed in Supplemental Table S1 and Section S9. Initial QC was performed using PLINK (*116*) 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).

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 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 non-typed 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.

The second module consisted of genotype PCA using peddy (*112*) to identify ancestry structure (Fig. S32). 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 (*117*). 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. For details, see Section S5.1.

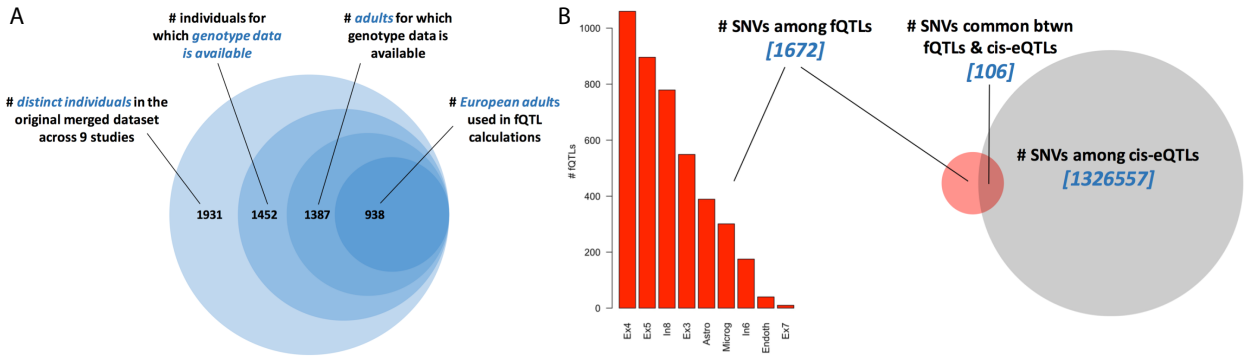


Fig. S34. Datasets, counts, and cis-eQTL overlaps associated with fQTLs. A. In calculating fQTLs, we restricted our analyses to 938 European adult samples for which genotype data was available. **B.** The histogram on the left represents the counts for the number of fQTLs across ten different cell types. These fQTLs encompass 1,672 distinct SNVs, of which 106 (6.3%) also appear among the cis-eQLTs. For details, see Section S5.2-4.

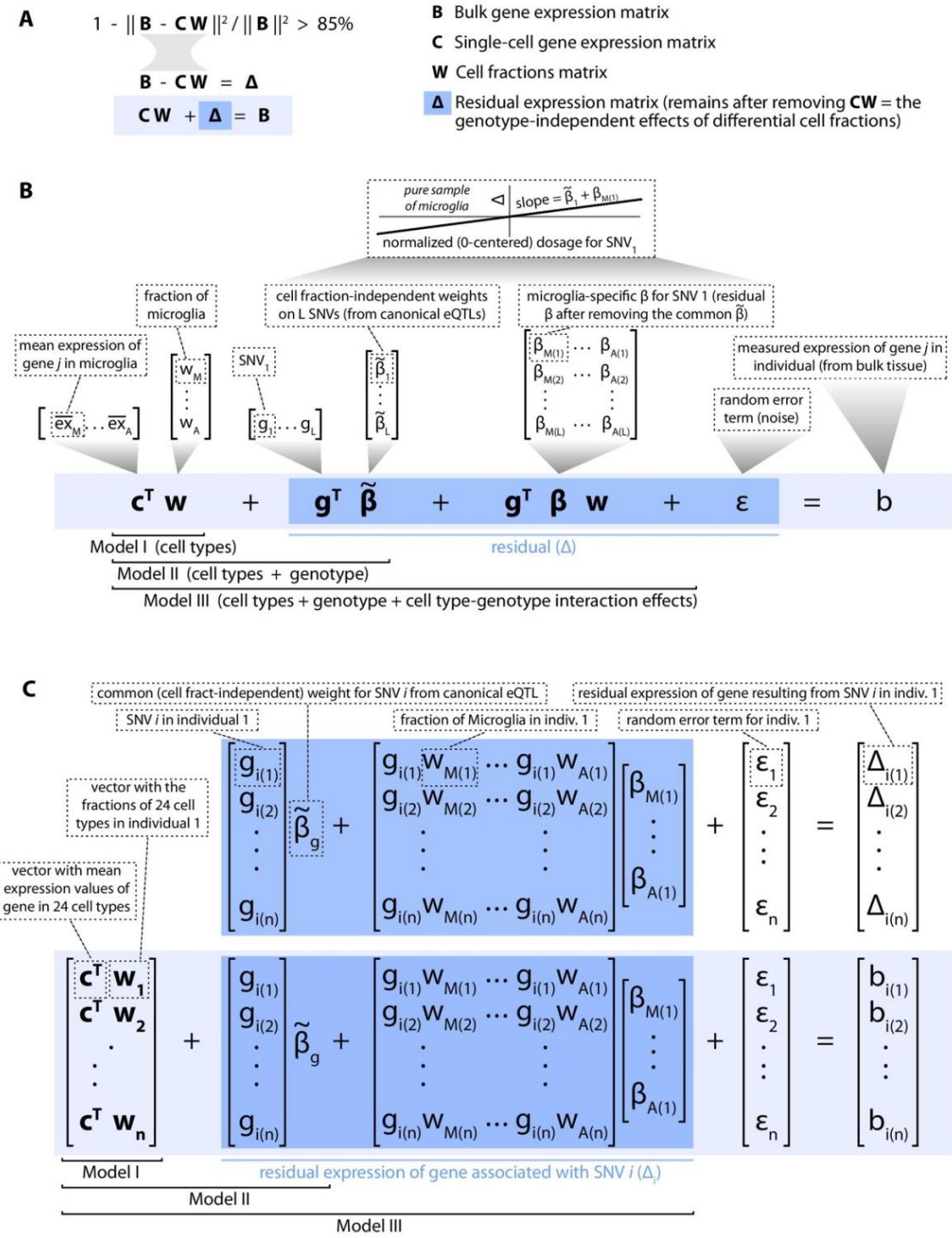


Fig. S35. Formalistic overview of residual expression and associated terms.
A: As reported in Fig. 2 of the main text, the estimated cell fractions explain more than 85% of bulk tissue expression variation: 1- ||**B** – **CW**||^2 / ||**B**||^2 > 0.85, where **B** designates the bulk tissue gene expression matrix, **C** designates the single-cell (i.e., cell type-specific) gene expression matrix, and **W** designates the matrix of estimated cell fractions for individuals. The matrix **Δ** can be thought of as the bulk tissue gene expression that cannot be accounted for by cell type-specific metrics alone (as represented by **CW**, a combination of the mean gene expression values for each cell type and the fractions of different cell types). **Δ** thus represents the contributions of *genetic components* to bulk tissue

gene expression after controlling for the cell type-specific contributions: **Δ** = **B** - **CW**. It thus evaluates expression as a function of both the genetic components and their interaction effects with cell type-specific metrics.
B: Modeling bulk tissue expression of ***one gene*** in ***one individual*** with ***many (L) SNVs***. Shown are the details of a three-layered (or three-model) nested framework for understanding each term in the expression **CW** + **Δ** = **B**. Terms within each component are detailed in dashed boxes. The box at the very top illustrates the meanings of $\tilde{\beta}_1$ and $\beta_{M(1)}$: in calculating a QTL, the genotype dosage for a particular SNV (SNV_1 in this case) is normalized across all individuals to have a mean of 0. $\tilde{\beta}_1$ represents the common effect size associated with the QTL that corresponds to SNV_1 (i.e., it represents the common *slope* in the linear regression that is used to identify this QTL, without considering cell-specific expression or cell type fractions). $\tilde{\beta}_1$ thus represents a type of universal, generic, averaged slope value associated with SNV_1. $\beta_{M(1)}$ represents the deviation from the common slope $\tilde{\beta}_1$ that would result from measuring gene expression in a pure sample of microglia (M). Thus, in a pure sample of microglia, the slope that would ultimately be evaluated in a QTL for a pure sample of microglia is $\tilde{\beta}_1 + \beta_{M(1)}$. Note that, because the gene expression values in this linear regression include the genotype component, the measured expression is actually the residual expression value for this gene in a pure sample of microglia.
C: Modeling the contribution of ***one SNV*** to the bulk tissue expression of ***one gene*** in ***many (n) individuals***. As in Panel A, shown are the details of the three-layered (or three-model) nested framework, with dashed boxes providing details for each component. At the top right, the residual expression resulting from the SNV in a particular individual can be thought of as the contribution that this SNV makes to the expression, after factoring out the cell type-specific contributions alone (i.e., **CW**).

For details, see Section S5.4.

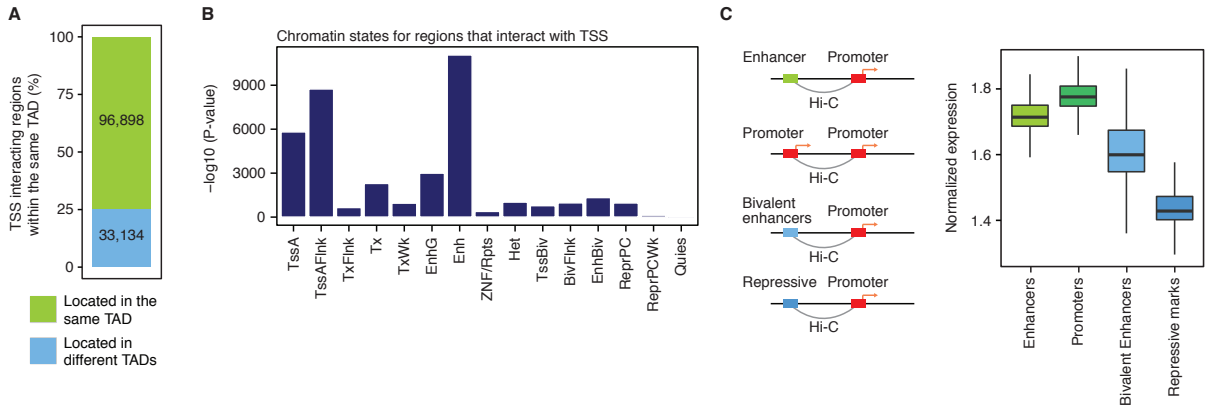


Fig. S36. Regulatory relationships in the adult cortex. A. The majority of promoter-based interactions reside within the same topologically associating domains (TADs). **B.** Regions that interact with transcription start sites (TSS) are enriched with other TSS and enhancers. **C.** Genes that interact with enhancers or promoters are more highly expressed than genes that interact with bivalent enhancers or repressive marks. For details, see Section S6.1-3.

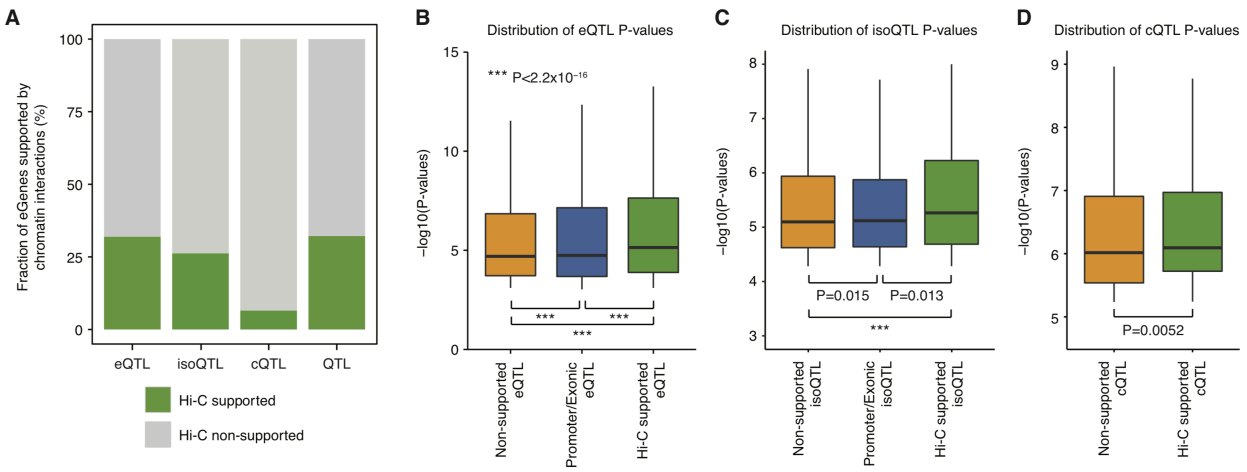


Fig. S37. Chromatin interactions mediate cis- and trans-regulatory relationships. A. A proportion of QTL-associated genes (eQTLs), isoforms (isoQTLs) and chromatin marks (cQTLs) that have Hi-C evidence. **B.** eQTLs supported by Hi-C evidence show stronger associations not only to eQTLs without genomic 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). For details, see Section S6.5.

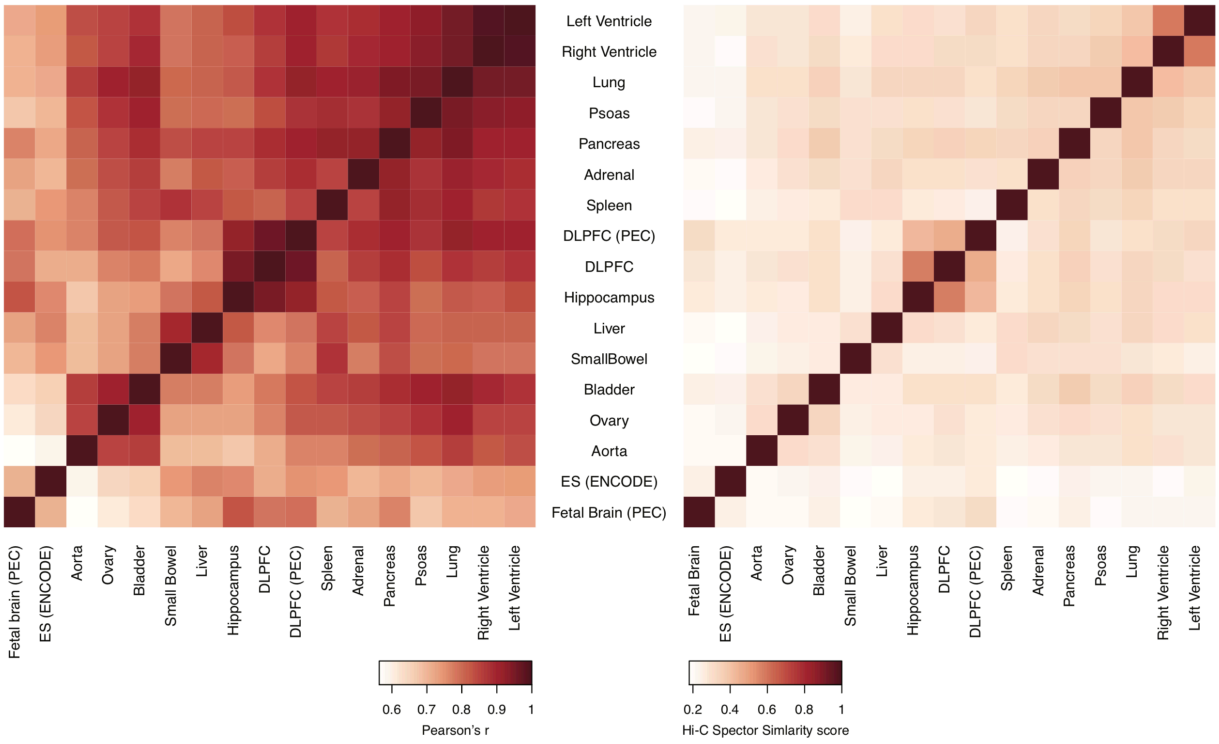
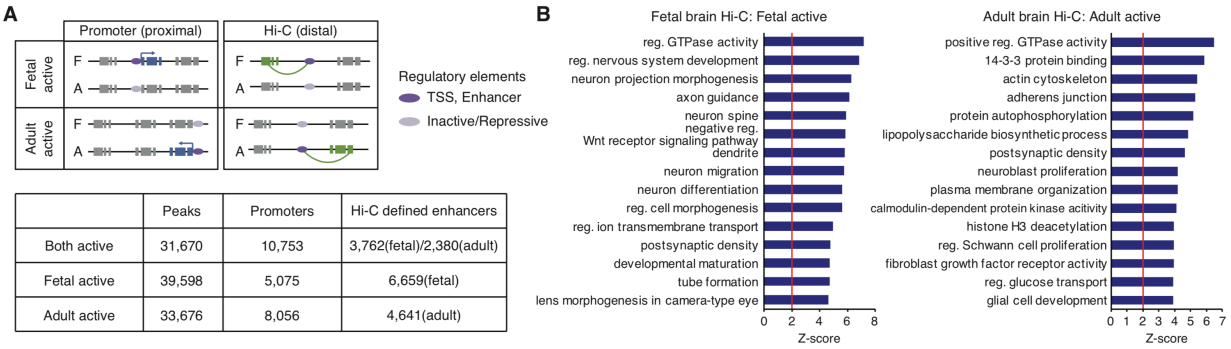


Fig. S38. Cross-tissue comparison of chromatin architecture. Pearson's correlation coefficients (left) and HiC-Spector similarity scores (right) calculated from contact matrices at 1Mb resolution show chromatin architecture from similar tissue types (e.g., DLPFC from PEC and Roadmap and hippocampus from Roadmap, right ventricle and left ventricle) is more highly correlated than non-related tissue types. Strikingly, fetal brain shows distinct chromatin architecture to adult brain, indicating extensive rewiring of chromatin structures during brain development. For details, see Section S6.4.



	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. S39. 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 two categories: 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), whereas 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 table. **B.** 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. For details, see Section S6.5.

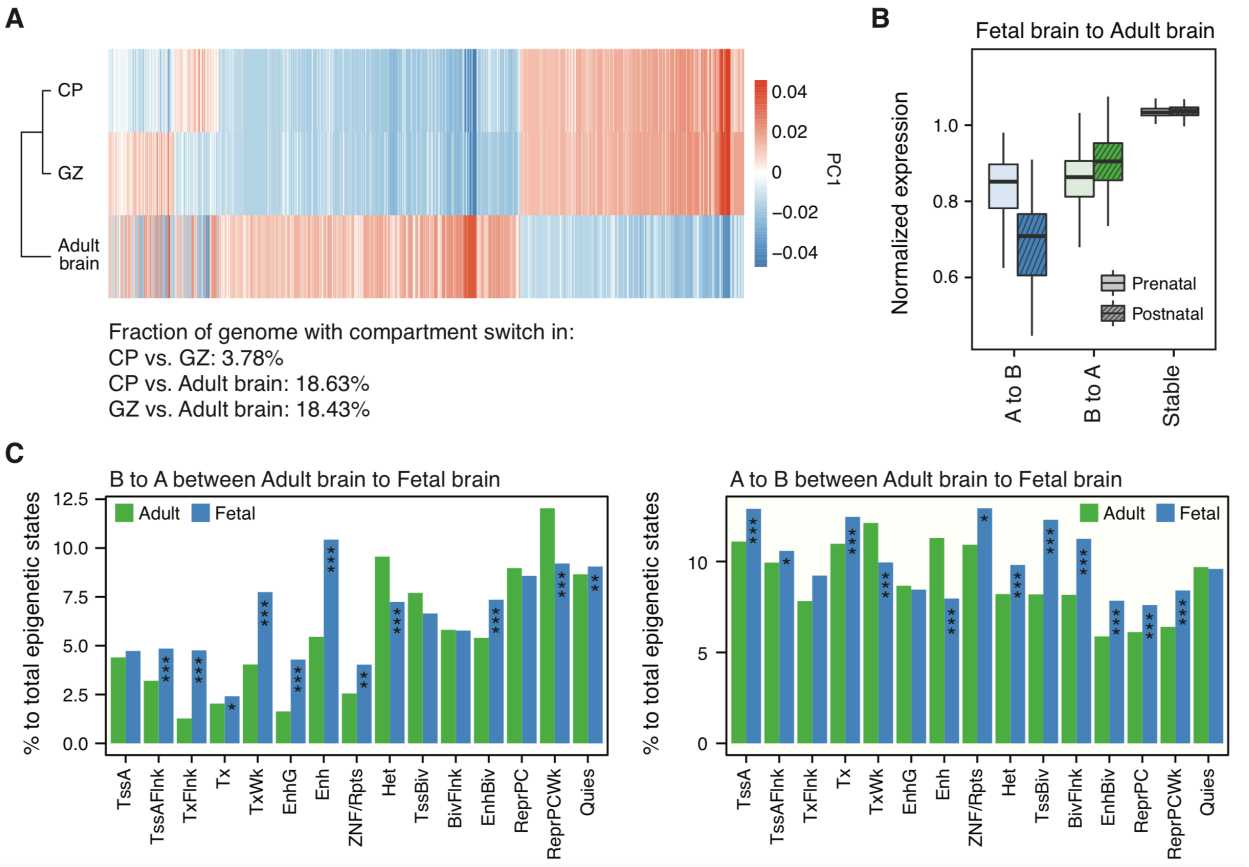


Fig. S40. Compartment switching across brain development is associated with expression and epigenetic changes. A. Heat map of the first principal component (PC1) values for regions that undergo compartment switching between fetal brain (CP, cortical plates which represent cortical layers with post-mitotic neurons and GZ, germinal zones which represent cortical layers with neural progenitors) 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. For details, see Section S6.5.

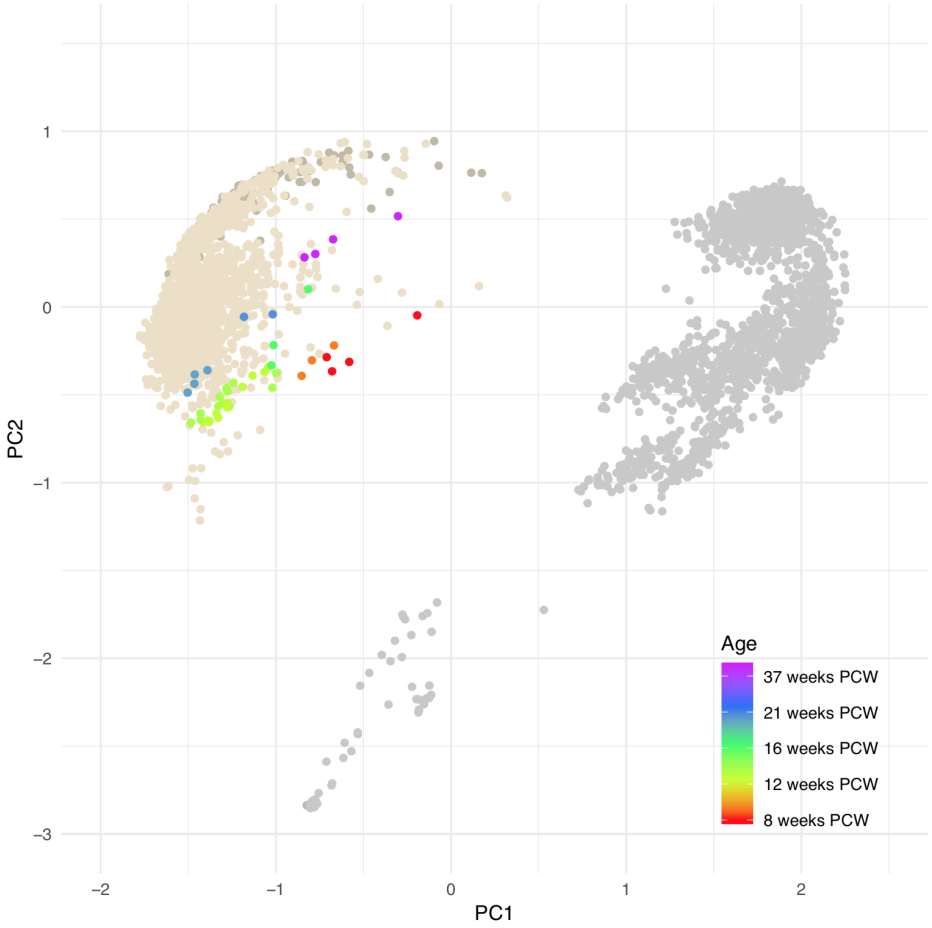


Fig. S41. Assessment of fetal samples in RCA space. All analyzed RNA-seq samples are displayed. Beige and gray samples are other tissues, respectively. Fetal samples are colored in respect of their age. We can observe a trajectory from young (red – 8PCW) to older (purple – 37PCW). For details, see Section S4.1 and S6.5.

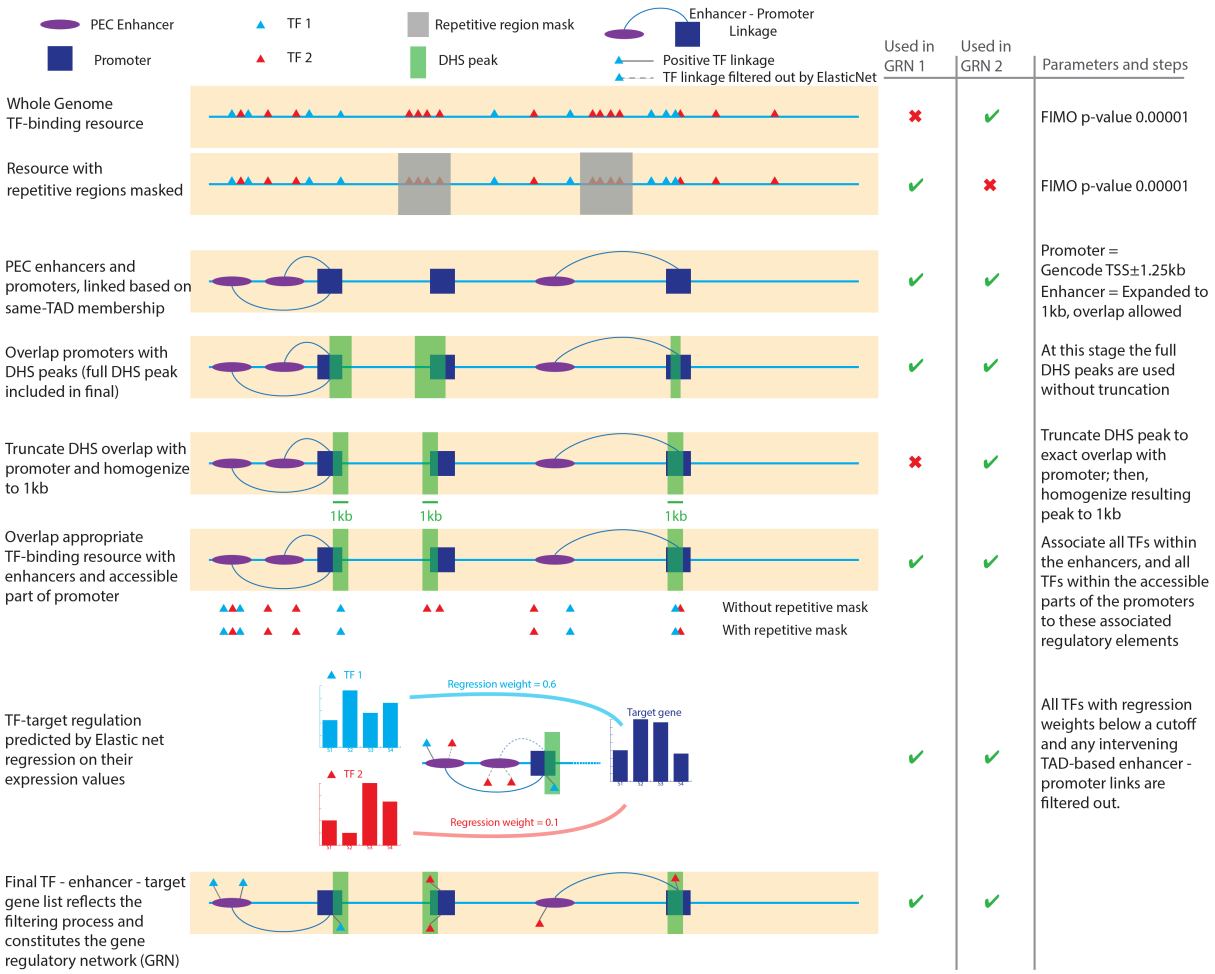


Fig. S42. Schematic of procedure for constructing gene regulatory networks. The analysis begins with the construction of two different TF-binding site (TFBS) maps of the whole genome, one with repetitive regions masked (for GRN1) and the other without such masking (for GRN2). PEC enhancers and promoters are tentatively linked by their membership within the same TAD region determined from the HiC matrix of the adult DLPFC. Subsequently, promoters are overlapped with DNase hypersensitivity site (DHS) peaks, which are either used as is (for GRN1), or truncated and homogenized to 1kb (for GRN2). Subsequently, the TFs from the corresponding TFBS map are linked to the promoters and enhancers, which are in turn linked as described above. The final step involves carrying out an ElasticNet regression using the TF expression levels to fit the target gene expression level. Those promoter- and enhancer-bound TFs whose regression weights are below a designated cutoff are removed from the TF-enhancer-promoter-TF linkage, as are any enhancers that are entirely bound by TFs deemed to have insignificant regression weights for the given target gene. The remaining linkages constitute the respective GRNs. For details, see Section S6.5.

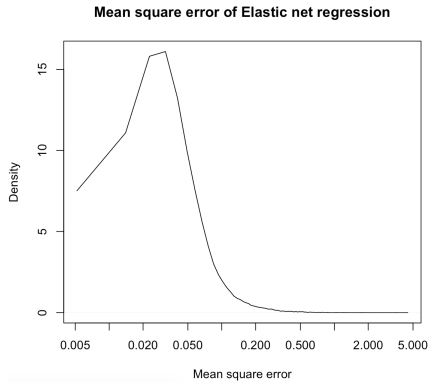


Fig. S43. Mean square error distribution of Elastic net regression predicting target gene expression from TF expression. The x-axis is the mean square error range across protein-coding target genes, using GRN1. The y-axis is the density of target genes. An associated data file with the mean square error values for each gene with an Elastic Net prediction is available on the website (resource.psychencode.org). For details, see Section S6.5.

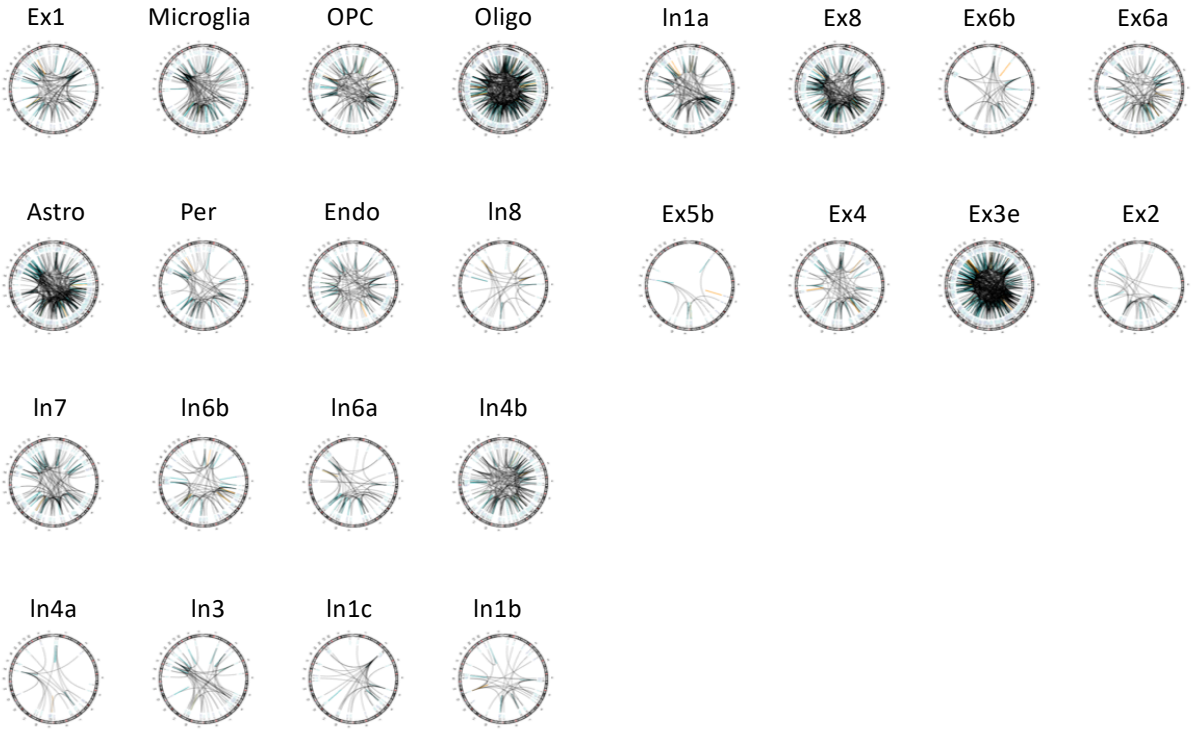


Fig. S44. The gene regulatory linkages for single cell marker genes across various cell types. The circos plots show the linkages from the full regulatory network targeting the cell-type-specific biomarker genes for various cell types (*64*). For details, see Section S6.5.

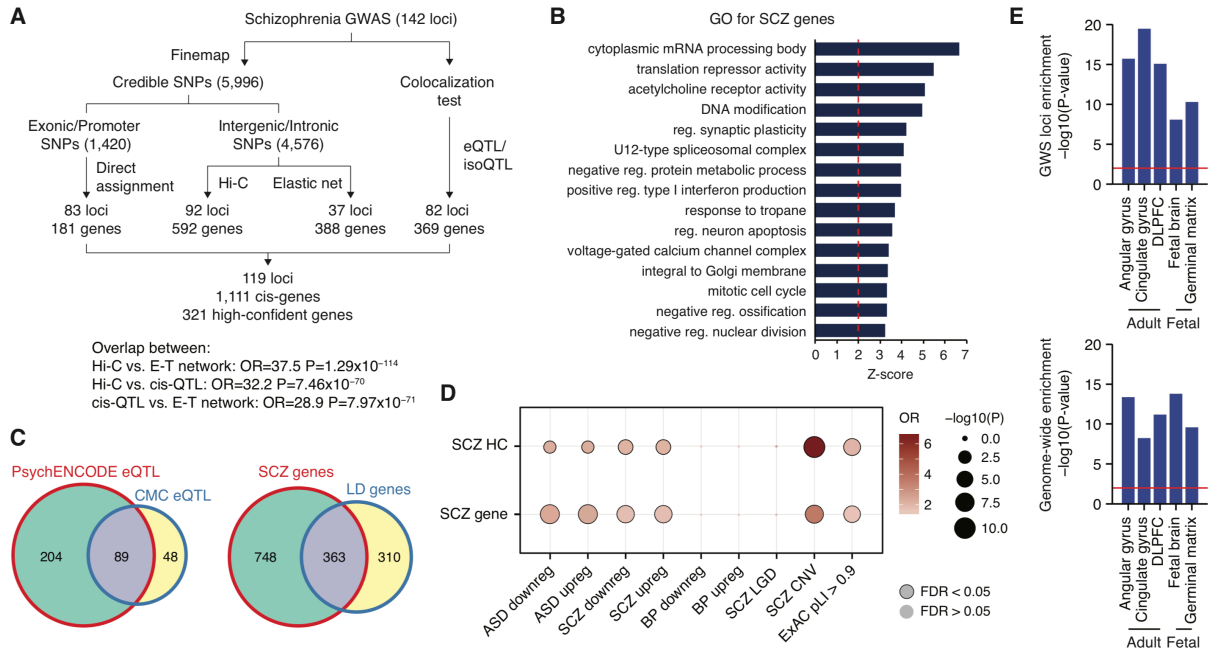


Fig. S45. Identification of schizophrenia risk genes. A. A schematic depicting how SCZ GWAS loci were assigned to putative genes. **B.** Gene ontology enrichment for SCZ-genes demonstrates that cholinergic receptors, synaptic genes, calcium channels, immune response-related genes, translational regulators, 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 (*99*). 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.** SCZ genome-wide significant (GWS) loci are more strongly enriched in regulatory elements in adult cortices than those in fetal brain (top). This is different from the genome-wide heritability enrichment (bottom), where genome-wide heritability enrichment is stronger in regulatory elements in fetal brain than in adult DLPFC. This result suggests that genome-wide significant SNPs are active in adult brain, while many sub-thresholded SNPs are active in fetal brain. Based on this result, we used regulatory relationships in adult brain to annotate SCZ GWS loci. For details, see Section S7.

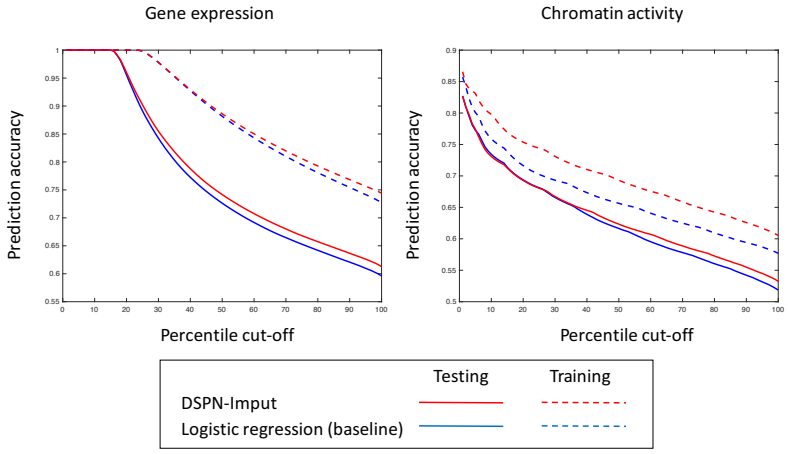


Fig. S46. Accuracy of intermediate phenotype imputation using DSPN-impute model. Figure compares prediction accuracy for gene expression and chromatin activity using the DSPN-impute model (with GRN structure included) vs. prediction with a logistic model (independent prediction). Performance on training and testing partitions is shown. For details, see Section S8.3.

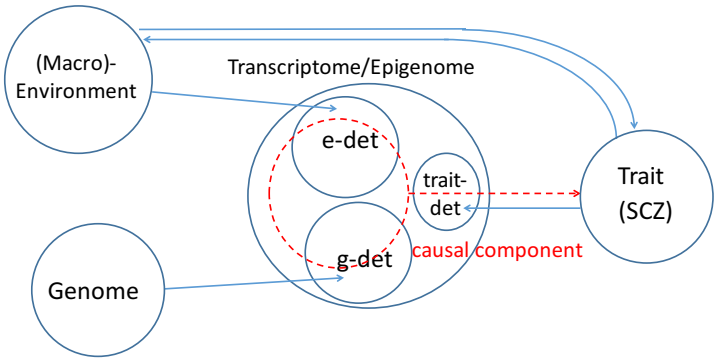


Fig. S47. Potential causal relationships between genome, transcriptome/epigenome, macro-environment and high-level traits. A schematic of a possible decomposition of variation in the indicated variables. Large circles represent total entropy of each variable, and smaller circles (e-det, g-det, trait-det) 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. For details, see Section S8.2.

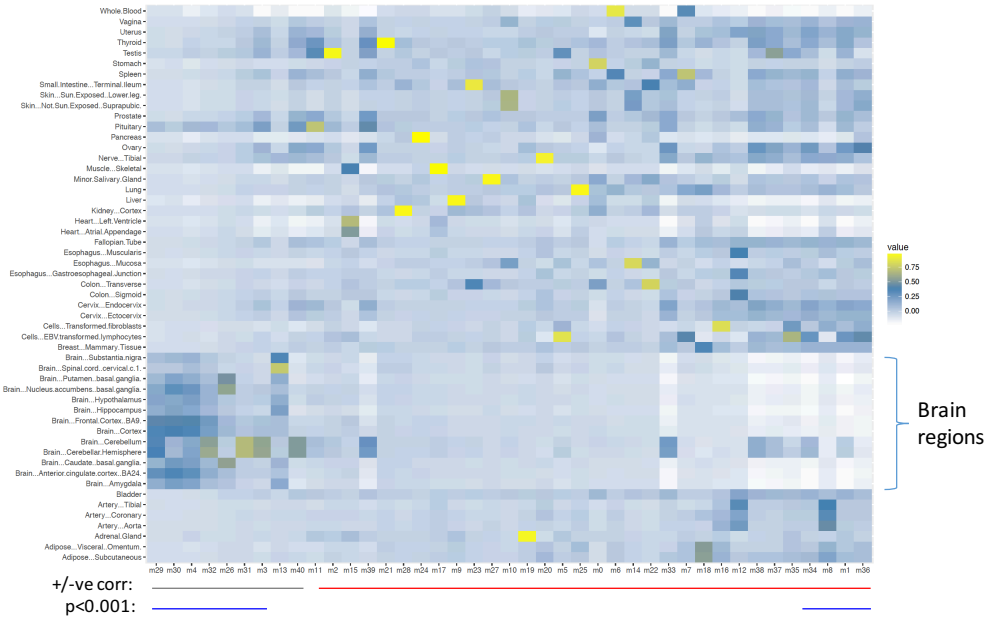


Fig. S48. Brain-specific co-expression modules and submodules. Module eigengenes are plotted as columns, which are ordered by the degree to which their expression is specific to the brain (see Section S2.6). The modules/sub-modules shown are a subset of the full 5024 modules, calculated using WGCNA on all GTEx expression data. 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.

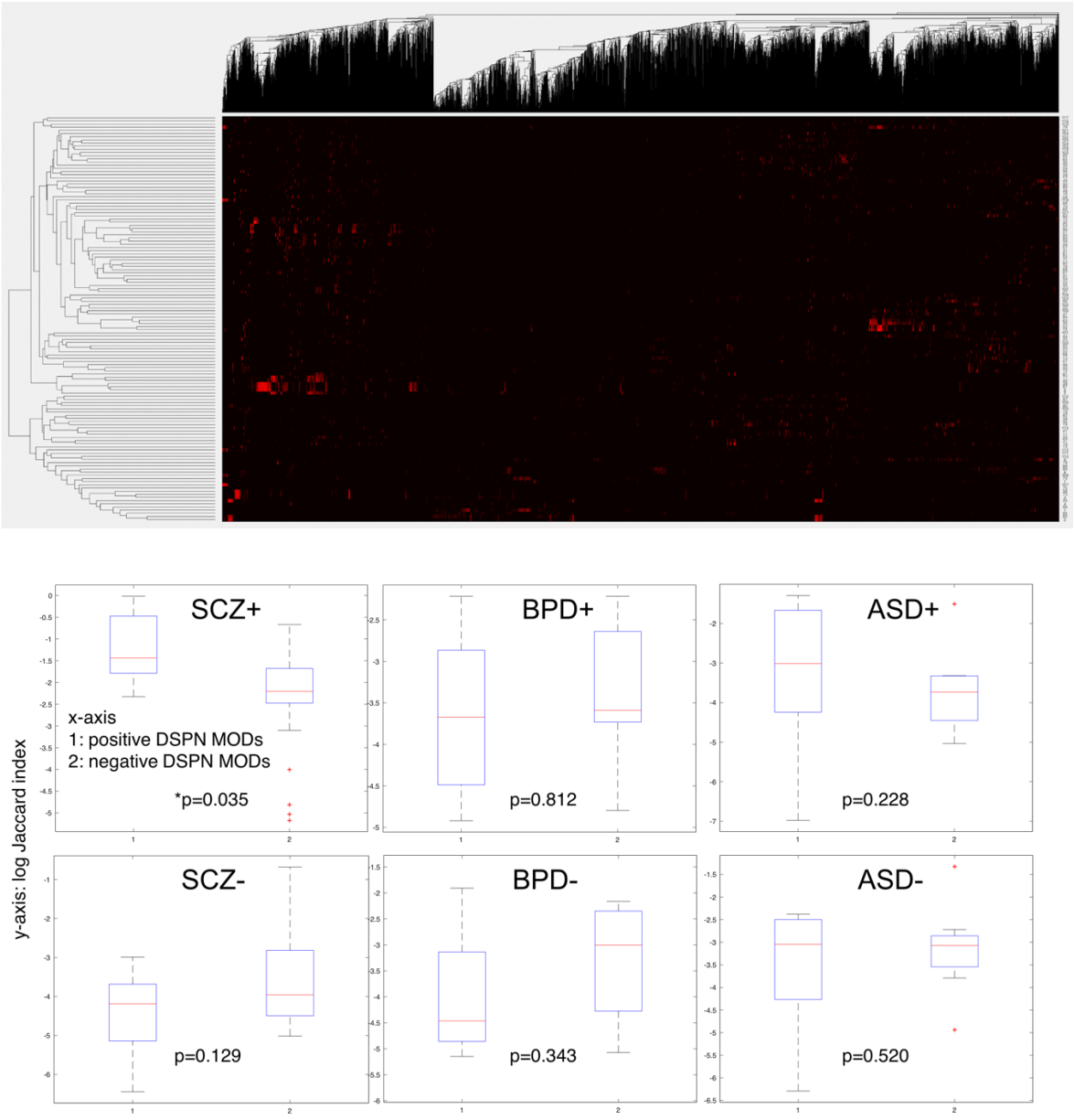


Fig. S49. Comparison of prioritized modules with those in up- and down-regulated disease modules in *(16)*. Top panel shows a clustergram of the Jaccard index between all pairs of modules used in our DSPN analysis (5012 modules, horizontal axis), and those from the analysis of *(16)* (128 modules, vertical axis; red shows high Jaccard index). Bottom panel compares the log Jaccard index scores between positive and negative DSPN modules, with up-regulated (SCZ+, BPD+, ASD+) and down-regulated (SCZ-, BPD-, ASD-) modules from *(16)*. As expected, the trend is for an increased score between positive and up-regulated modules, and negative and down-regulated modules. Although a significant p-value is achieved individually only for SCZ+ (1-tailed Wilcoxon rank-sum p-value shown), across pairs from all disorders a test for higher Jaccard indices for matching directions remains significant (*p=0.05, 1-tailed Wilcoxon rank-sum test). Boxes show 25th and 7th percentile, with red line at median, and whiskers show range after excluding outliers. For details, see Section S8.5.

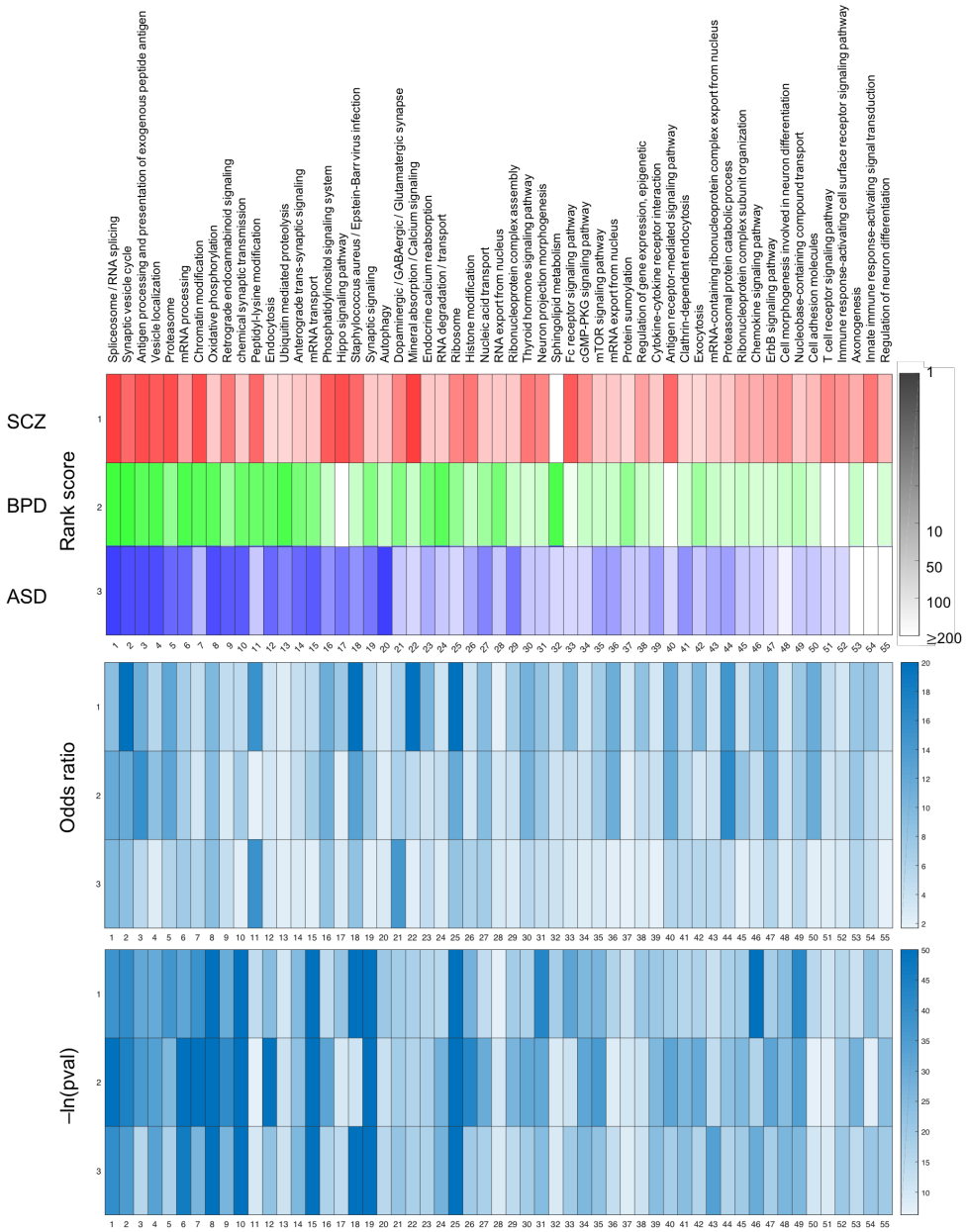


Fig. S50. Extended ranking of functional terms from DSPN enrichment analysis. The figure shows the full ranking of KEGG and GO (biological process) functional terms corresponding to Fig. 8C. All terms that have ranks between 1-10 in any disorder individually (in either positive or negative rankings) are included. The odds ratio and –ln(pval) corresponding to the module most enriched for this term in each disorder is also shown from the gene-set enrichment analysis. Functional terms associated with all prioritized modules are found on the website (using $p<0.05$ and $q<0.1$). For the rankings shown, KEGG terms were excluded from neurological diseases, known confounders due to treatment effects, terms that are near duplicates of others already listed, and associations which were judged as likely to be spurious, leading us to exclude the following terms: Parkinson's disease, Huntington's disease, Alzheimer's disease, Cardiac muscle contraction, Pertussis, Leishmaniasis, Thermogenesis, Non-alcoholic fatty liver disease (NAFLD), Salivary secretion, Osteoclast differentiation, Gastric cancer, Inositol phosphate metabolism, Hematopoietic cell lineage, Ferroptosis, Platelet activation, and Leukocyte transendothelial migration. For details, see Section S8.5.

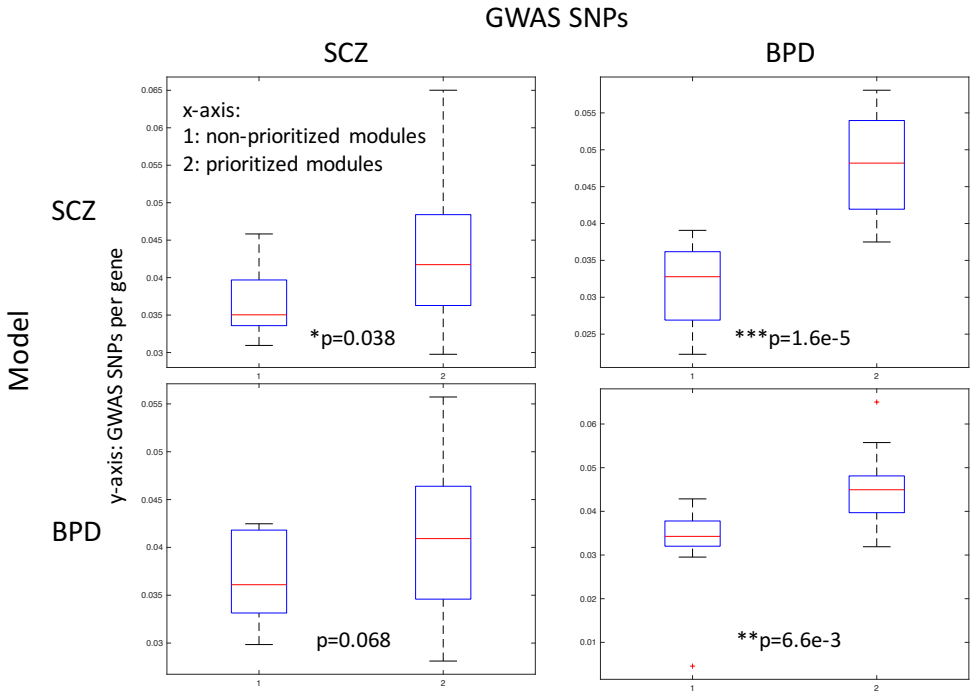


Fig. S51. Enrichment of GWAS SNPs in DSPN prioritized modules. The figure shows enrichment of GWAS SNPs associated with SCZ and BPD in the DSPN modules prioritized in the SCZ and BPD models. SNPs are linked with prioritized modules using all eQTLs associated with genes they contain. Enrichment is tested using a one-tailed Mann-Whitney test for an increase in the number of GWAS SNPs per gene in prioritized versus non-prioritized modules. We observed 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. Boxes show 25th and 7th percentile, with red line at median, and whiskers show range after excluding outliers. For details, see Section S8.5.

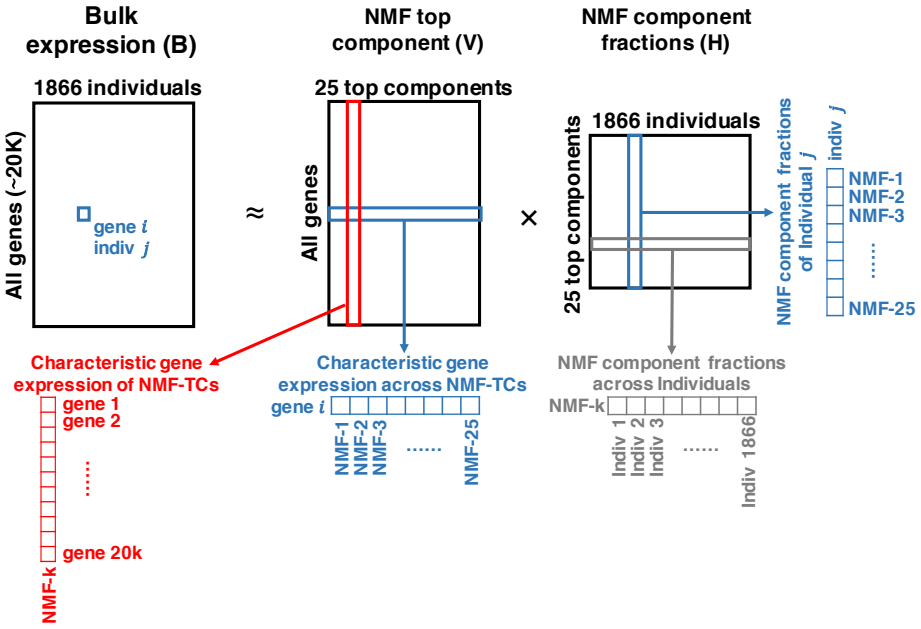


Fig. S52. NMF decomposition of bulk gene expression. The bulk gene expression is decomposed into the product of two matrices: an NMF component matrix (V, genes by top NMF components; i.e., NMF-TCs) and a component fraction matrix (H, top NMF components by individuals); i.e., B≈VH. For details, see Section S2.3.

S11. Supplementary Tables

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 S1. 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; the number of total individuals of these three studies is 896. For details, see Section S5.

Abbreviation	Adult/Developmental	Full name
Gran	Adult	Granule Neuron
Purk1/Purk2	Adult	Purkinje Neuron
Microglia/Micro	Adult	Microglia
OPC	Adult	Oligodendrocyte progenitor cells
Endothelial/Endo	Adult	Endothelial cells
Astrocytes/Astro	Adult	Astrocytes
Oligo	Adult	Oligodendrocyte
Perictyes/Peri	Adult	Pericytes

Table S2. Summary of cell types of cerebellum dataset. This table includes public adult cerebellum cell types from (*64*). For details, see Section S2.

Abbreviation	Adult/Developmental	Full name	Source
Ex	Adult	Excitatory Neuron	(*63, 64*)
In	Adult	Inhibitory Neuron	(*63, 118*)
OPC	Developmental	Oligodendrocyte progenitor cells	(*93*)
Trans	Developmental	Transient cell type (nascent neurons)	(*93*)
NEP	Developmental	Neuroepithelial cells	(*93*)
IPC	Developmental	Intermediate progenitor cells	(*93*)
Quiescent/Quies	Developmental	Quiescent newly born neurons	(*65*)
Replicating/Repli	Developmental	Replicating neuronal progenitors	(*65*)
IntN	Developmental	Inhibitory Neuron	(*93*)
ExtN	Developmental	Excitatory Neuron	(*93*)
Oligo	Developmental	Oligodendrocyte cells	(*93*)
Astrocytes/Astro	Developmental	Astrocytes	(*93*)
Pericytes/Peri	Developmental	Pericytes	(*93*)
Endothelial/Endo	Developmental	Endothelial cells	(*93*)
Microglia/Micro	Developmental	Microglia	(*93*)
Microglia/Micro	Adult	Microglia	(*64, 65*)
OPC	Adult	Oligodendrocyte progenitor cells	(*65*)
Endothelial/Endo	Adult	Endothelial cells	(*65*)
Astrocytes/Astro	Adult	Astrocytes	(*65*)
Oligo	Adult	Oligodendrocyte	(*65*)
OtherNeuron	Adult	Mixed of excitatory and inhibitory neuronal cells	(*65*)

Table S3. Summary of cell types of read count dataset. This table includes PsychENCODE developmental cell types, and public adult cell types from (*63, 65*). For details, see Section S2.

Abbreviation	Adult/Developmental	Full name	Source
Ex	Adult	Excitatory Neuron	(*64, 93*)
In	Adult	Inhibitory Neuron	(*64, 93*)
Microglia/Micro	Adult	Microglia	(*64, 93*)
OPC	Adult	Oligodendrocyte progenitor cells	(*64, 93*)
Endothelial/Endo	Adult	Endothelial cells	(*64, 93*)
Astrocytes/Astro	Adult	Astrocytes	(*64, 93*)
Oligo	Adult	Oligodendrocyte	(*64, 93*)
Perictyes/Peri	Adult	Pericytes	(*64, 93*)

Table S4. Summary of cell types of UMI count dataset. This table includes PsychENCODE adult cell types, and public adult cell types from (*64, 93*). For details, see Section S2.

	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 S5. RCA centroid distances. Euclidean distance was calculated between all tissue centroids in RCA space. For details, see Section S4.1.

	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 S6. 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. For details, see Section S4.1.

Reactome pathways	**# of genes in ref.**	**# or genes obser.**	**expected**	**Fold Enrichment**	**Enriched (+) or Depleted (-)**	**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 S7. Reactome pathway enrichment for most impactful genes in the RCA PC1 dimension. Pathway enrichment for the top genes selected in the Fig. S28 analysis. For details, see Section S4.1.

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	(*87*)	855,987,816	1,834,759,860

Table S8. Summary of Hi-C datasets. For details, see Section S6.1-3.

Disorders	Source
ADHD	(*119*)
ASD	(*120*)
Bipolar disorder (BPD)	(*121*)
Major depression	(*122*)
Schizophrenia (SCZ)	(*97*)
Educational attainment	(*123*)
Intelligence	(*124*)
Alzheimer's disease	(*125*)
Parkinson's disease	(*126*)
Type 2 diabetes (T2D)	(*127*)
Coronary artery disease (CAD)	(*128*)
Inflammatory bowl disease (IBD)	(*129*)

Table S9. GWAS datasets used for heritability enrichment analysis. The enrichment analysis is described in detail in Section S7.4.

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%)	63.3% (10.8%)	67.4%	85.7%	81.5%
DSPN-impute (a)	56.4% (1.2%)	61.7% (5.4%)	58.8% (3.2%)	-	-	-
DSPN-full (a)	67.9% (16.3%)	66.1% (30.0%)	68.3% (11.3%)	69.7%	92.7%	83.9%

Table S10. 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-impute and DSPN-full models are as in Fig. 7C for all phenotypes except ASD. For details, see Section S8.2.

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 (*76*)	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 data (*64*)	CTL	Frontal Cortex (BA6 and BA10)	scRNA-seq	Bulk Tissue Deconvolution and Decomposition	10319
Published Single-cell data (*63*)	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 data (*65*)	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 (Developmental)	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	459
	CTL	Dorsal Pallium (Developmental)	scRNA-seq	Bulk Tissue Deconvolution and Decomposition, fQTL	473
	CTL	Dorsolateral Prefrontal Cortex	scRNA-seq	Bulk Tissue Deconvolution	17093

		(Adult)		and Decomposition	
PEC: Reference Brain	CTL	Dorsolateral Prefrontal Cortex	ChIP-seq: H3K27ac	Enhancer Definition	1
	CTL	Dorsolateral Prefrontal Cortex (pooled HiC matrices from 3 reference brains)	HiC (10kb resolution)	Enhancer Definition	1
	CTL	Dorsolateral Prefrontal Cortex (pooled HiC matrices from 3 reference brains)	HiC (40kb resolution)	TAD and compartment 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

Table S11. 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. This table is provided in downloadable form as RAW-01 on the website (resource.psychencode.org). For details, see Section S1.

Cell type	ASD vs CTL	BPD vs CTL	SCZ vs CTL	ASD vs BPD	ASD vs SCZ	BPD vs SCZ	M vs F	Age trend
Adult-Ex1	1	1	8.81E-02	1	1	4.75E-02	6.20E-02	1.40E-01
Adult-Ex2	1	1	1	1	1	1	1	3.13E-01
Adult-Ex3	8.53E-02	1.32E-02	2.96E-01	1	1	7.22E-06	7.21E-01	6.30E-10
Adult-Ex4	8.00E-03	9.51E-02	7.62E-01	2.89E-01	8.80E-02	5.27E-04	4.72E-04	1.47E-06
Adult-Ex5	4.65E-06	2.34E-03	1.39E-01	3.67E-03	1.70E-03	2.35E-03	2.22E-02	9.32E-02
Adult-Ex6	1	1	1	1	1	1	1	1.25E-01
Adult-Ex7	5.79E-03	1	1	8.72E-01	1	1	1	6.68E-04
Adult-Ex8	8.28E-08	1	8.81E-02	4.54E-06	2.44E-06	8.66E-03	3.16E-02	1.90E-01
Adult-In1	1	1	1	1	1	1	1	1
Adult-In2	1	1	1	1	1	1	1	1
Adult-In3	1	1	1	1	1	1	1	1
Adult-In4	1	1	1	1	1	1	1	1
Adult-In5	1	1	1	1	1	1	1	1
Adult-In6	1	1	1	1	1	1	1	8.03E-11
Adult-In7	1	1	1	1	1	1	1	1
Adult-In8	4.65E-06	1.62E-01	5.18E-01	6.38E-04	2.57E-04	1	1.50E-04	7.16E-04
Adult-Astrocytes	2.73E-01	7.79E-01	4.24E-07	8.72E-01	1	2.91E-08	1.58E-01	2.05E-14
Adult-Endothelial	5.79E-03	2.34E-03	1.51E-02	8.72E-01	1	5.27E-01	1.94E-01	1.34E-05
Fetal-Quiescent	1	1	1	1	1	1	2.87E-01	6.25E-02
Fetal-Replicating	1	1	1	1	1	1	2.29E-01	6.56E-01
Adult-Microglia	3.86E-01	1.38E-01	2.81E-05	5.75E-01	2.35E-02	3.60E-03	1.10E-01	4.75E-30
Adult-OtherNeuron	1.90E-01	9.81E-02	1.37E-05	1.37E-01	8.11E-01	4.29E-03	1.72E-06	3.76E-17
Adult-OPC	1	1	1	1	1	1	5.83E-01	3.71E-06
Adult-Oligo	4.65E-06	2.35E-01	8.25E-01	4.79E-02	6.43E-03	1.95E-01	1	2.05E-14

Table S12. Summary of statistical test FDR of cell fractions. This table provides the FDR (corrected p-values) of the statistical test of cell fractions. The tests including ks-test between disorders and genders, and age trend analysis of control samples. For details, see Section S2.4.

A

Gene	eQTL SNPs
GRIN1	9:140306414, 9:140744847
CLU	8:27704027
RBFOX1	16:5387873, 16:5935585
ANK2	4:113369725, 4:113376696, 4:113381943, 4:113387640, 4:113313848, 4:11332091, 4:113320989
RELA	11:65453678, 11:65492891, 11:65571606, 11:65573587, 11:65576788, 11:65577427, 11:65578759, 11:65579600, 11:65583340, 11:65592935, 11:65596546, 11:65664346 11:65665256
HOMER1	5:78837521

B

Enhancer	cQTL SNPs
EH37E0947082	8:27524060
EH37E0245781	11:125311674

Table S13. eQTL and cQTL coordinates for genes and enhancers of interest. This table shows genome coordinates for all eQTLs and cQTLs shown in the model trace examples from Fig. 8D. See the website resource.psychencode.org, files DER-08 and DER-09 for full eQTL and cQTL files. For details, see Section S5.5.