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 Navarro[1,2*], Declan Clarke[1,2*], Mengting Gu[1*], Prashant Emani[1,2*], Yucheng T. Yang[1,2], Min Xu[1,2], Michael 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 Hoffman[18], PsychENCODE Consortium[‡], Panos Roussos[17,18], Schahram Akbarian[17,19], Andrew E. Jaffe[21], Kevin White[16], Zhiping Weng[15], Nenad Sestan[20], Daniel H. Geschwind[7-9†], James A. Knowles[10†], Mark Gerstein[1,2,11,12†]

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]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[20]Department of Neuroscience and Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT 06520, USA
[21]Lieber Institute for Brain Development, Johns Hopkins Medical Campus; Departments of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health Baltimore, MD, 21205, USA

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

Abstract

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, ~2.5M 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 enables construction of a gene-regulatory network, linking GWAS variants to genes (e.g., 321 for schizophrenia). We embed the 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 with brain genomics data, at scale (*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 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. Broadly, 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 imputed 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 by individual users 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*). For instance, we found 12,080 genes were transcribed in the brains of 95% of the individuals surveyed, and >16,000 protein-coding and >9,000 non-coding genes were detected in total (in PFC) (*20, 21*).

Brain tissues are 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 remains unknown how these changes in cell proportions can quantitatively contribute to the variation in tissue-level gene expression observed across a population of individuals. To address this question, we used two complementary strategies examining expression 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**), were consistently associated with the single-cell signatures (Figs. 2B and 2C) (*20*). A number of NMF-TCs were correlated with cell types from matrix **C** for both TPM and UMI data -- e.g., component NMF-17 is highly correlated with the Ex2 cell type (r=0.63, Figs. 2C and S9). This demonstrates that an unsupervised analysis

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

We then examined how variation in 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 error = 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 ~89% (Fig. 2D, $1 - || \mathbf{B} - \mathbf{CW} ||^2 / || \mathbf{B} ||^2 = 89\%$), 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 at both population and individual levels 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; the fraction of In6, for instance, was significantly higher in females (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*) . SST exhibits increasing promoter methylation with age, perhaps explaining its decreasing expression. Other genes known to be associated with brain aging exhibit different trends, such as 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 (Fig. S22). Consistent with the ENCODE annotations, 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.

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 (e.g., H3K27ac) at enhancers across different individuals and tissues high, but the boundaries of enhancers can

grow and shrink, sometimes disappearing altogether (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 ChIP-seq data across the cohort resulted in a total of 37,761 H3K27ac "peaks" (enriched regions) in PFC, 42,683 in TC, and 26,631 in CB -- 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 signal). As expected, not every reference enhancer was active in each cohort 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 proportion of the reference enhancers (~68%) was active in more than half the cohort (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 the identified 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 the transcriptome variation to that of the epigenome. Several approaches, including PCA, t-SNE, and reference component analysis (RCA) were tested to determine the most appropriate 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 using 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 focused on inter-tissue differences in annotated regions (i.e., genes, promoters, and enhancers). However, in addition to the canonical expression differences in protein-coding genes, we also found differences in unannotated non-coding and intergenic regions (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, adhering closely to the GTEx pipeline (Figs. S31, S32, and S33; (*34*)). 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 conservative 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 testing for overlap 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). We also applied the same QTL pipeline to splicing and identified 2,628,259 isoQTLs associated with the changes in isoform usage, which are, in turn, associated with 19,790 distinct genes (*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 the effect of a fQTL may involve other QTL types, such as eQTLs. An illustrative example is the gene 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, taking into account our 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 and isoQTLs 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 eQTL overlap with cQTLs was the largest while that with fQTLs was lowest (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). Finally, 1391 SNPs, which we dubbed multi-QTLs, functioned as QTLs in at least 3 different capacities (e.g. as eQTLs, cQTLs and isoQTLs, Fig. 4F).

Regulatory networks

We next integrated the genomic elements described above at the regulatory-network level. We created a network revealing how the genotype and regulators relate to target gene expression. We first processed a Hi-C dataset for adult brain in the same reference samples used for enhancer identification, providing a physical basis for interactions between enhancers and promoters (Fig. 5A, 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 eQTLs and isoQTLs; 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 that 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 for cross-tissue transcriptome comparison, with fetal samples included (Fig. S41). We did a number of other comparisons with 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, to create a connection, we required that (i) it has a "good binding site" (matching the TF's motif) in gene-proximal open chromatin regions (either promoters or brain-active enhancers) and that (ii) it has 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 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). We provide both the full network and a high confidence subset (*20*). In particular, the subset 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*).

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. 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. 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*). 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*). 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*).

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 between intermediate phenotypes. 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 and submodules in the model, referring to this modification as "DSPN-mod" (Fig. S48). Using it, we determined which co-expression modules were prioritized, as well as the sets of genes associated with latent nodes that were uncovered 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 (Fig. 8B-C) and to intersect substantially with the modules from more disease-focused analyses (Fig. 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 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 key 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 a given cell type. In the future, larger-scale single-cell studies will allow us to examine this question in greater 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 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, the separation

we make between eQTL detection and integrative modeling allows us to compare our eQTL methods directly to previous analyses such as GTEx, which use univariate approaches. However, multivariate-based methods for identifying QTLs have been used elsewhere in the literature 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, such as the transcriptome; 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 to a given main text section within the supplementary section is named in a parallel fashion. Supplementary materials and methods are contained within their respective sections. Detailed data protocols are available in the supplement. The associated and derived data files are also available on the resource website, Resource.psychENCODE.org.

ACKNOWLEDGMENTS

We would like to acknowledge the National Institute of Mental Health (NIMH) for funding. Also, we acknowledge program staff, in particular T. Lehner, L. Bingaman, D. Panchision, A. Arguello and G. Senthil, for providing institutional support and guidance for this project.

Figures

Figure 1. Comprehensive data resource for functional genomics of the adult 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 >85% 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) Numbers of genes with at least one eQTL (eGenes) are shown across different studies. The number of eGenes increased as the sample size increased. The eGenes of PsychENCODE 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)** Similarity between PsychENCODE brain DLPFC eQTLs and GTEx eQTLs of other tissues are evaluated by π_1 values and SNP-eGene overlap rate. Both π_1 values and SNP-eGene overlap rate are higher in brain DLPFC than the other tissues. **(C)** 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 a 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)** Enrichment of genomic regions annotations of QTLs is 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. Asteriks (*) indicate that for cQTLs we only show the number of top SNPs for each enhancer. The isoQTLs were calculated based on isoform percentage (i.e., the percentage of the ratio between the transcript abundance and its parent gene's abundance). Overlaps of all QTL SNPs and overlap

of eQTL and isoQTL eGenes are shown in heatmaps (square rows). The linked circles show the overlap of QTL types. Sharing of other QTLs vs. eQTLs are evaluated using π_1 values in the orange bar plot. This indicates the highest sharing is between cQTLs and eQTLs. An example on the right for the MTOR gene shows the overlapping of eQTL SNPs of the gene and cQTL SNPs for the H3K27ac signal on an enhancer ~50kb upstream of the gene. Hi-C interactions (bottom of the plot) 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**.

‡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;

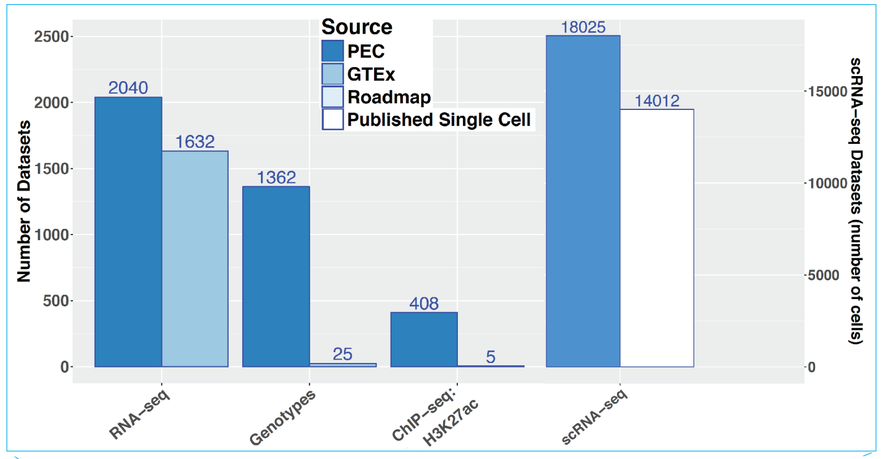
SUPPLEMENTARY MATERIALS

Supplementary Text

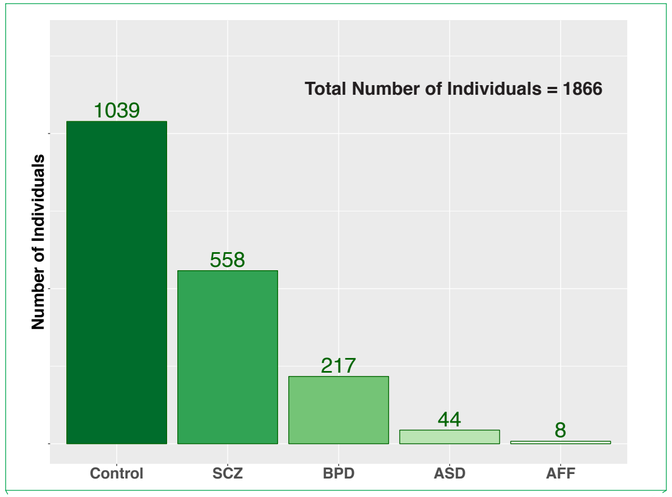
Figs. S1 to S52

Tables S1 to S13

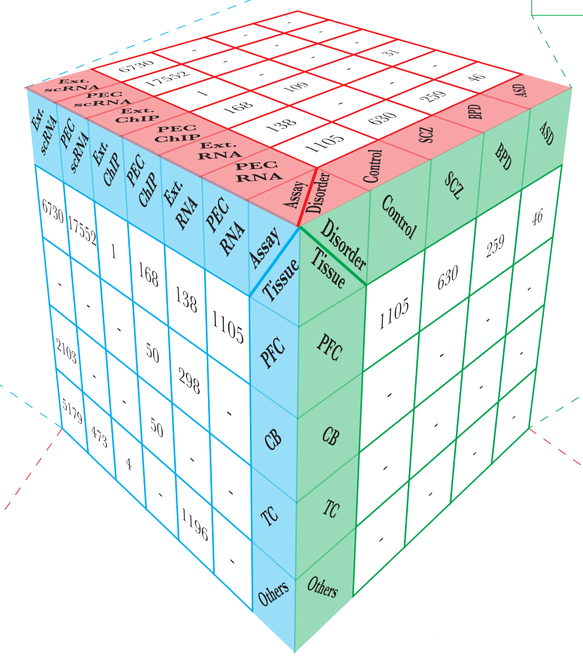
References (62–129)

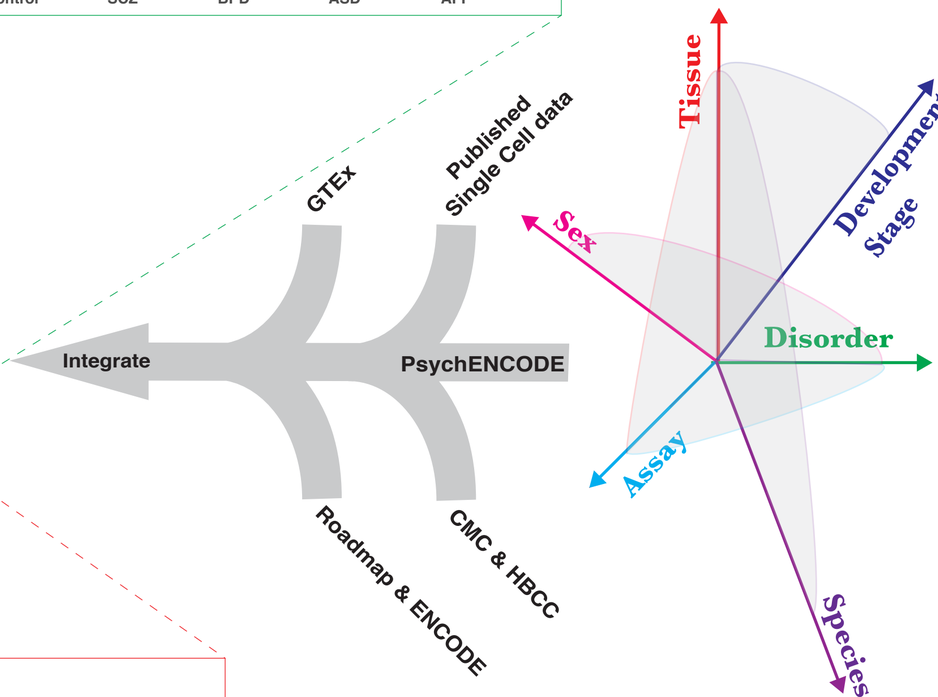


Assay

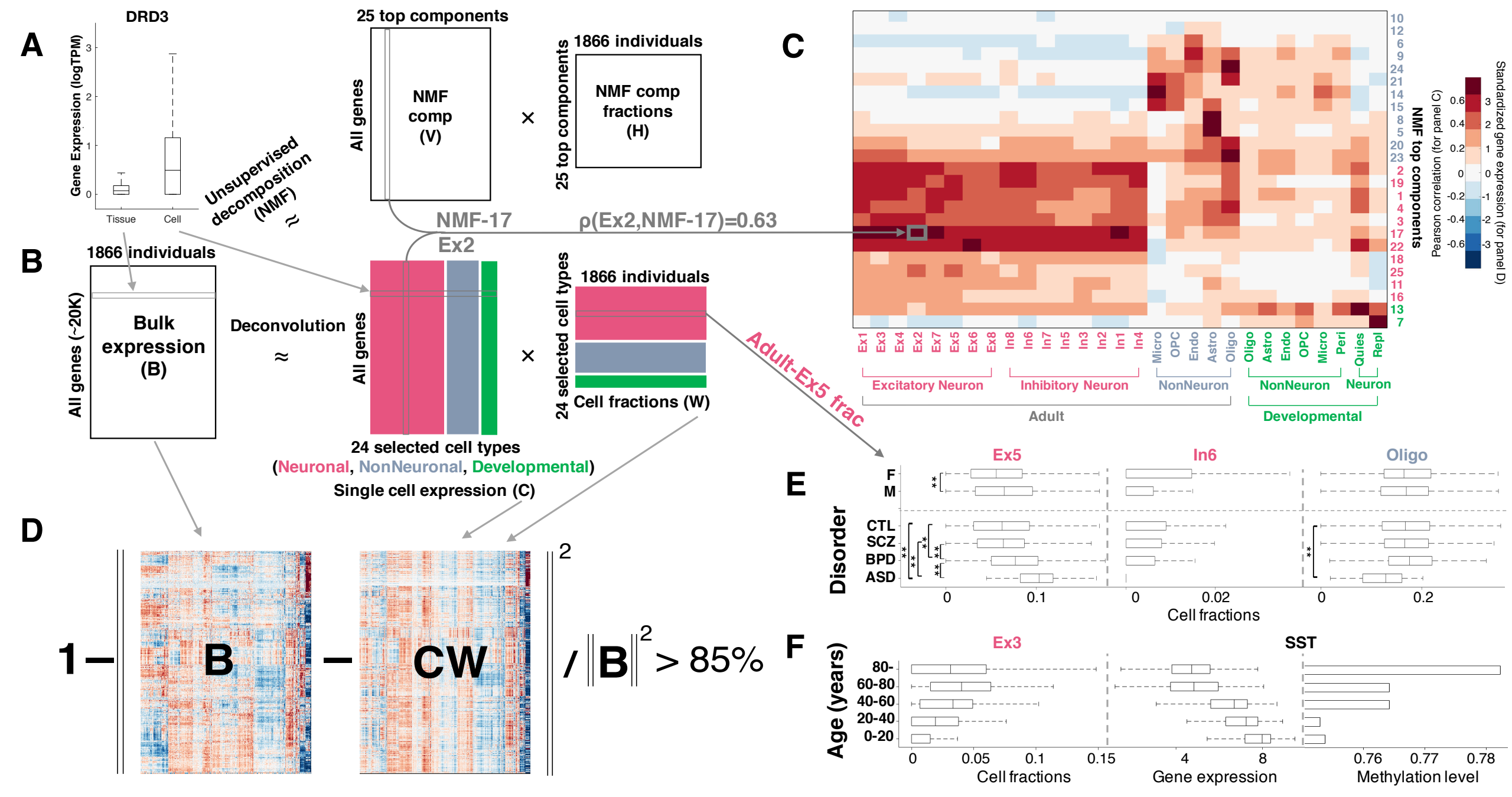


Disorder

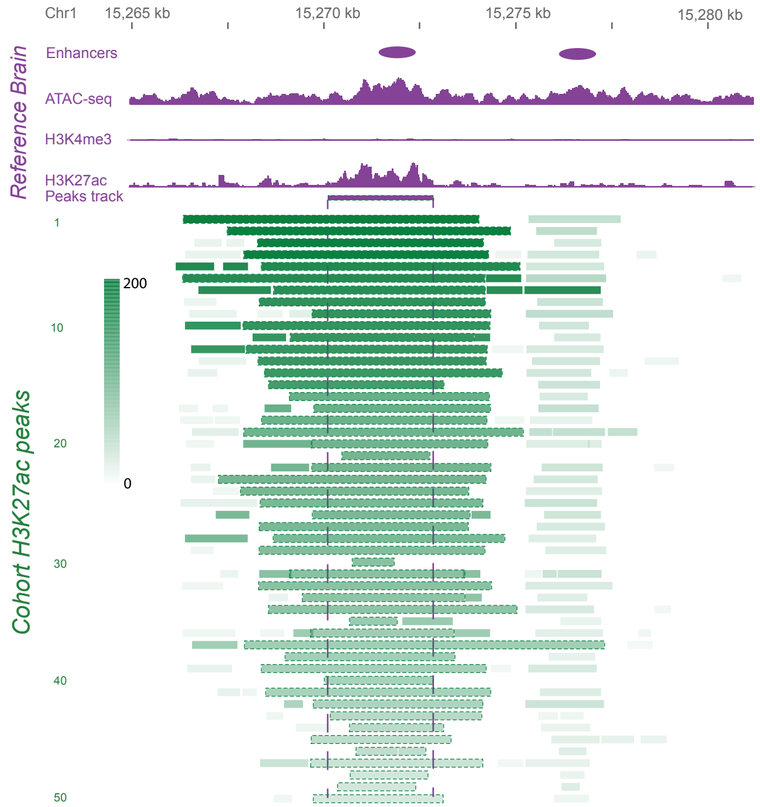




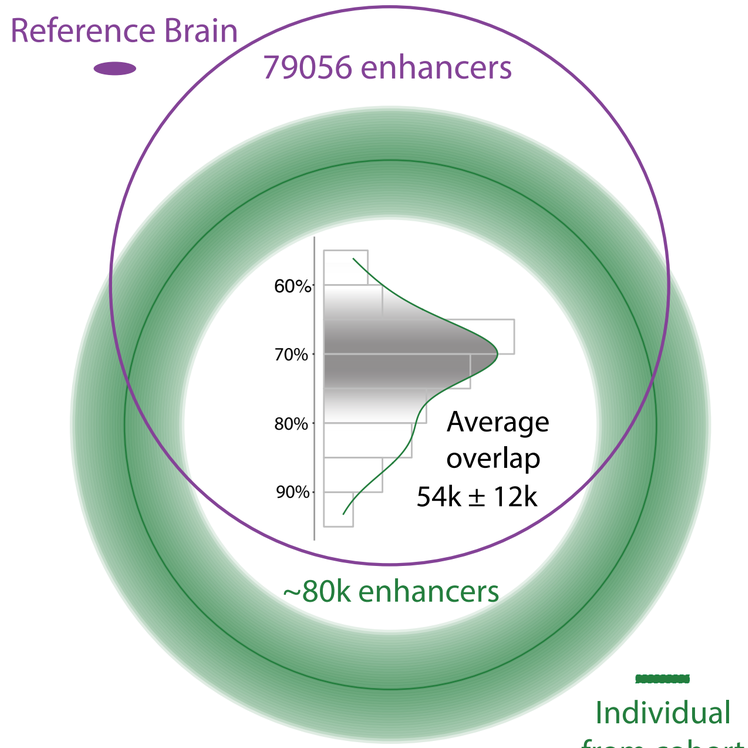
Tissue


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)
B
1866 individuals
All genes (~20K)
Bulk expression (B)
Deconvolution
≈
All genes
24 selected cell types
×
1866 individuals
24 selected cell types
Cell fractions (W)
24 selected cell types
(Neuronal, NonNeuronal, Developmental)
Single cell expression (C)
NMF-17
Ex2
ρ(Ex2,NMF-17)=0.63
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 > 85%
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

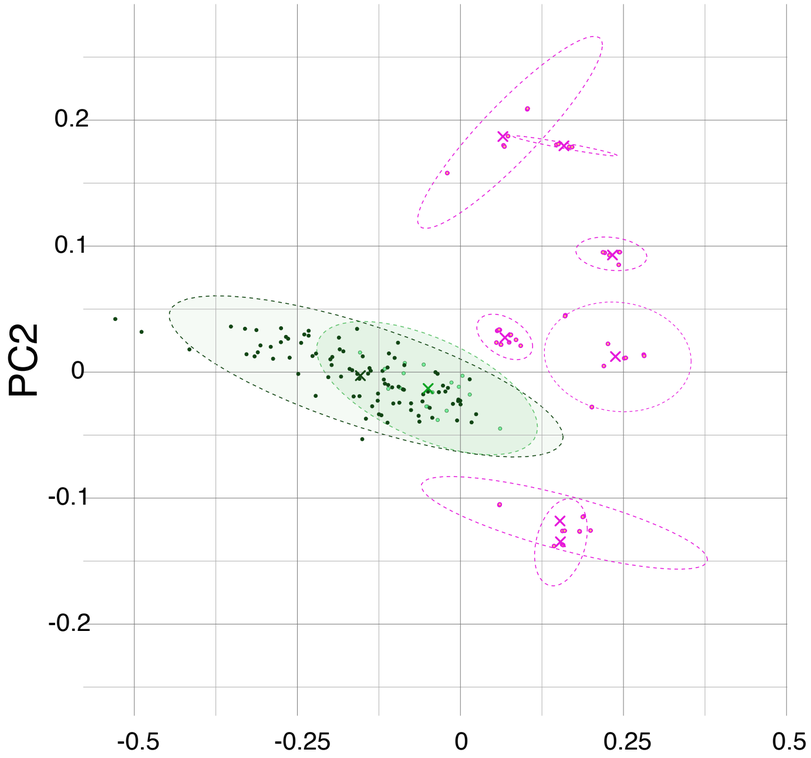
A


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

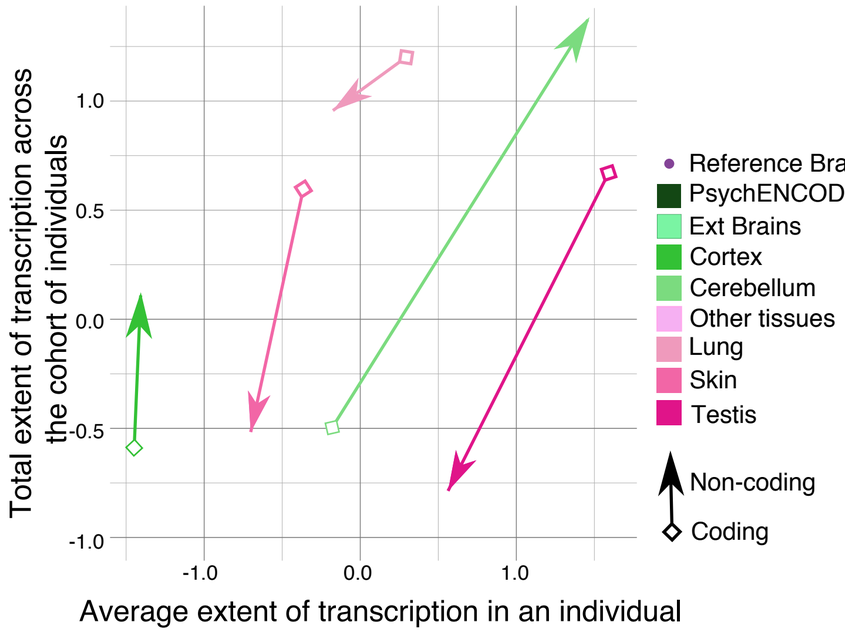
B


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

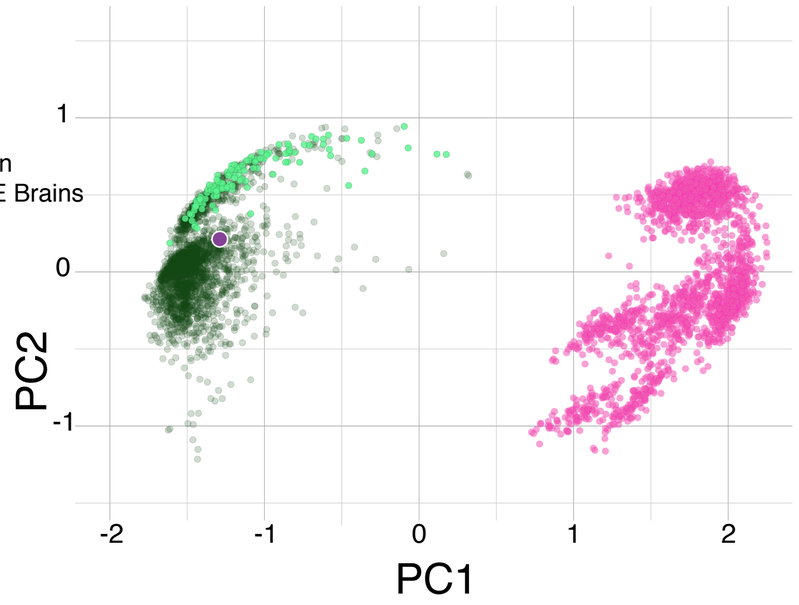
C


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

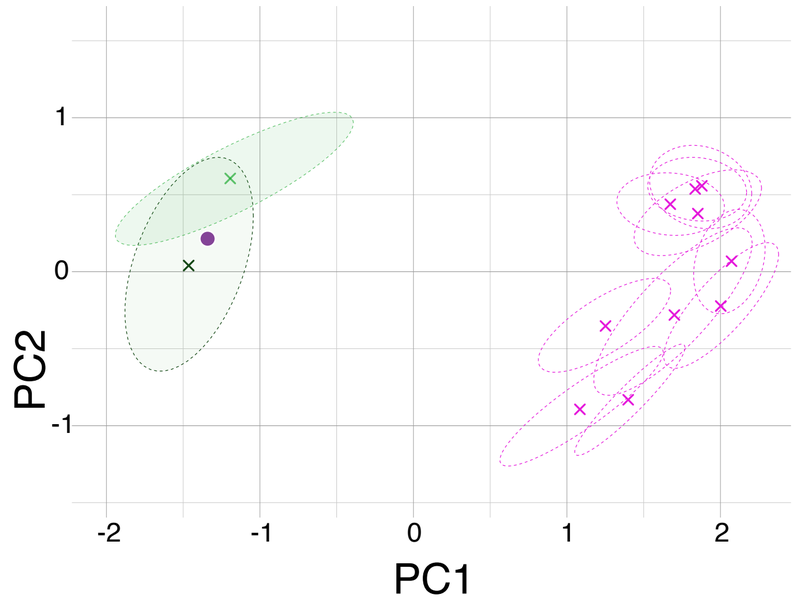
D

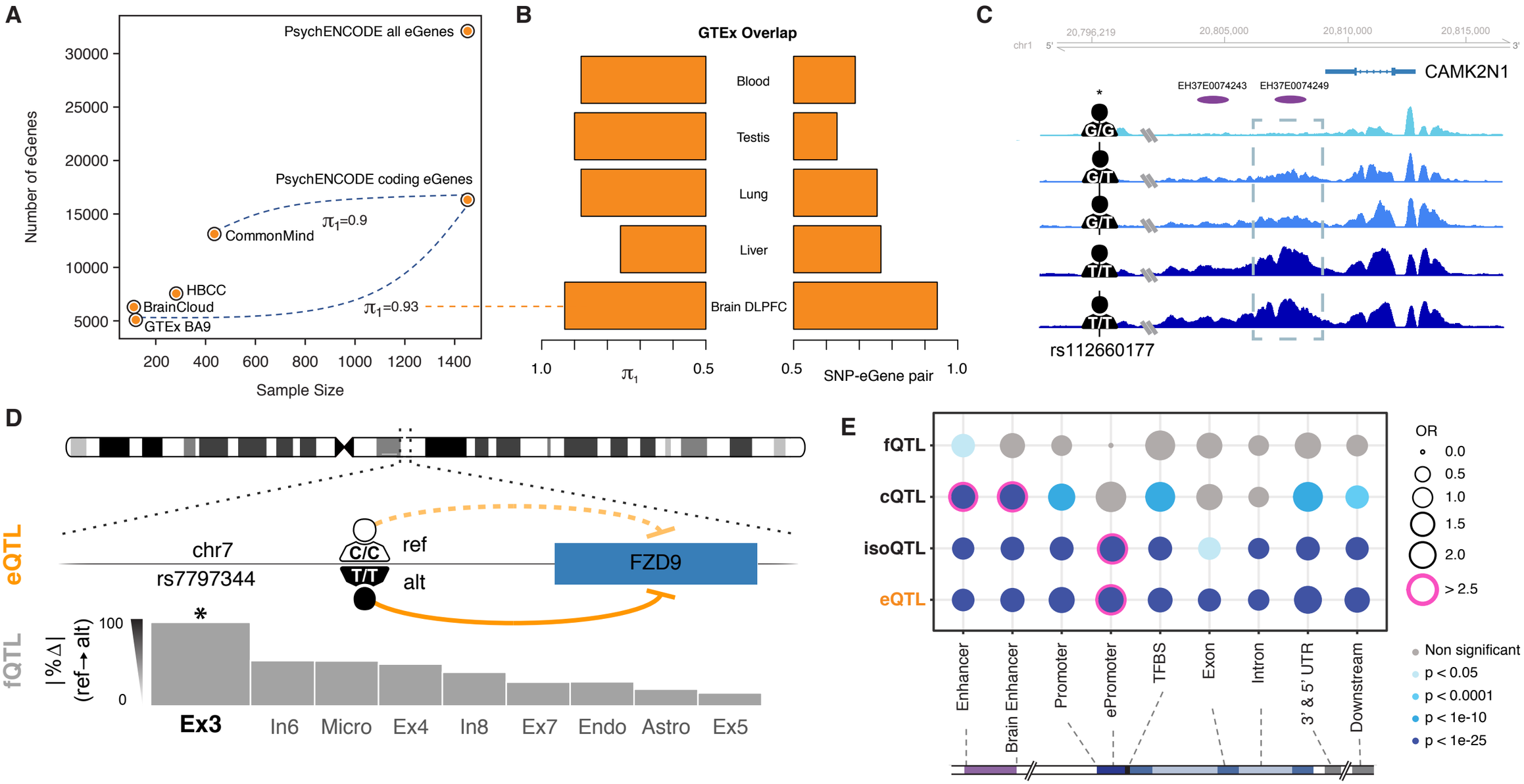

Total extent of transcription across
the cohort of individuals
1.0
0.5
0.0
-0.5
-1.0
-1.0
0.0
1.0
Average extent of transcription in an individual
Reference Brain
PsychENCODE Brains
Ext Brains
Cortex
Cerebellum
Other tissues
Lung
Skin
Testis
Non-coding
Coding

E

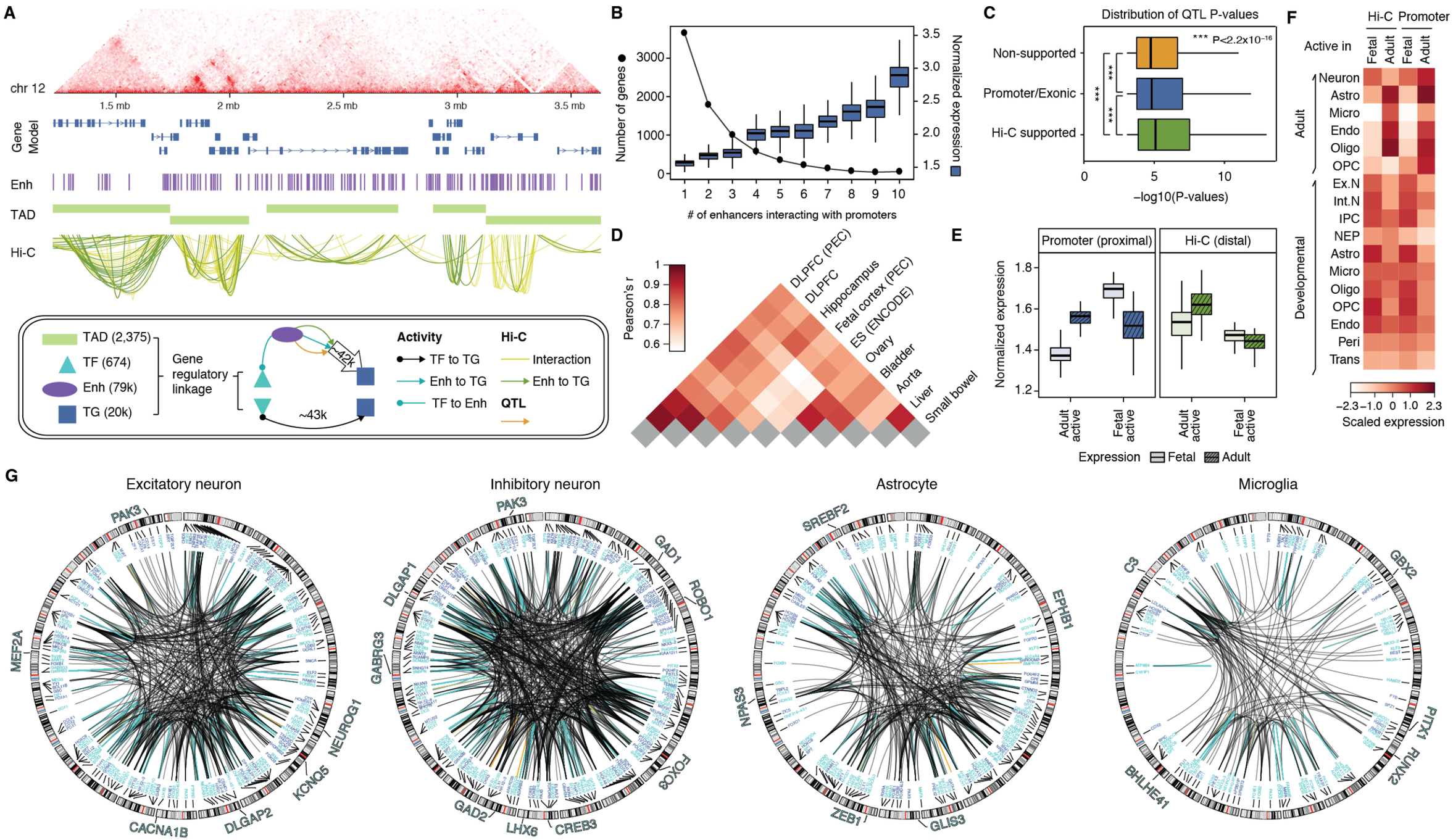

PC2
1
0
-1
-2
-1
0
1
2
PC1

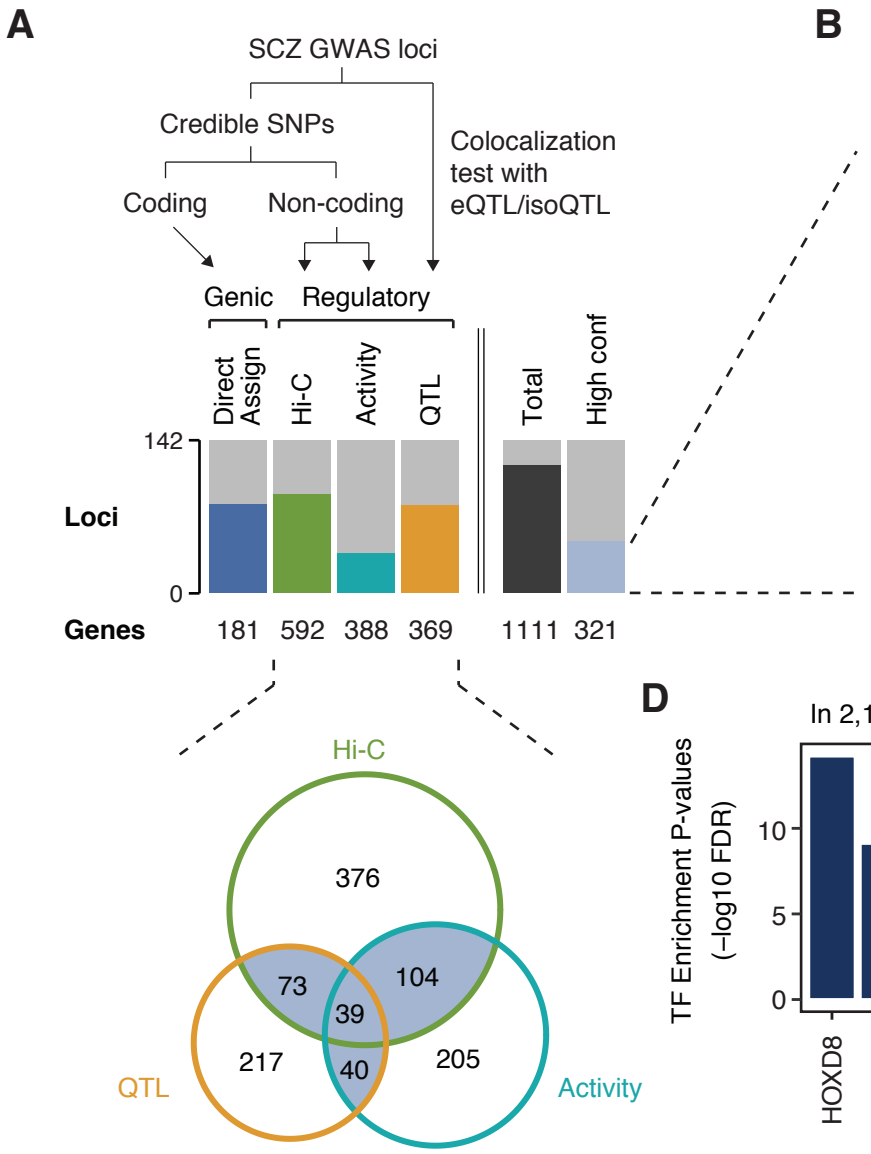
F

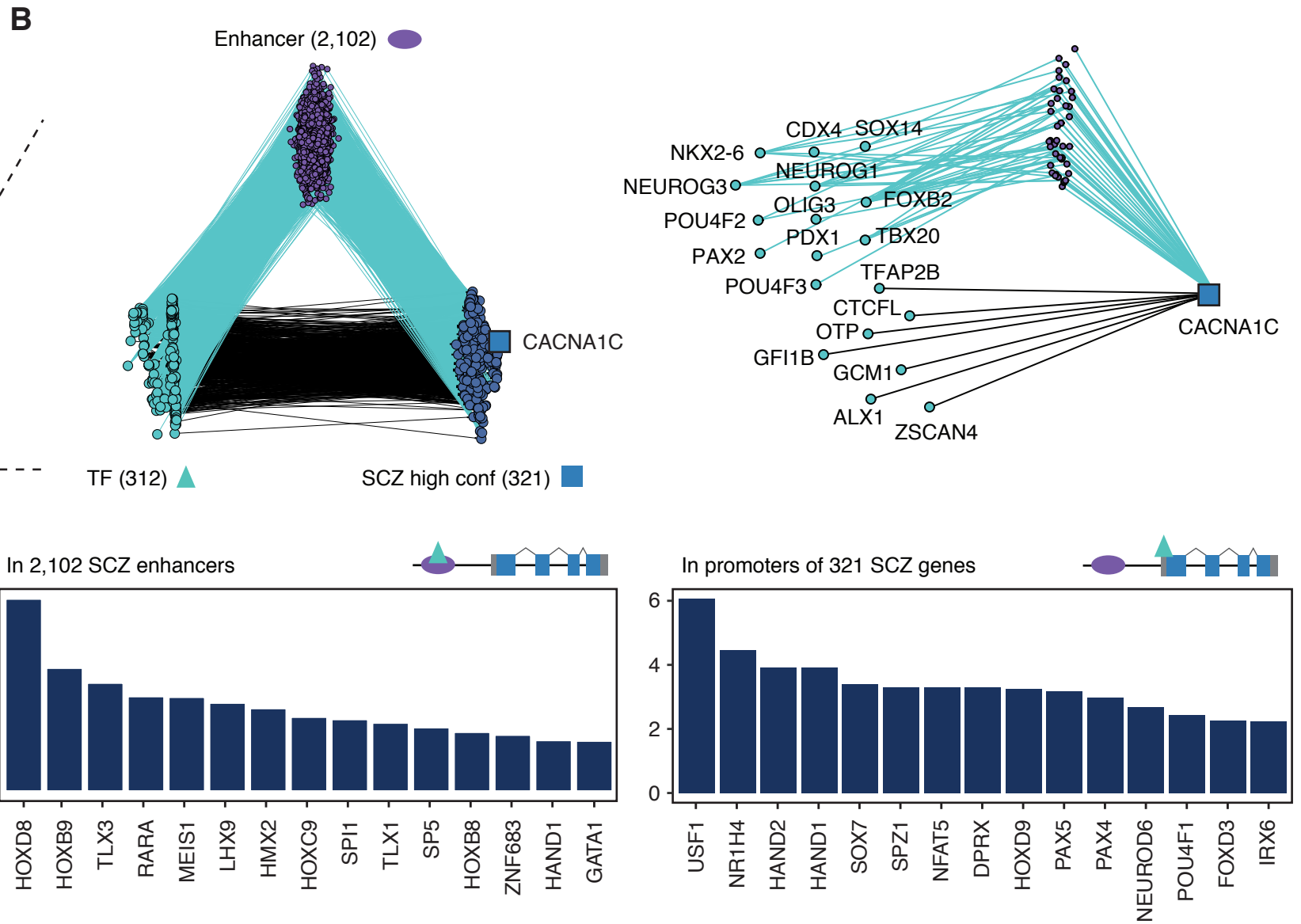

PC1
PC2
1
0
-1
-2
-1
0
1
2
PC1



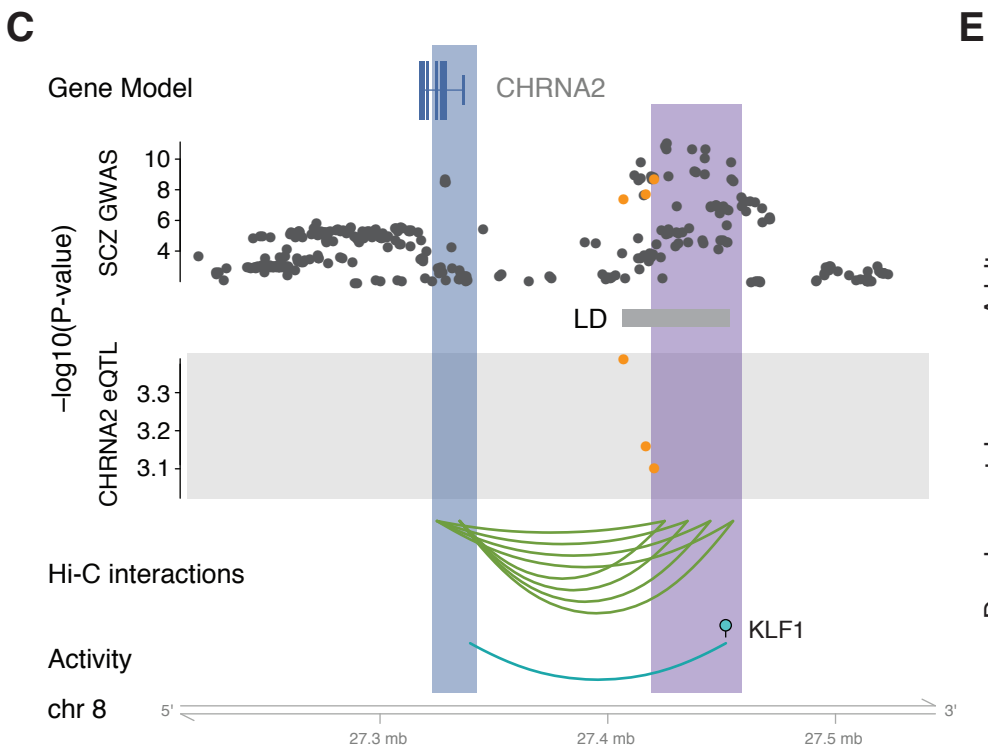
	Numbers of QTLs	eGenes Enhancers Cell types	SNPs
eQTL	2,542,908	32,944	1,341,182
isoQTL	2,628,259	19,790	1,052,939
cQTL*	8,464	8,484	7,983
fQTL	4,199	9	1,672

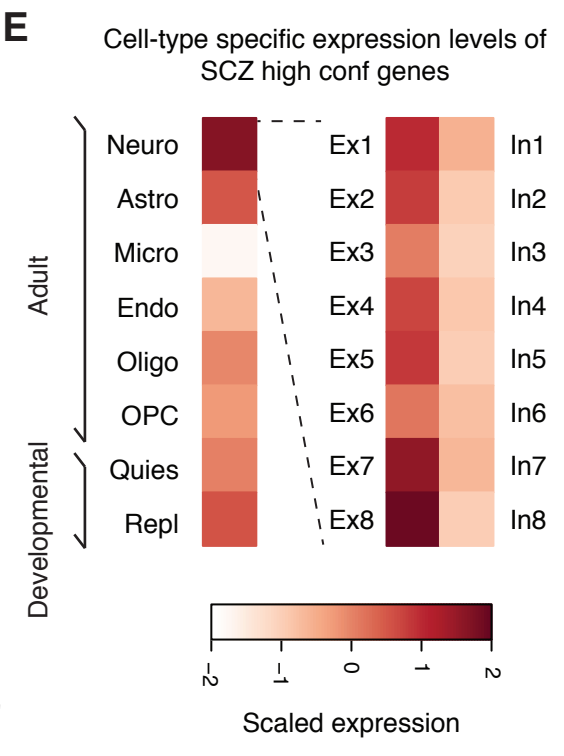

A
chr 12
1.5 mb
2 mb
2.5 mb
3 mb
3.5 mb
Gene Model
Enh
TAD
Hi-C
TAD (2,375)
TF (674)
Enh (79k)
TG (20k)
Gene regulatory linkage
~42k
~43k
Activity
TF to TG
Enh to TG
TF to Enh
Hi-C
Interaction
Enh to TG
QTL
B
Number of genes
Normalized expression
# of enhancers interacting with promoters
C
Distribution 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

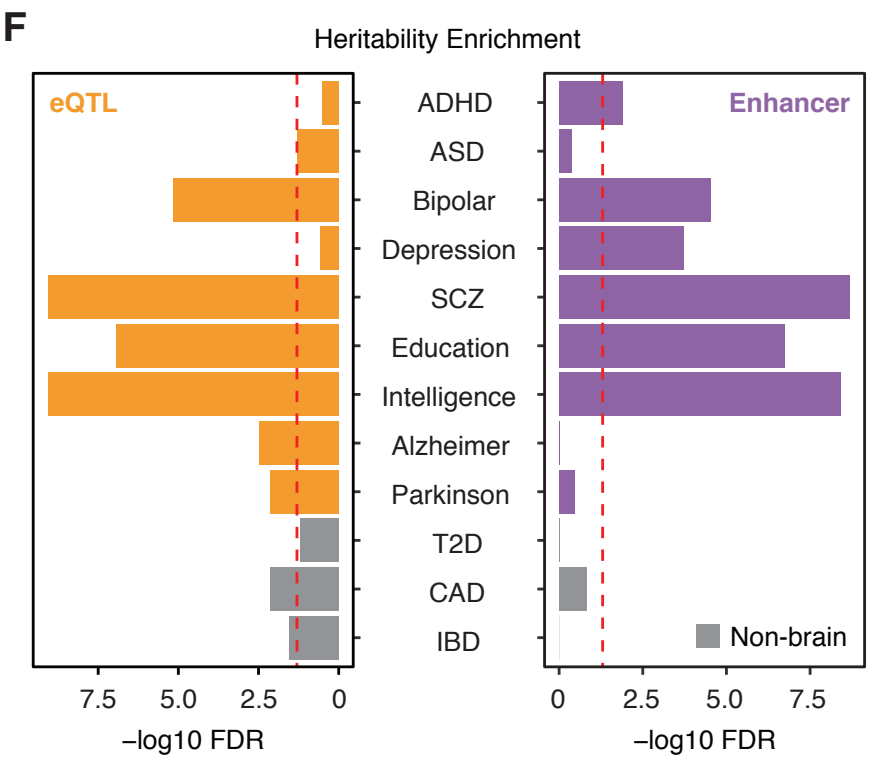

A
SCZ GWAS loci
Credible SNPs
Coding
Non-coding
Colocalization test with eQTL/isoQTL
Genic
Regulatory
Direct Assign
Hi-C
Activity
QTL
Total
High conf
142
Loci
0
Genes
181
592
388
369
1111
321
Hi-C
376
73
104
39
217
40
205
QTL
Activity


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
CACNA1C
OTP
GFI1B
GCM1
ALX1
ZSCAN4

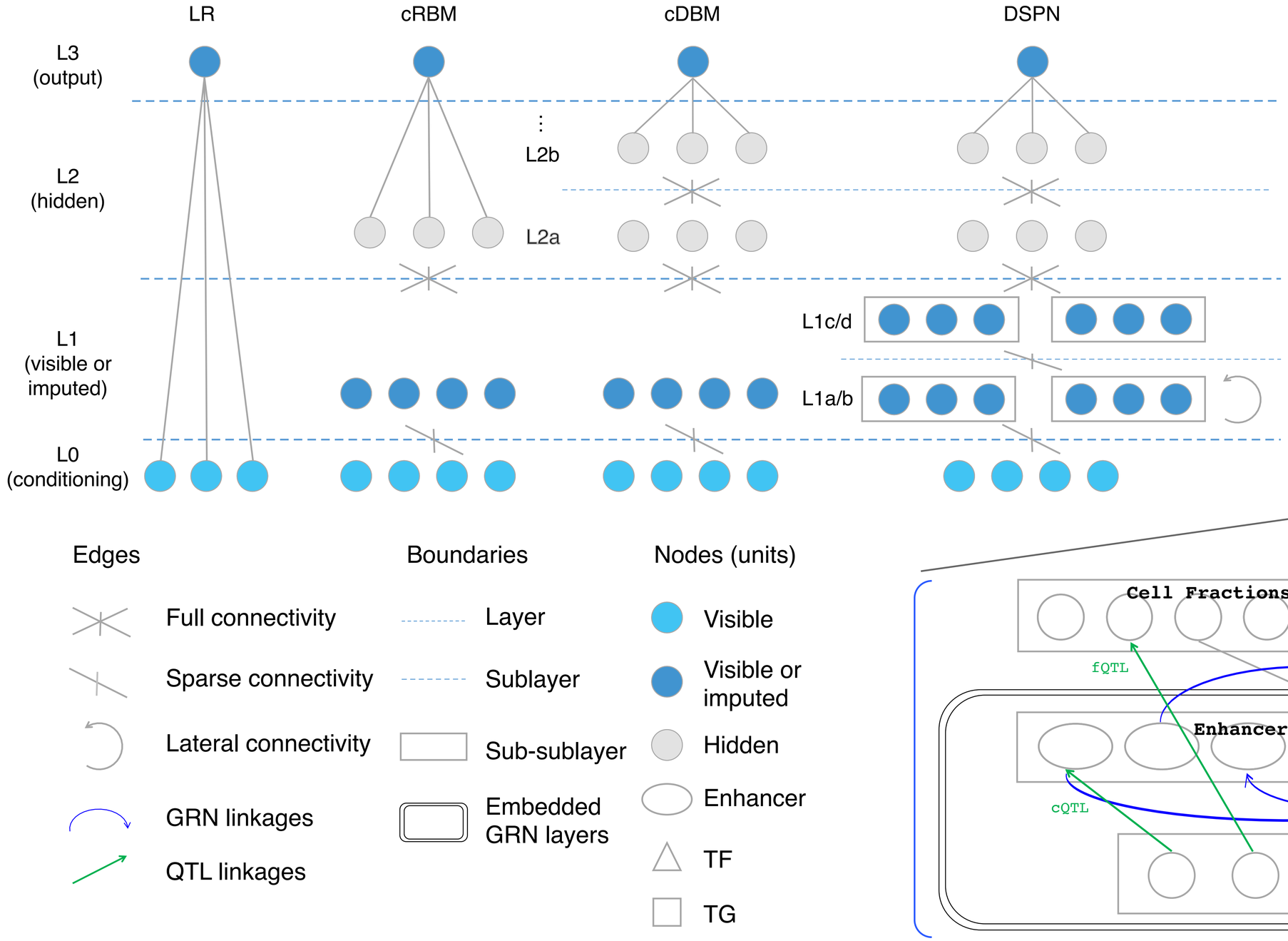
D
In 2,102 SCZ enhancers
TF Enrichment P-values (–log10 FDR)
10
5
0
HOXD8
HOXB9
TLX3
RARA
MEIS1
LHX9
HMX2
HOXC9
SPI1
TLX1
SP5
HOXB8
ZNF683
HAND1
GATA1
In promoters of 321 SCZ genes
6
4
2
0
USF1
NR1H4
HAND2
HAND1
SOX7
SPZ1
NFAT5
DPRX
HOXD9
PAX5
PAX4
NEUROD6
POU4F1
FOXD3
IRX6


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

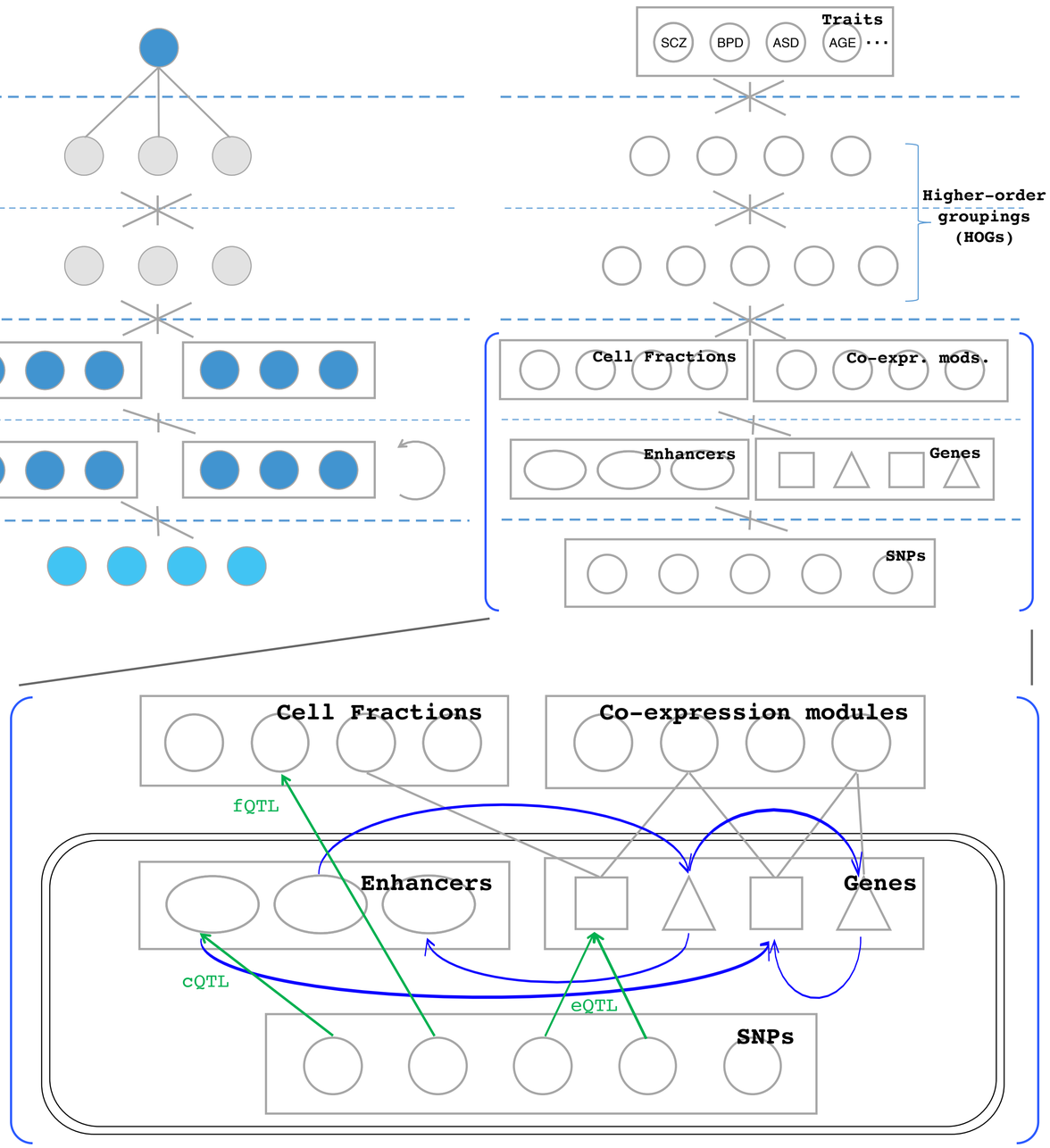

E
Cell-type specific expression levels of SCZ high conf genes
Adult
Developmental
Neuro
Astro
Micro
Endo
Oligo
OPC
Quies
Repl
Ex1
Ex2
Ex3
Ex4
Ex5
Ex6
Ex7
Ex8
In1
In2
In3
In4
In5
In6
In7
In8
–2
–1
0
1
2
Scaled expression


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

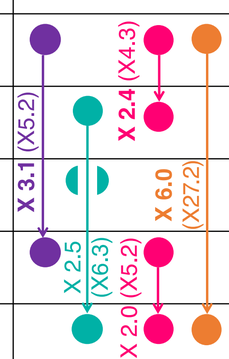


B



C

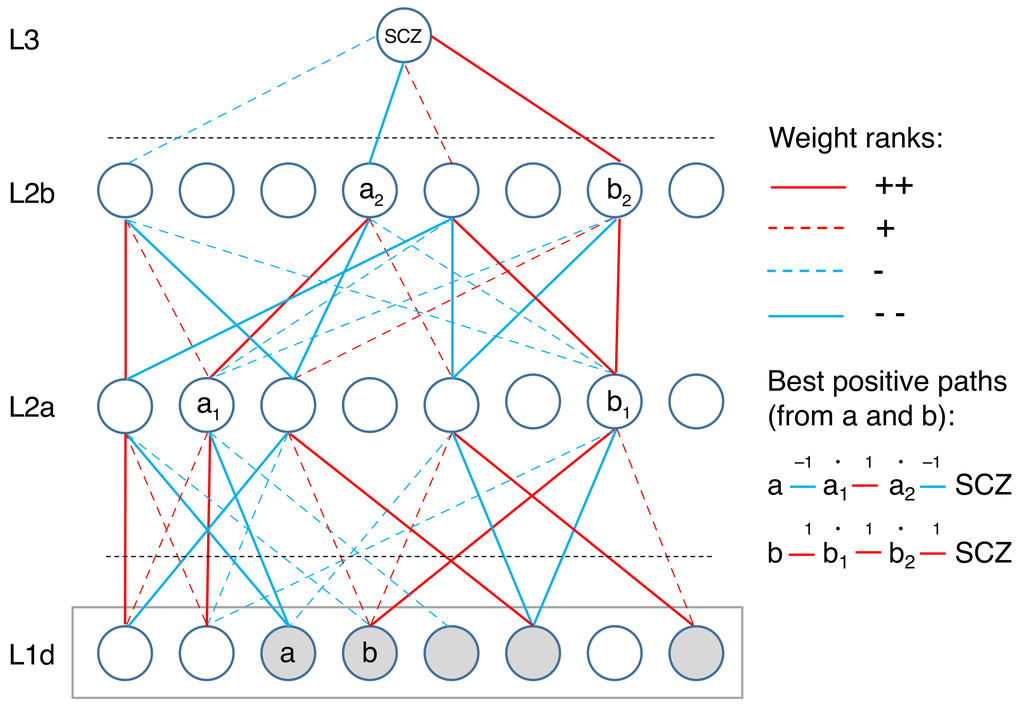
Method	SCZ	BPD	ASD	AVG (SCZ+BPD+ASD)	GEN	ETH	AGE
LR-gene	54.6% (0.5%)	56.7% (2.5%)	50.0% (0.0%)	53.8% (1.0%)	50.0%	99.0%	61.9%(AOD)
LR-trans	63.0% (4.8%)	63.3% (6.3%)	51.7% (1.8%)	59.3% (4.3%)	69.7%	86.0%	81.2%
cRBM	70.0% (31.0%)	71.1% (22.6%)	63.3% (10.8%)	68.1% (21.5%)	71.5%	89.0%	83.1%
DSPN-impute	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%



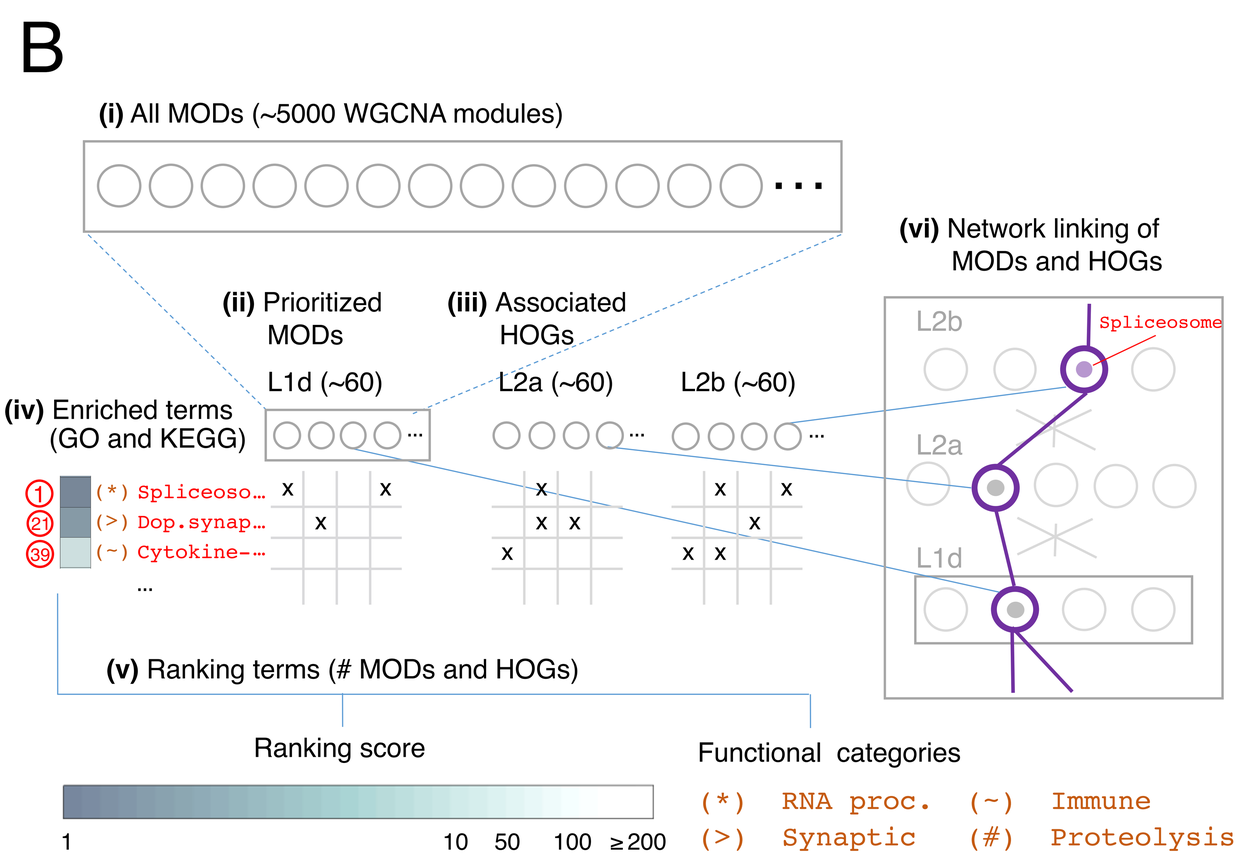
Model complexity	increasing	increasing	constant	increasing
Predictors	genotype	transcriptome	genotype->transcriptome	genotype->transcriptome

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

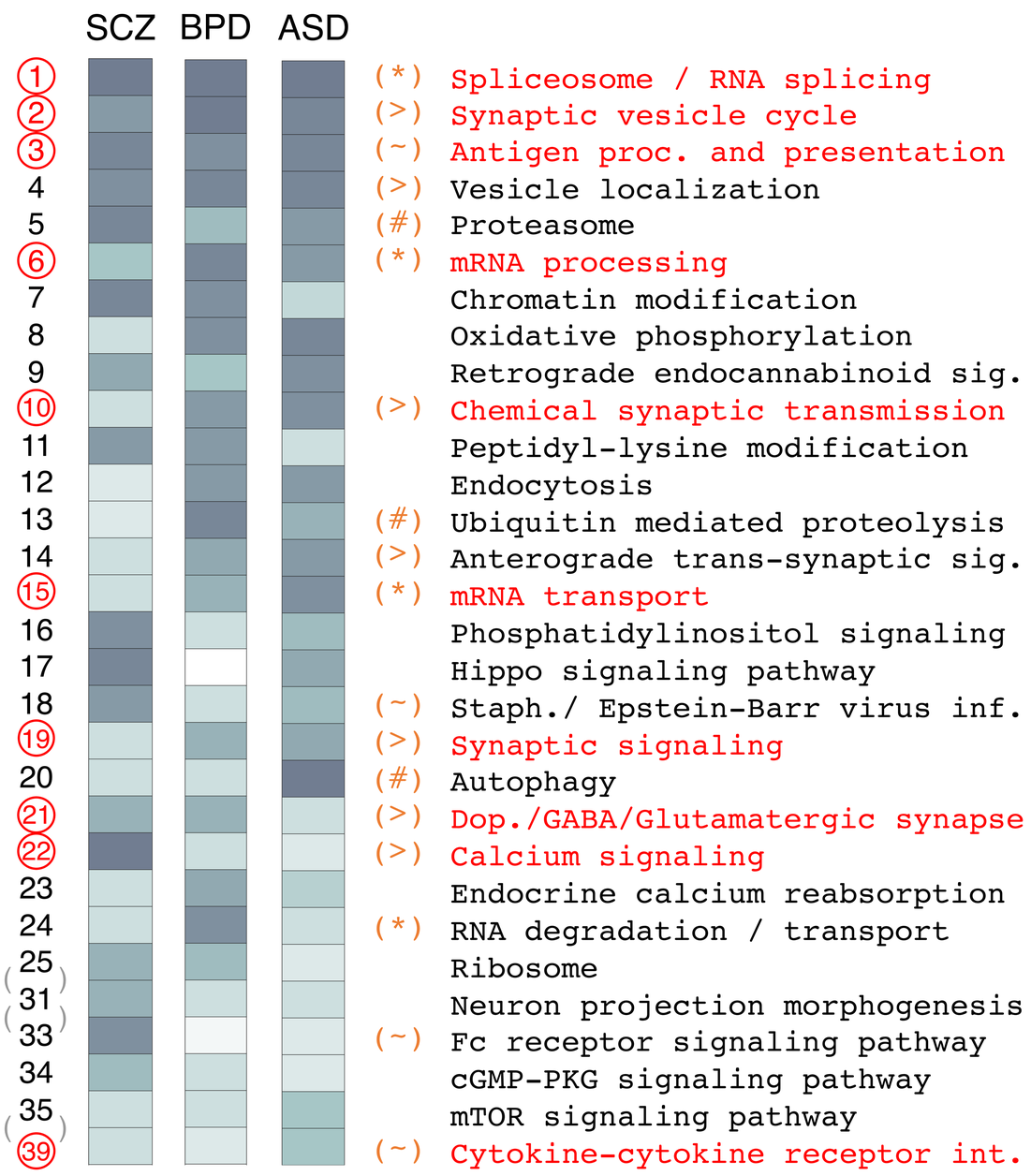
A


L3
L2b
L2a
L1d
SCZ
a2
b2
a1
b1
a
b
Weight ranks:
++
+
-
- -
Best positive paths (from a and b):
a −1 a1 1 a2 −1 SCZ
b 1 b1 1 b2 1 SCZ

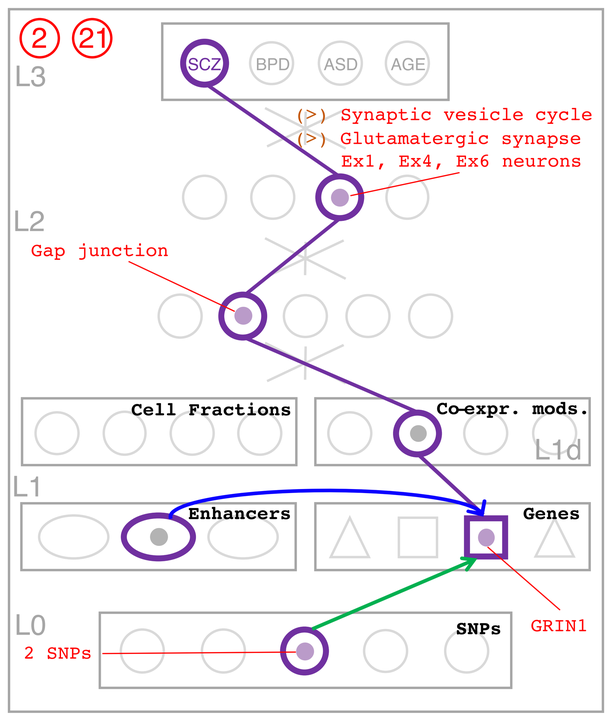
B

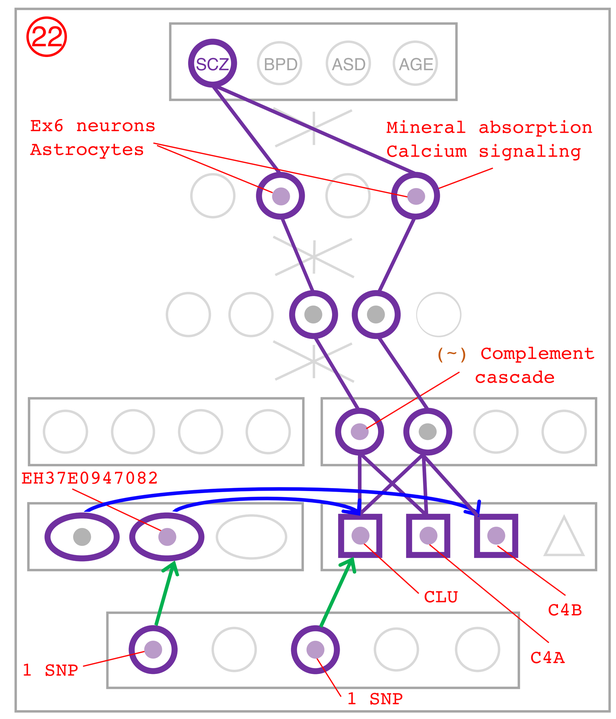

(i) All MODs (~5000 WGCNA modules)
(ii) Prioritized MODs
(iii) Associated HOGs
L1d (~60)
L2a (~60)
L2b (~60)
(iv) Enriched terms (GO and KEGG)
1 (*) Spliceoso…
21 (>) Dop.synap…
39 (~) Cytokine-…
(v) Ranking terms (# MODs and HOGs)
Ranking score
1
10 50 100 ≥200
(vi) Network linking of MODs and HOGs
L2b
L2a
L1d
Spliceosome
Functional categories
(*) RNA proc. (~) Immune
(>) Synaptic (#) Proteolysis

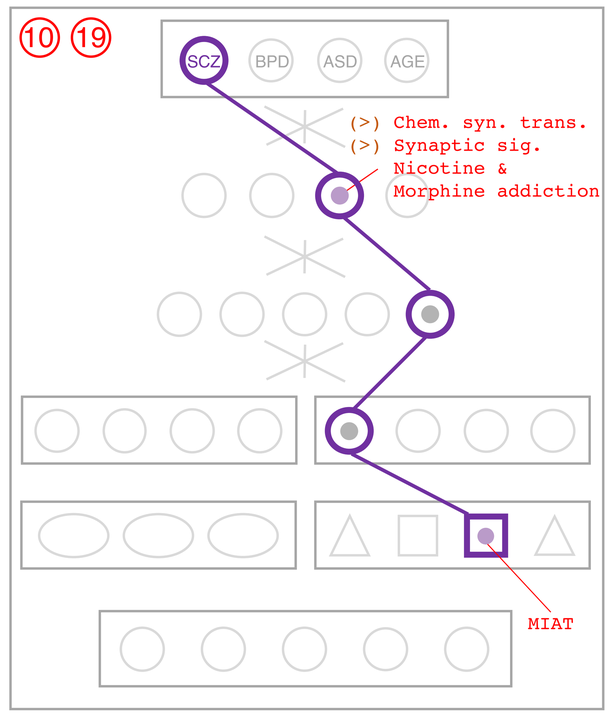
C

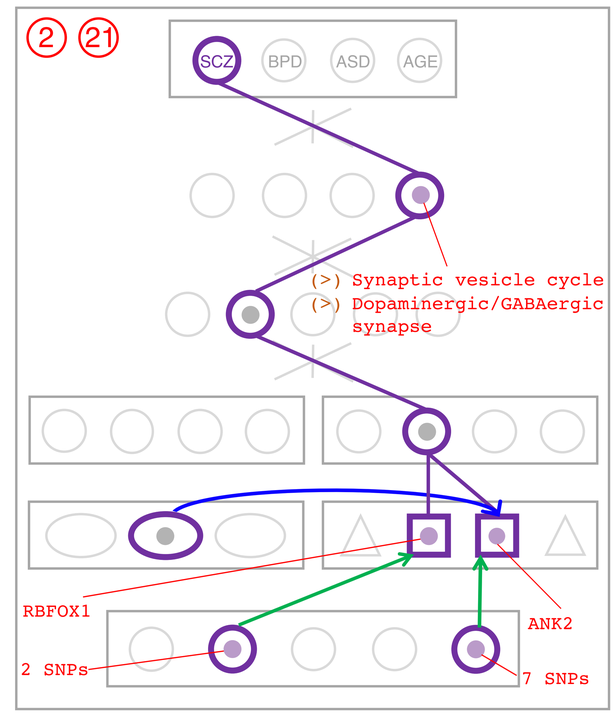

SCZ BPD ASD
1 (*) Spliceosome / RNA splicing
2 (>) Synaptic vesicle cycle
3 (~) Antigen proc. and presentation
4 (>) Vesicle localization
5 (#) Proteasome
6 (*) mRNA processing
7 Chromatin modification
8 Oxidative phosphorylation
9 Retrograde endocannabinoid sig.
10 (>) Chemical synaptic transmission
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.

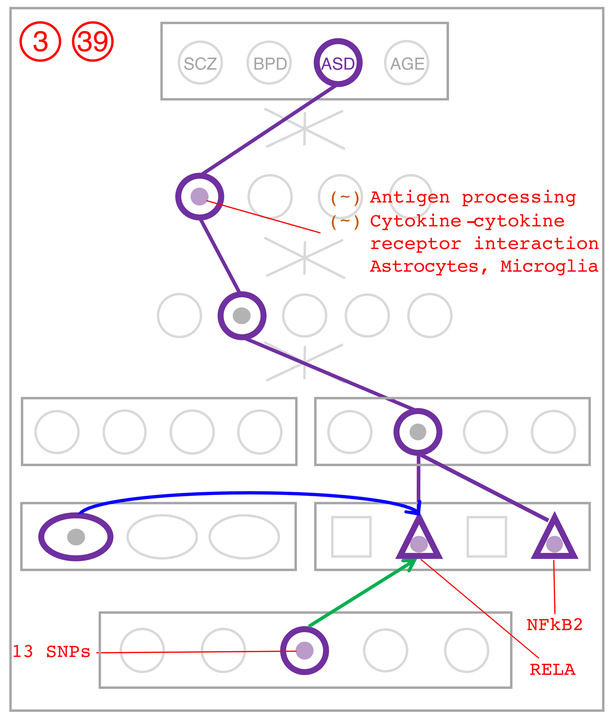
D

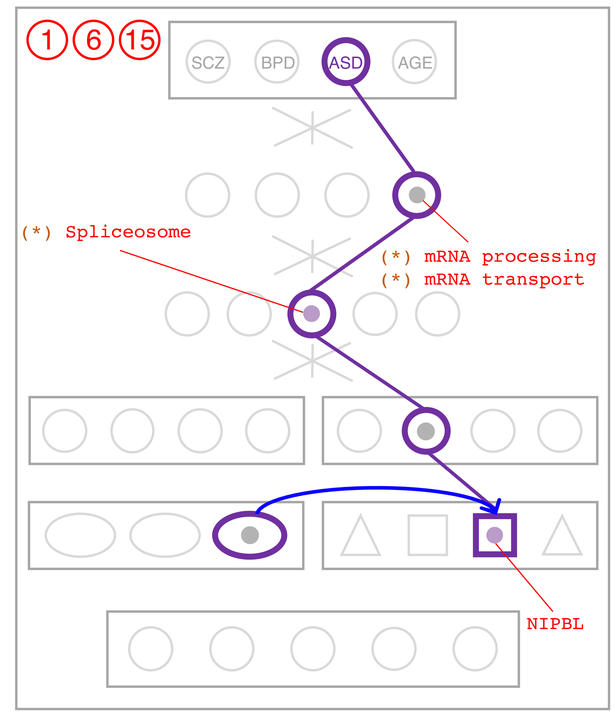

2 21
L3
L2
L1
L0
L1d
SCZ BPD ASD AGE
(>) Synaptic vesicle cycle
(>) Glutamatergic synapse
Ex1, Ex4, Ex6 neurons
Gap junction
Cell Fractions
Co-expr. mods.
Enhancers
Genes
SNPs
GRIN1
2 SNPs

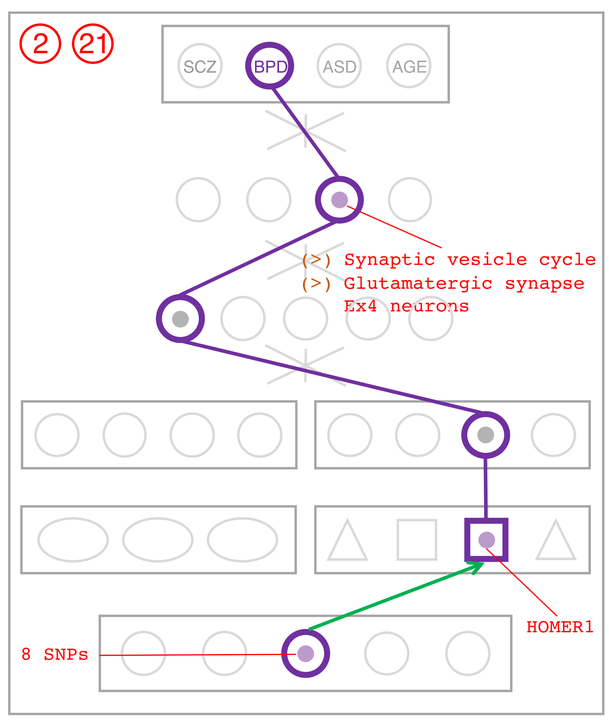

22
SCZ BPD ASD AGE
Ex6 neurons
Astrocytes
Mineral absorption
Calcium signaling
(~) Complement cascade
EH37E0947082
CLU
C4B
C4A
1 SNP
1 SNP

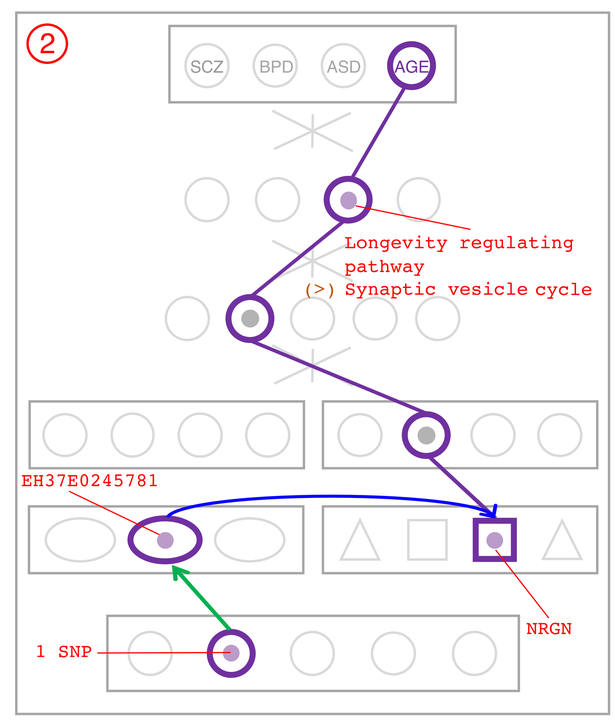

10 19
SCZ BPD ASD AGE
(>) Chem. syn. trans.
(>) Synaptic sig.
Nicotine & Morphine addiction
MIAT


2 21
SCZ BPD ASD AGE
(>) Synaptic vesicle cycle
(>) Dopaminergic/GABAergic synapse
RBFOX1
ANK2
2 SNPs
7 SNPs


3 39
SCZ BPD ASD AGE
(~) Antigen processing
(~) Cytokine-cytokine receptor interaction
Astrocytes, Microglia
NFkB2
RELA
13 SNPs


1 6 15
SCZ BPD ASD AGE
(*) Spliceosome
(*) mRNA processing
(*) mRNA transport
NIPBL


2 21
SCZ BPD ASD AGE
(>) Synaptic vesicle cycle
(>) Glutamatergic synapse
Ex4 neurons
HOMER1
8 SNPs


2
SCZ BPD ASD AGE
Longevity regulating pathway
(>) Synaptic vesicle cycle
EH37E0245781
NRGN
1 SNP

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 Navarro*, Declan Clarke*, Mengting Gu*, Prashant Emani*, Yucheng T. Yang, Min Xu, Michael 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 Hoffman, PsychENCODE Consortium‡, Panos Roussos, Schahram Akbarian, Gregory Crawford, Andrew E. Jaffe, Kevin White, Zhiping Weng, Nenad Sestan, Daniel H. Geschwind†, James A. Knowles†, Mark 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: