**Comprehensive functional genomic resource and integrative model for the adult brain (4481 + fig captions ~1400)**

**Abstract (225) (Ready for MG)**

Understanding how genomic variants influence brain phenotypes and disorders remains a challenge. To this end, the PsychENCODE consortium has generated large-scale sequencing datasets over 1866 adult brains with various phenotypes and diseases (e.g., schizophrenia, bipolar, autism), including genotype, transcriptomics, chromatin, and single-cell. By uniformly processing and consistently analyzing these datasets, we developed a comprehensive resource on functional genomics of the adult brain. We deconvolved the bulk tissue expression using single-cell data and found that the cell fraction changes can explain >85% tissue covariance and associate with psychiatric phenotypes. Moreover, we found that the brain differs from other tissues most notably in gene expression compared to chromatin activities. We used the chromatin and Hi-C data from reference brain samples to generate ~80,000 active enhancers in the adult prefrontal cortex, and linked them to genes and transcription factors. We also identified approximately two million eQTLs and xx additional with chromatin, splicing and cell fractions. In addition, we built a full gene regulatory network linking all possible genomic elements, and leveraged it to discover 112 additional genes associated with GWAS variants with psychiatric disorders. Finally, we developed a deep-learning model embedding the regulatory network to connect genotypes and phenotypes on multiple layers, achieving ~4.6X improvement in disease-trait prediction above an additive genetic model. This model enables highlighting intermediate genes and functional modules and imputing missing transcriptome and epigenome from genotype data only.

**Introduction (460) (Ready for MG)**

Disorders of the brain affect nearly one fifth of the world’s population \cite{19507169}. Decades of research has led to little progress in our fundamental understanding of the molecular causes of psychiatric disorders. This contrasts with cardiac disease, for which lifestyle and pharmacological modification of environmental risk factors has had a profound effect on morbidity, or cancer, which is now understood to be a direct disorder of the genome \cite{9603539, 24071849, 16461820, 19360079}. Although genome-wide association studies (GWAS) have identified many genomic variants associated with psychiatric disease risk, for the vast majority we have little understanding of how molecular mechanisms affect the brain, thereby altering risk \cite{19339359}.

To this end, a number of studies have begun to elucidate the molecular steps on the path from genomic alteration to risk. For instance, the Psychiatric Genomics Consortium (PGC) has identified 142 GWAS loci associated with schizophrenia \cite{26404826}. Many of these are in non-coding regions \cite{26404826}, suggestive of roles in gene regulation. Other consortia have annotated these non-coding regions using expression quantitative trait loci (eQTLs) from the Genotype-Tissue Expression (GTEx) project and enhancers in the Encyclopedia of DNA Elements (ENCODE) and Epigenomics Roadmap projects. However, none of these projects have specifically tailored their efforts toward comprehensively identifying the functional elements in the brain. The initial projects to focus on identifying brain-specific genomic elements have provided greater insight into brain-specific functional genomics \cite{27668389, 22031444}, but could be enhanced with larger sample sizes from both healthy and diseased samples. Moreover, many new assays for functional elements have been recently developed, such as Hi-C and single-cell sequencing, which have yet to be integrated with brain genomics data, at scale \cite{27760116,27339989, 26060301}.

Hence, the PsychENCODE Consortium has designed, generated and assembled a large-scale dataset for providing insights into the adult human brain and psychiatric disorders, including data derived through genotyping, bulk and single-cell RNA-seq, ChIP-seq, ATAC-seq, and Hi-C using brains from thousands of individuals \cite{26605881}. All raw and uniformly processed data at both tissue and single-cell level have been placed into a central, publically available resource for adult brain functional genomics, that also integrates relevant re-processed data from the other related projects, including ENCODE, CommonMind (CMC), GTEx, Epigenomics Roadmap, with nearly ~12,000 data samples in total. By leveraging this resource, we were able to identify functional elements and QTLs specific to the adult brain, including novel psychiatric GWAS and gene linkages. Moreover, we combined these elements to build an integrated deep-learning model. This tool can utilize the richly structured data of the resource to identify interactions between genotype and molecular phenotypes at multiple layers, as well as predict high-level traits.

**Resource construction (196) (Ready for MG)**

We designed the resource Adult.PsychENCODE.org to provide a coherent structure to a large amount of data on functional genomics of the brain \cite{Supplement}. Broadly, it organizes data hierarchically, with a large base of raw data files (many of which have restricted access, such as individual genotyping and raw next-generation sequencing of transcriptomics and epigenomics), a middle layer of uniformly processed and easily shareable results (such as open-chromatin regions and gene-expression quantifications), and a compact cap that consists of an integrative model based on imputed regulatory networks and QTLs. As shown in Fig. 1, to build the base layer we included all the adult data from PsychENCODE (~5,500 datasets derived from 1,866 individual brains) and merged these with relevant data from ENCODE, CMC, GTEx, Roadmap, and recent single-cell studies (~5,000 additional datasets) \cite{27339989, 26060301} . These data cover a representation of phenotypes and psychiatric disorders (schizophrenia, bipolar, autism). Furthermore, the PsychENCODE project developed a specific "reference brain" project on adult prefrontal cortex (PFC) utilizing many matched assays on the same set of brain tissues, which we used (below) to develop an anchoring annotation for the entire resource \cite{Supplement}.

**Transcriptome analysis: Bulk & single-cell (660) (Ready for MG)**

To identify the genomic elements exhibiting transcriptional activities specific to the adult brain, we used the ENCODE pipeline to uniformly process RNA-seq data of all available samples from PsychENCODE, GTEx and Roadmap. Using these data, we identified a wide variety of interpretable brain functional elements, such as non-coding regions of transcription, and sets of differentially expressed and co-expressed genes as part of resource - e.g., 12,080 genes were transcribed in the brains 95% of the individuals surveyed and over 16,000 protein-coding and 9000 non-coding genes were detected in total \cite{cap1, Supplement}.

e.g., 12,080 genes transcribed in the brains 95% of the individuals surveyed \cite{cap1, Supplement}. [[DG:We detected 16,541 protein-coding and 9,233 non-coding genes detected in 1695 unique subjects; 12,080 genes were transcribed in the brains 95% of the individuals surveyed. Details of the transcriptomic analyses including DE and network analyses are presented in \cite{cap1, Supplement}. ]]

e.g., 12,080 genes transcribed in the brains 95% of the individuals surveyed \cite{cap1, Supplement} and over 16,000 protein-coding and 9000 non-coding genes detected in total \cite{cap1, Supplement}. ]]

Brain tissues are composed of a variety of cell types, including neuronal and non-neuronal cells. Previous studies have suggested that gene-expression changes at the tissue level may be associated with changing proportions of basic cell types \cite{21614001, 29439242,18849986, 27409810}. However, studies have not systematically revealed how different cell types can quantitatively contribute to population-level expression variation. Here, we addressed whether expression changes over our cohort of 1,866 adults were driven by expression changes in a particular cell type or whether they resulted from changing proportions of various cell types.

We used two complementary strategies. First, we used the standard pipeline to uniformly process single-cell RNA-seq data in PsychENCODE, in conjunction with a number of other single-cell studies on the brain, in order to assemble a list of brain cell types for the project. This includes previously identified neuronal types, major non-neuronal types, and a number of additional cell types involved in brain development \cite{Supplement}. The results constitute a matrix (C) of expression signatures, mostly concordant with what has been published (Fig. S5 and Discussion). A number of genes had expression levels varying more substantially across these cell types than they did across individuals in a population (e.g., the dopamine receptor DRD3, Fig. 2A). This implies that the changes in bulk expression can readily result from cell fraction variations.

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

We then tried to understand how variation in proportions of cell types contributes to variation in bulk expression. In particular, we de-convolved the expression matrix of tissue (B) using the single-cell signatures (C) to estimate the cell-fractions W, solving the equation B≈CW \cite{Supplement} (Fig. 2B). As validation, our estimated fractions of NEU+/- cells matched the experimentally determined fractions from the reference brain samples (Median error = 0.04, Fig. S6) and compared our results with previous deconvolution methods \cite{Supplement}. Overall, we found that single-cell expression signatures could explain much of the population-level expression variation (Fig. 2D, i.e., across tissue samples from different individuals 1-||B-CW||2/||B||2>85%) \cite{Supplement}.

Finally, we found that cell-fraction changes were associated with different observed phenotypes and disorders (Fig. 2E, SXXX). For example, particular excitatory and inhibitory neurons exhibited different fractions between male and female samples (i.e., Ex3 and In6). The fraction of Ex3 was also reduced in autism spectrum disorder (p=0.0077), where non-neuronal cells (e.g., oligodendrocytes) were represented in greater abundance. Another interesting association was with age. In particular, the fractions of neuronal types Ex3 and Ex4 significantly increased with age; by contrast, some non-neuronal types (e.g. oligodendrocytes) decreased. These changes are potentially associated with differentially expressed genes. For example, a gene involved in the early growth response was down-regulated in older groups, whereas ceruloplasmin was down-regulated in the middle-aged (Fig. 2F).

**Enhancers (440) (Ready for MG)**

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

Assessing the variability of enhancers across individuals and tissues is more difficult than performing the analogous comparison for gene expression. Not only does the chromatin signal change across the population, but the boundaries of enhancers grow and shrink and sometimes disappear altogether (Fig. 3A). To investigate chromatin variability across the population, we uniformly processed the H3K27ac data from PFC, temporal cortex (TC), and cerebellum (CBC) on a cohort of 50 individuals \cite{Supplement}. Aggregating ChIP-seq data across the population resulted in a total of 37,761 H3K27ac "peaks" (enriched regions) in PFC, 42,683 in TC, and 26,631 in CB -- each of them present in more than half of the cohort. Comparing aggregate sets for these three brain regions, the PFC was more similar to TC than CB (~90% vs 34% overlap in H3K27ac peaks), consistent with previous reports \cite{27863250}.

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

**Consistent comparison: transcriptome & epigenome (270) (Ready for MG)**

As we uniformly processed the transcriptomic and epigenomic data across PsychENCODE, ENCODE, GTEx, and Roadmap datasets, we could compare the brain to other organs in a consistent fashion and also to compare across transcriptome and epigenome. We tried several methods, including PCA, t-SNE, and reference component analysis (RCA) for an appropriate comparison. Although popular, PCA tends ignore most of the local structure and can be easily influenced by outliers. By contrast, t-SNE analysis preserves local relationships but “shatters” global structure \cite{Supplement}. We selected RCA as a compromise. RCA projects gene expression in an individual sample against a reference panel, and then essentially reduces dimensionality of each individual projection.

For gene expression, our comparison revealed that the brain separates from the other tissues in the first component (Fig. 3E). Also, inter-tissue differences were larger than intra-tissue ones (Fig. SXXX). A different picture emerged for chromatin: comparison showed that the chromatin levels at all regulatory positions were, overall, less distinguishable between brain and other tissues (Fig. 3C) \cite{Supplement}. At first glance, this is surprising as one expects great differences in epigenetics between tissues. Yet our analysis compares chromatin signals over all ~1.3M non-coding regulatory elements (including enhancers and promoters), which is consistent with a comparison across all protein-coding genes (Fig 3F). The total number of regulatory elements is much larger than the the brain-active enhancers (~79K), so there are proportionately far fewer brain-active regulatory elements than the brain-active genes (6% vs. 21%).

This analysis focused on inter-tissue differences in well-annotated regions (i.e., genes, promoters, and enhancers). In addition to the expression differences in protein-coding genes, we also found transcriptional diversity across tissues in unannotated non-coding and intergenic regions. In particular, testes and lung tend to have the largest transcriptional diversity in terms of the degree of transcription of protein-coding genes (Fig. 3D); however, when we shifted to unannotated, intergenic regions, we found brain tissues (such as cortex and cerebellum) stood out by exhibiting a greater extent of transcription than other tissues.

**QTL analysis (408)(Ready for MG)**

We used the PsychENCODE data to identify QTLs affecting gene expression and chromatin activity. In particular, we calculated expression, chromatin, splicing-isoform, and cell-fraction QTLs (eQTLs, cQTLs, isoQTLs and fQTLs, respectively). For eQTLs, we adopted a standard approach, adhering to the established GTEX pipeline. We identified 2,542,908 cis-eQTLs (2,097,741 after linkage-disequilibrium pruning) and 32,944 eGenes (including non-coding ones) in PFC. We found 1,341,182 (25%) unique SNPs involved in cis-eQTLs from 5,297,875 SNPs in 1 Mb cis window of genes. This conservative estimate has a substantially larger number of eQTLs and eGenes than previous eQTL studies and reflects the large PsychENCODE sample size and statistical power \cite{Supplement}. We think the eQTL number is close to saturation, in terms of associating almost every variant with an expression-modulating characteristic. We also applied the same QTL calculation pipeline to splicing and identified 157,592 isoQTLs with 6480 unique genes which have at least one isoQTL. Finally, we also identified 1,798,602 trans-eQTLs involving 659,952 unique SNPs.

For cQTLs, the situation is more complicated: no established methods exist for calculating these on a large scale, although there have been a variety of previous efforts \cite{25799442, 26300125}. To properly identify cQTLs, we focused on our reference set of enhancers and then examined how H3K27ac chromatin activity varied in these regions across 292 individuals \cite{supplement}. Overall, we identified ~2,000 cQTLs in addition to the 6,200 identified using individuals from the CMC cohort \cite{<https://doi.org/10.1101/141986>}.

Next, we determined if any SNPs were associated with changes in the relative fractions of cell types across individuals (fQTLs). In total, we identified 3,720 distinct SNPs constituting 4,186 different fQTLs. Of these, the proportions of microglial cells and excitatory neuron Ex8 were associated with the most variants. Next, after factoring out these cell type differences, we identified 260,280 SNPs significantly associated with gene expression changes across individual tissues; these "residual trans-eQTLs" represent variant-expression associations largely unexplained by changing proportions of cell types.

To further dissect the associations between genomic elements and the QTLs, we intersected our QTL lists with each other and a comprehensive set of genomic annotations (Fig. 4D). As expected, eQTLs tended to be enriched at promoter regions, and cQTLs, at enhancer and TFBS regions; fQTLs tended to occur in many different elements. For the overlap among different QTLs, we expected that most cQTLs, isoQTLs and fQTLs would be a subset of the much larger number of eQTLs; somewhat surprisingly, an appreciable number of these did not overlap (Fig. 4C).

**Regulatory networks (390) (Ready for MG)**

We next integrated the genomic elements described above at the regulatory-network level. We created a putative network revealing how the genotype and regulators relate to target gene expression in the brain. 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) \cite{27760116, Supplement}. In total, we identified 2,735 topologically associating domains (TADs) and 149,098 enhancer-promoter interactions. This Hi-C dataset substantially differed from an earlier fetal brain Hi-C dataset (e.g. only 31% of the enhancer-promoter interactions in the adult were detected in the fetal dataset) \cite{27760116}, highlighting the importance of the developmental stage for chromatin topology {Fig. SXXX}.

As expected, ~75% of enhancer-promoter interactions occurred in the same TAD, and genes with more associated enhancers tended to have higher expression (Fig. 5B). We next integrated the Hi-C dataset with QTLs to assess how much of the common variation-associated gene regulation is corroborated by chromatin interactions. Surprisingly, QTL SNPs on enhancers supported by Hi-C evidence showed stronger associations than ones on exons and promoters (Fig. 5C).

In addition, we tried to predict further regulatory relationships (beyond those from Hi-C and QTLs) based on directly relating the activity of transcription factors (TFs) to target genes. In particular, for each potential target of a TF, we demanded that it have good binding motifs in open chromatin regions (either in promoters or brain-active enhancers) and that it have a high "coefficient" in a regularized, elastic-network regression relating TF activity to target expression \cite{Supplement}. Overall, we found the subset of interactions that met these two criteria could predict the expression of 8930 genes with minimal mean-square error (MSE) < 0.05 (Fig. Sxx). For example, we could predict the expression of the ASD risk gene CHD8 with an MSE<0.034. Moreover, the further subset of these interaction that involved the TF interacting with a motif on a nearby enhancer to a gene, necessarily instantiated a third set of enhancer-to-gene links.

Collectively, we generated a full regulatory network, linking enhancers, TFs, and target genes, based on the union of (i) activity relationships, (ii) physical chromatin interactions, and (iii) QTLs \cite{Supplement}. This network has a scale-free and hierarchical structure (Fig. Sxx), and consists of ~150,000 Hi-C enhancer-promoter interactions, ~2 million QTL-based linkages (QTL-SNP-to-eGene), and ~211k TF-to-target and ~577k enhancer-to-target-promoter linkages based on activity relationships.

**Linking GWAS variants to genes (266) (Ready for MG)**

We used our above regulatory network to connect non-coding GWAS loci to potential genes. We exploited all possible linkages from including Hi-C, QTL, and activity-correlation relationships. For the newly identified 142 schizophrenia GWAS loci \cite{29483656}, we identified a set of 1,054 putative schizophrenia-associated genes, covering 119 loci (hereby referred as "SCZ-genes"). 302 of these constituted a high-confidence set supported by more than two sources of evidence (i.e., QTL and Hi-C, Fig. 5D, Fig. Sxx). These SCZ-genes identified from multi-sources represent a substantial increase from the previously reported 22 genes across 19 loci based on a smaller QTL set\cite{29483656, 27668389}. The majority of SCZ-genes were not in linkage disequilibrium with index SNPs (698 genes [~66%] with r2>0.6, Fig. Sxx), consistent with previous observations that regulatory relationships often do not follow linear genome organization \cite{27760116}.

We then looked at the characteristics of the SCZ-genes. As expected, these shared many characteristics with known schizophrenia-associated genes. In particular, they were enriched in genes intolerant to loss-of-function mutations \cite{29483656}, translational regulators, cholinergic receptors, calcium channels, synaptic genes, and genes that are known to be differentially expressed in schizophrenia (Fig. Sxx). Next, we integrated SCZ-genes with single-cell profiles, where we found that they are highly expressed in neurons with the highest expression in excitatory neurons (Fig. 5G).

Finally, in a more general context, we found significant aggregate associations between our eQTLs and all brain-disease GWAS variants, not just schizophrenia. In particular, we found overlap enrichments of our cis-eQTL SNPs and GWAS SNPs for Schizophrenia, Bipolar disorder, and Parkinson’s disease to be more significant than that for GWAS SNPs for non-brain related disorders (Fig. 4E).

**Integrative deep-learning model (848) (Ready for MG)**

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

Using the full model with the genome and transcriptome data provided, we demonstrated that the extra layers of structure in the DSPN allowed us to achieve substantially better prediction of disease and other observed traits than traditional genotype-to-phenotype models, and that the transcriptome carries additional trait information, which the DSPN is able to extract (Fig. 6D). For instance, a logistic predictor was able to gain a 2.4X improvement when using the transcriptome vs. the genome alone (+9.3% for transcriptome vs. +3.8% for the genome, above 50% random baseline). In comparison, the DSPN was able to gain a 4.6X improvement (+17.4% vs. +3.8%), which may reflect the fact that the DSPN is able to incorporate non-linear interactions between intermediate phenotypes. Moreover, the DSPN also allows us to perform joint inference and imputation of intermediate phenotypes (i.e., transcriptome and epigenome) and observed traits from just genotype alone, achieving a ~2.7X improvement over a logistic predictor (Fig. 6D). These results demonstrate the usefulness of even a limited amount of functional genomic information for unraveling gene-disease relationships and show that the structure learned from such data can be used to make more accurate predictions of observed traits even when absent.

We transformed our results above to the liability scale in order to compare with narrow-sense heritability estimates (Fig. 6D) \cite{Supplement}. Prior studies have estimated that common SNPs explain 25.6%, 20.5%, and 19% of the genetic variance for schizophrenia, bipolar disorder, and autism, respectively \cite{bioRxiv\_10.1101/048991}. These may be taken as limits to which an additive predictive model using the common SNPs will approach given unlimited training data; by contrast, a non-linear predictor can potentially exceed these limits. Our best liability scores based on genotype are substantially below these levels, implying that incorporating further data will be beneficial. The variance explained by the full DSPN model was of a similar order to that explained by common SNPs for all three conditions (16.3%, 30%, and 14.4%). Improved imputation of intermediate phenotypes may thus capture most of the variance due to common-SNP, narrow-sense heritability, although this is limited by the proportion of variance in the imputed variables which is genetically determined (Fig. SA1).

A key aspect of the DSPN is that its structure is interpretable. In particular, we examined the specific connections learned by the DSPN between intermediate and high-level phenotypes for potentially relevant biological associations. We included known co-expression modules in the model and examined which of these the DSPN prioritized, as well as new sets of genes associated with DSPN latent nodes that were uncovered at each hidden layer (Supplement and Fig. SA2). We provide a full summary of the functional enrichment analysis for the prioritized modules and highlight some of the associations found using the schizophrenia model (Fig. 6c) \cite{Supplement}. Overall, we showed that the modules prioritized by the DSPN were strongly enriched for known GWAS variants (Supplement). In particular, among the highest schizophrenia-prioritized modules and higher-order groupings, we found enrichments for (i) glutamatergic-synapse pathway genes, (ii) calcium-signaling pathways and astrocyte-marker genes, and (iii) complement cascade pathway genes including C4A, C4B, and CLU -- confirming and extending previous analyses \cite{26814963}. Further, we found enrichment in the prioritized modules/groupings for aging, the marker genes for Ex4 cell types, and the gene NRGN (in a module associated with synaptic and longevity functions), consistent with our differential expression analysis (Fig. S XXXX).

**Discussion (318) (Ready for MG)**

Here, we uniformly integrated PsychENCODE datasets with other datasets, developing a comprehensive resource for functional genomic elements of the adult brain. [[DG: The integrative analyses of genetic association data here and with respect to the transcriptome in (Ref capstone 1), demonstrate that functional annotation of gene regulatory elements is necessary to understand the molecular mechanisms of disease related genetic variation.]]

Overall, our study identified a set of eQTLs several fold greater than previous studies, achieving close to saturation for protein-coding genes. Our data are consistent with the stage and tissue specific nature of gene regulation, indicating that it will be valuable to profile different regions and developmental stages at similar scale. This also includes increasing individual sample size and quality of chromatin data, such as identifying enhancers via STARR-seq, will help with cQTLs. More fundamentally, one-dimensional fluctuations in chromatin signal reflect changes in three-dimensional changes in architecture and new metrics beyond cQTLs may need to be developed to measure this better. In addition, some other epigenetic marks might exhibit distinguishable patterns in the brain, e.g. the brain methylation landscape compared to the other tissues. Likewise, the inter-tissue differences might be further amplified when comparing the transcription level of microRNAs.

Another area for future development is single-cell analysis. In particular, it will be very important to consider the cell heterogeneity in more detail [[DG: remove]]. In this study, we found that a set of basic and known cells with very different expression signatures and showed that they could explain large expression variations across tissues. This assumes that expression signatures, at least for biomarker genes, are fairly constant over same cell types. Larger-scale single cell studies will allow us to examine this assumption in more detail, perhaps quantifying environment-associated transcriptional variability of certain cell types. In addition, current techniques suffer from low capture efficiency; therefore, it remains challenging to reliably quantify low-abundant transcripts and genes \cite{26949524, 25053837}, with particular challenges for specific cell parts such as axons and dendrites \cite{25053837}. Thus, we may see novel techniques such as single-nucleus RNA-seq techniques that emerge to address these problems.

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

**Figures (1435)**

**Figure 1. Comprehensive data resource of functional genomics in adult brain**. The functional genomics data generated by the PsychENCODE consortium (PEC) constitute a multidimensional exploration across tissue, developmental stage, disorder, species, assay, and sex. From this larger corpus of PEC samples, we focused on adult datasets, integrated with those from consortia such as GTEx, the Roadmap Epigenomics Consortium, ENCODE, CMC and Human Brain Collection Core studies, and previously published single-cell transcriptomic data. The central data cube represents the results of this integration for the three dimensions of disorder, assay, and tissue, where only the numbers of datasets used in the current analysis are depicted. Projections of the data onto each of these three parameters are shown in graph form for assay and disorder, and in schematic form for the primary brain regions of interest. **Assay:** The bars represent datasets across a subset of the assay types, including RNA-seq (N = 2040 PEC + 1632 uniformly processed GTEx samples), genotypes (N = 1753 PEC + 113 GTEx = 1866 individuals), scRNA-seq (N = 932 PEC + 3693 external datasets), and H3K27ac ChIP-seq (= 408 PEC + 5 uniformly processed Roadmap samples). **Disorder:** The number of individuals under the control category include the 113 from GTEx and 926 from PEC, while individuals from PEC provide data on the remaining disorders of schizophrenia (SCZ, N = 558), bipolar disorder (BPD, N = 217), ASD (N = 44), and affective disorder (AFF, N = 8), resulting in a total of 1,866. **Tissue:** In this schematic, we focus on the datasets derived from three primary brain regions evaluated in our integrative study: the PFC (N = 3521), the ex TC (N = 2153), and the CB (N = 348).

**Figure 2. Deconvolution analysis of bulk and single-cell transcriptomics reveals cell fraction changes across tissue phenotypes and disorders**. **(A)** Genes had significantly higher expression variability across single cells than tissue samples. Left: dopamine gene, DRD3. **(B)** Top: the bulk tissue gene expression matrix (B, genes by individuals) can be decomposed by NMF to the product of two matrices: NMF component matrix (V, genes by top NMF components; i.e., NMF-TCs) and component fraction matrix (H, top NMF components by individuals); i.e., B≈VH. Bottom: the bulk tissue gene expression matrix B can be also deconvolved by the single-cell gene expression matrix (C, genes by cell types) to estimate the cell fractions across individuals (the matrix, W); i.e., B≈CW. Three major cell types were neuronal cells (blue), non-neuronal cells (red), and developmental (dev) cells (green), as highlighted by columns groups in C (also row groups in W). **(C)** The heatmap shows the Pearson correlation coefficients of gene expression between the NMF TCs and single cell types for the biomarker genes (N=457). For example, NMF-7 highly correlated with the Ex3 cell type (r=0.6597). **(D)** The estimate cell fractions contributed to >85% bulk tissue expression variations; i.e., 1-||B-C\*W||/||B||>0.85. **(E)** The cell fractions changed across brain phenotypes and disorders. The neuronal cell types (e.g., Ex3 and In6) had significantly higher fractions in male than female samples (p=0.0036, and p=6.74e-5). Disorder types showing significant changes compared to control samples are labeled (\*\*). For example, neuronal cells and oligodendrocytes had lower fractions in ASD than other cell types. **(F)** The cell fractions and gene expression (EGR1) changed across ages. The excitatory neuronal cell type Ex3 had a significant increase increase with age (trend analysis p<6.3e-10).

**Figure 3. Comparative analysis for transcriptomics and epigenomics between brain and other tissues**. **(A)** Chromatin features of the reference brain were used to identify active enhancers. Enhancers were located in the open chromatin region (high ATAC-seq signal), with strong H3K27ac signal and a lack of H3K4me3 signal. Enhancer activity varied among individuals, as indicated by the varying H3K27ac peak signal in the cohort. Each row corresponds to an individual, with shallow to dark blue indicating low to high signal peak value. **(B)** The number of brain active enhancers achieved a saturation of over 70,000 with 20 samples. **(C)** The coefficients (PC1 vs. PC2) of RCA analysis for gene expression data of PsychENCODE samples are shown in dark green. Other external brain samples (light green) including GTEx and other tissue samples (magenta) are shown as well. **(D)** The center (cross) and ranges of different tissue clusters (dashed ellipse) are shown on an RCA scatterplot of (C). **(E)** The tissue clusters of RCA coefficients (PC1 vs. PC2) for chromatin data of active brain enhancers are shown. The brain cluster (green ellipse) includes the PsychENCODE samples and other tissue clusters (magenta ellipses) consisting of Epigeomics Roadmap samples. **(F)** The transcriptional diversity on coding (circle) and non-coding (triangle) regions among the tissue samples (inter-sample on x-axis) is shown compared to the diversity on cumulative tissue samples (y-axis) for select major tissue types including CB (lightblue), cortex (xxx), lung (xxx), ovary (xxx), skin (xx), and testes (xx), using PolyA RNA-seq data.

**Figure 4. Summary of QTLs of human adult brain PFC**. (**A**) Example of H3K27ac signal of individual brains in a representative genomic region showing largely congruent identification of regions of open chromatin. Region in the dashed frame represents a chromatin QTL; the signal magnitudes of individuals with a G/G or G/T genotype were lower than the ones with a T/T genotype. (**B**) Numbers of genes with at least one eQTL (eGenes) are shown compared to sample size in different studies. The number of eGenes increased as the sample size increased. The eGenes of PsychENCODE is close to saturation. (**C**) Overlap of eQTL, sQTL, fQTL, and cQTL SNPs is shown. In total, 80% of eQTLs overlapped with other QTLs. The percentage of overlapped SNPs of eQTL with other QTLs was the highest among all QTLs. A total of 31% of fQTL SNPs overlapped with other QTLs, which was the lowest among all QTLs; 36% of sQTL and cQTL SNPs overlapped with other QTL SNPs. fQTLs overlapped more with sQTLs (17%) than eQTLs (9%) (p<1e-15, odds ratio:2.07). (**D**) Enrichment of genomic regions annotations of QTLs is shown. (E) Enrichment of GWAS SNPs on cis-eQTL SNPS is shown. Enrichment for GWAS SNPs of brain disorders on cis-eQTLs SNPs were more significant than those of non-brain disorder GWAS SNPs. Schizophrenia GWAS SNPs had the highest enrichment on cis-eQTLs SNPs among the three brain disorders analyzed.

**Figure 5. Data integration and modeling predicts gene regulatory network, revealing additional GWAS genes for psychiatric disorders**. **(A)** The full Hi-C data from adult brain reveal the folding principle of the genome, ranging from contact maps (top), TADs, and enhancer-promoter interactions. We leveraged gene regulatory linkages involving TADs, TFs, enhancers, and target genes to a full gene regulatory network consisting of ~150,000 Hi-C enhancer-promoter interactions, ~2 million eQTL-eGene linkages, ~211k TF-to-target and ~577k enhancer-to-target-promoter linkages based on activity relationships. **(B)** We compared the number of genes (left y-axis, dotted line) and the normalized gene expression levels (right y-axis, boxes) with the number of enhancers that interact with the gene promoters. **(C)** QTLs that were supported by Hi-C evidence showed more significant P-values than those that were not. **(D)** An overlap between putative target genes of schizophrenia GWAS loci defined by eQTL associations (eQTL), chromatin interactions (Hi-C), and activity relationships (Elastic net), including 302 highly confident and 1054 risk genes with schizophrenia. **(E)** A gene regulatory network of TFs (green), eQTLs (red), enhancers (blue), and 302 highly confident schizophrenia genes as targets is shown. **(F)** Evidence depicting that GWAS SNPs that overlap with CHRNA2 eQTLs also have chromatin interactions and activity correlations with the same gene. **(G)** SCZ genes show higher expression levels in neuronal cell types (excitatory neurons) than others.

**Figure 6. Deep-learning model predicts genotype-phenotype and reveals intermediate molecular mechanisms**. **(A)** The schematic outlines the model structures for Logistic Regression (LR), conditional Restricted Boltzmann Machine (cRBM), conditional Deep Boltzmann Machine (cDBM), and DSPN models. Nodes are partitioned into four possible layers (L0-L3) and colored according to their status as (i) conditioning nodes visible during training and testing (light blue); (ii) nodes visible during training and visible or imputed during testing (dark blue); or (iii) hidden nodes (grey). **(B)** The DSPN structure is shown in further detail, with the biological interpretation of layers L0, L1, and L3 highlighted. The gene regulatory network structure learned previously is embedded in layers L0 and L1, with different types of regulatory linkages and functional elements shown. **(C)** Shown are examples of associations found: model traces are shown for three co-expression modules and associated higher-order groupings prioritized by the DSPN schizophrenia model, along with functional annotations enriched at each level. Genes, enhancers, and SNPs associated with each module are shown. **(D)** The performance of different models is summarized, comparing performance across models of different complexity; using different predictors (genotype/transcriptome); and with or without imputation (colors highlight relevant models for each comparison). Performance accuracy on a balanced sample is shown first, with variance explained on the liability scale shown in brackets. LR-gen and LR-trans are logistic models using the genotype and transcriptome as predictors respectively; DSPN-Imput and DSPN-full are the DSPN model with imputed intermediate phenotypes (genotype predictors only) and fully observed intermediate phenotypes (transcriptome predictors) respectively. Differential performance of models is shown in terms of improvement above chance, for instance comparing LR-gen and DSPN-Imput accuracy improves from 53.8% to 60.2%, which can be expressed as a 2.7X improvement above chance (+10.2% vs. +3.8%, blue). Corresponding improvements in liability variance scores are shown in brackets.

####

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | **Hyejung Won** | | Mar 26 (2 days ago) |  |  |
| |  | | --- | | to Daniel, Daifeng, Mark, Shuang | | | |

Hi Daifeng,

A few comments in Figures.

**Figure 1**: Why are we having a species axis in the right?

Some characters have outer lines (especially the characters in the tissue box), making them difficult to read.

I'd have all characters as Arial or Helvetica.

**Figure 2**:

C: Some characters are not aligned.

D: The colormap on the right is too big

E: Can you point out which groups show significant difference (put \* between CON and ASD etc)? Also, use outer tickmarks for the plot as some of the x axis ticks are too close to the box plots.

**Figure 3**:

A: Is there a reason using the color gradient for H3K27ac bars?

B: This figure is unclear what message it would like to convey

**Figure 4**:

I will make D and E in the same format as both of them are enrichment plots.

D: Why not using multiple correction?

**Figure 6**:

Fonts into all Arial. B and C hard to read.

Best,

Hyejung

####