**Comprehensive resource and integrative model for functional genomics of the adult brain**

**Abstract**

Understanding how genomic variation influences brain phenotypes and disorders remains a key challenge. To this end, the PsychENCODE consortium has generated large-scale datasets on the adult human brain, including genotyping, RNA-seq, ChIP-seq, ATAC-seq, Hi-C and single-cell data on healthy and diseased tissues of thousands of individuals with different phenotypes. Using this, we developed a comprehensive resource on functional genomics of adult brain including ~2 million QTLs for expression and chromatin, and ~80k active enhancers. Leveraging single cell data explaining >85% tissue expression variation, we deconvolved the tissue-level gene expression and found ~500 QTLs significantly associated with the cell fraction changes for various phenotypes. Comparing this resource with others using spectral analysis, we show that the brain has unique expression and greater non-coding transcription than most other tissues, but not for chromatin. Moreover, we integrated the Hi-C and regulatory data to predict the gene regulatory network linking all possible functional genomic elements, identifying 112 novel linkages between psychiatric GWAS SNPs and genes from known 22 linkages. Finally, based on this network, we developed a deep-learning model to predict genotype-phenotype associations, with ~3.5X accuracy improvement achieved by using functional genomics to learn to impute intermediate phenotypes first compared to directly predicting high-level phenotypes from the genotype. This model highlights intermediate genes and functional modules, revealing potential mechanisms, and also enables quantitative imputation of missing transcriptome and epigenome information from genotype data only.

**I. Introduction**

Disorders of the brain [[dc: change to “Psychiatric disorders”]] affect nearly a 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 to cardiac disease, for which lifestyle and pharmacological modification of environmental risk factors has had a profound effect on disease morbidity \cite{}, or cancer, which is now understood to be a direct disorder of the genome \cite{}. Though GWAS studies have identified many genomic variants associated with psychiatric disease risk, a detailed understanding of the precise molecular mechanisms behind these associations remain elusive \cite{}.

To this end, a number of genomic studies have recently focused on discovering genomic variants relating to psychiatric diseases. As such, a variety of common and rare variants and linked genes have been found to be associated with brain and psychiatric disorders \cite{26404826}. For instance, the Psychiatric Genomics Consortia (PGC) identified 142 GWAS loci associated with schizophrenia \cite{29483656}. Many variants for psychiatric disorders have been found to lie in non-coding regions \cite{26404826}, suggesting potential roles in gene regulation. Other large consortia have annotated many of these non-coding elements -- e.g., expression quantitative trait loci (eQTLs) and eGenes in GTEx, and enhancers from ENCODE and Epigenomics Roadmap that are associated with various human cells and tissues. Though some of these elements relate to the brain, none of the consortia have specifically tailored their efforts toward comprehensively identifying [[dc: functional elements]] elements functional in the brain. Pioneering work from the CommonMind consortium \cite{27668389} has attempted to identify brain eQTLs using bulk tissue RNA-Seq data but has not reached maximal scale such as missing QTLs with chromatin activities and single cell fractions. Thus, comprehensive data integration for functional genomic data is essential to fully understand the molecular mechanisms in adult brain. In particular, these genomic elements [[dc: delete]] relating to adult brain from previous genomic studies suggest that the gene regulation is a key molecular mechanism [[change to: implicate gene regulation as a key molecular mechanism]] revealing how genomic variants affect gene expression in brain, so a number of new technologies have been developed to detect different types of gene regulatory activities at particular scales. For example, recent HiC studies have been used to identify specific chromatin structural and regulatory elements, such as active enhancers in fetal brain \cite{ 27760116}. Single-cell sequencing techniques also offer great promise for studying the transcriptome. However, each of the studies that leverage such technologies have generally focused on individual aspects of brain functional genomics. Therefore, these data have not yet been fully integrated at scale.

To this end, we have built a central, publically available resource for adult brain functional genomics, including all the raw and uniformly processed data at both tissue and single cell levels from PsychENCODE and other related major projects \cite{27339989, 26060301} with up to ~6k samples. By leveraging this resource, our analyses identified various functional genomic elements and QTLs [[dc: be less general?]]specific to the adult brain, including novel psychiatric GWAS and gene linkages. We also combined these elements and built an integrated deep-learning model to predict the molecular relationships from genotype to phenotype with high accuracy and impute missing data. The results obtained from this model are then studied in relation to specific brain phenotypes and psychiatric disorders.

## **II. Comprehensive resource for adult brain functional genomics [[dc: title sounds clunky. Maybe change to something like “A comprehensive functional genomics resource for studying of the adult brain”]]**

We designed this resource to provide a coherent and comprehensive structure to the data (http://adult.psychencode.org/, details in supplement). Broadly, it organizes a large amount of data for brain functional genomics pyramidally, with a large base of raw data files (much of it as restricted-access data, such as individual genotyping and raw next-generation sequencing data of transcriptomics and epigenomics), a middle layer of uniformly processed and easily shareable results (such as open chromatin peaks and gene expression quantifications), and a compact cap at the top, consisting of an integrative model (based on imputed regulatory networks and QTLs) [[dc: text very similar to this appears in the pre-amble. Take text directly from pre-amble, or maybe delete this text altogether here (it does not seem to be essential that this text be included here in the main text, since it seems very redundant w/Fig. 1]]. As shown in Figure 1, to build the base layer, we included all the datasets from PsychENCODE related to the adult and merged these with other relevant data from additional projects including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, and recent brain single cell studies. In total, this resource constitutes up to ~6k data samples derived from 1931 [[===dc: this number will need to be changed (here and elsewhere) since we’re only taking adult samples]] individual adult brains from multiple cohorts, which covers a large representation of brain phenotypes and psychiatric disorders. The major data types include RNA-seq, ChIP-seq, ATAC-seq, HiC, [[dc: change to single-cell gene expression data?]]single-cell data, and genotyping. (The later required large-scale imputation for all the PsychENCODE datasets, and we make full genotype sets available). Furthermore, the PsychENCODE project developed a specific "reference brain" project utilizing many assays on the same set of brain tissues, which we used to develop an anchoring annotation for the entire resource (Supplement).

**III. Bulk and single cell transcriptome analysis: deconvolution explains gene expression in terms of cell fraction changes**

To identify the genomic elements that exhibit transcriptional activities specific to the adult brain, we used the ENCODE standard pipeline to uniformly process the RNA-seq data of all available samples from PsychENCODE and GTEx. Using these data, we identified interpretable functional elements, such as non-coding regions of transcription, sets of differentially expressed and co-expressed genes characterizing various brain regions, phenotypes and disorders, which are provided as part of our resource (see Supplement for details). In particular, the co-expressed genes are summarized as a list of gene co-expression modules for psychiatric disorders \cite{cap1} and for brain regions after clustering together with other tissues of GTEx (see Supplement).

Brain tissues has been found to comprise a variety of cell types, including neuronal and non-neuronal cells, and their expression differences might be attributed to the cell fraction changes \cite{18849986, 27409810, 21614001, 29439242}. Therefore, one question with measuring gene expression changes over a population in our brain tissue samples is reliably determining whether the changes are driven by gene expression in a particular cell type or whether they result from changes in relative proportions of various cell-types.

To address this, we integrated single cell transcriptome data to discover how the gene expression from cell types contribute to bulk gene expression. 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 basic cell types in the brain (i.e., 16 neuronal types, 5 major non-neuronal types and a dozen of additional developmental types from PsychENCODE; see Supplement). This list constitutes a matrix (C) of the gene expression signatures of 25 basic cell types, which are mostly concordant with what has been published, with some modifications (Figure S5 and Discussion). Across these cell types, we found that a number of genes whose expression levels vary much more substantially than they do across individuals in a population (e.g., the dopamine receptor gene DRD3, Figure 2A). This implies that small variations in cell types can readily give rise to substantial changes in bulk gene expression at the [[dc: insert “bulk”]] tissue level.

To explore this further, we performed an unsupervised learning analysis for bulk tissue expression data to identify the primary components as they relate to different single cell types. In particular, we decomposed the bulk gene expression matrix (B) from our resource using non-negative matrix factorization (NMF, Figure 2B, see Methods); i.e., B~=V\*H [[===dc: Here, the “H” and “V” matrices really need to be more explicitly defined (just as the B matrix is defined just above)]], and then determined whether the top components (TCs) of the NMF (ie, NMF-TCs) that capture the majority of covariance in the data and the 25 standard gene expression signatures [[===dc: The term “standard gene expression signatures” is being used prior to being defined – this may appear unclear to the reader]] of single cells are consistent. As shown in Figure 2B, we found a number of NMF-TCs are highly correlated with the gene expression signatures [[dc: see note above – how were these signatures defined? What are they? Need to make more explicitly clear to reader, even if only 1 sent is devoted to explaining this]] of neuronal, non-neuronal and developmental-related cell types. This demonstrates that an unsupervised analysis derived from the main components of the bulk tissue data roughly matches the single cell data, partially corroborating our standard 25 cell types.

We then want to understand how these cell types contribute to bulk tissue gene expression and relate to different phenotypes. Previous studies found that the gene expression changes at the tissue level can be significantly associated with various cell types; e.g., the cell type specific gene co-expression modules \cite{18849986, 19829370}, but have not systematically and quantitatively revealed how different cell types contribute to the tissue gene expression changes. Thus, we used the supervised method to estimate the cell fractions for tissue samples of individuals. In particular, as shown in Figure 2B, we de-convolved the normalized bulk gene expression matrix of tissue, B using the single cell signatures (C) to estimate the cell fractions W, by [[dc: strictly speaking, I don’t think we’re solving an equation – isn’t this just a factorization?] ]solving the equation “B~=W\*C” (See methods). As a validation, we found that our estimated fractions of NEU+/- cells match the experimentally determined fractions for the reference brain samples (Figure S6). Moreover, we found that using the single cell expression signatures weighted by the derived cell fraction can explain much of the population-level expression variation (i.e., across tissue samples of the same brain region taken from different individuals). Specifically, we find 1-||B-WC||2/||B||2>85% [[dc: maybe explicitly state the significance of this inequality]] (Methods).

Furthermore, we found cell fraction changes were found to be highly associated with different phenotypes and psychiatric disorders (Figure 2E, Supplement Fig XXX). For example, the particular excitatory and inhibitory neurons exhibit significantly different fractions between healthy male and female samples (Ex3 and In6). The fraction of Ex3 is also significantly reduced in ASD samples (p<2.74e-11, ANOVA test), while non-neuronal cells (e.g., oligodendrocytes) are represented in much greater abundance. Another interesting association we found was that cell fractions change with age. In particular, the fractions of neuronal type(s) (Ex3 and Ex4) significantly increase with age (trend analysis p<6.3e-10 and 1.5e-6), but some non-neuronal types such as oligodendrocyte are found to decrease (p<2.1e-14). Furthermore, these age-related cell fraction changes are also potentially associated with differentially expressed genes across age groups (Figure 2E). For example, the gene involved in early growth response is down-regulated in older age groups, whereas the gene ceruloplasmin is down-regulated among middle-aged groups.

**IV. Active enhancers in adult brain**

In addition to the transcriptome data, we uniformly processed chromatin data in the resource to give uniform peak calling lists and single tracks for adult brain epigenomics. This includes the H3K27ac ChIP-seq data of a cohort of 150 postmortem samples from normal individuals in three different brain regions, namely, prefrontal cortex (PFC), temporal cortex (TC) [[dc: I advise against using the acronym “TC” here, since “TC” has already been introduced as the acronym for “top components” above]] and cerebellum (CBC). We pooled together the processed peaks from the cohort [[dc: delete]] respectively in each region, generating a consistent set of 37,761 H3K27ac peaks in PFC, 42,683 peaks in TC, and 26,631 in CBC. Each of the pooled peaks is in more than half of the samples of the corresponding brain region. We also processed the H3K27ac and H3K4me3 ChIP-seq data for the reference brain. Combining the ChIP-seq data and the ATAC-seq data from the same reference brain, we identified a consistent set of ~80k brain active enhancers in the prefrontal cortex, >90% of which overlap with the Epigenomics Roadmap annotation (See details in the Supplement). We examined the overlaps of these enhancer regions with the pooled H3K27ac peaks and found that around 90% of the identified enhancers reside in PFC H3K27ac peaks and TC H3K27ac peaks, despite that temporal cortex has a slightly larger set of pooled peaks. However, only 34% of the PFC enhancers are found in the cerebellum pooled H3K27ac peaks, indicating very different epigenome landscape of cerebellum from prefrontal cortex. This is consistent with previous findings that prefrontal cortex and temporal cortex shared similar chromatin dysregulation signature in ASD and is largely different from that of cerebellum.

We then looked at the epigenetic signal variations across individuals at these enhancers. For this we examined the H3K27ac peaks across 50 PFC samples (Figure 3A). As expected, not all the identified enhancers have H3K27ac signal in every individual. In fact, on average only about 51K of the identified enhancers are found to be active in one individual person at a time, with an even smaller fraction of enhancers have H3K27ac signals across all 50 PFC samples. Yet the majority (~68%) of the identified enhancers are active in more than half of the population. The variation of enhancer state can be also observed from the cumulative numbers of active enhancers by increasing the sample size. For example, the cumulative number increase dramatically for the first 20 sample examined, but becomes nearly saturated at the 30th sample to more than 75K (Figure 3B). Again, this suggests that enhancer activity varies across individuals, yet the majority of brain enhancers are active in most of the population. We also compared the distribution of the saturation curve on the normal samples (N=50) with the ASD samples (N=43) and found no significant differences in overall enhancer activity (KS-test p>0.1).

**V. Consistently comparative analysis reveals the brain related transcriptomic and epigenomic activity**

[[dc: maybe more clearly state this as a highlight in the intro?]] One key aspect of our analysis is that we uniformly processed the transcriptomic and epigenomic data across PsychENCODE, ENCODE, GTEx and Roadmap. This allows us to compare the brain to other organs in a consistent fashion in order to delineate gene expression and chromatin activities unique to the brain. We attempted several methods including PCA and tSNE for an appropriate comparison, and finally used the Reference Component Analysis (RCA). PCA, though popular, tends to capture global structures, ignoring most of the local structure, but it can easily be influenced by outliers. On the other hand, t-SNE analysis preserves local structure but “shatters” global structure (see discussions in supplement). RCA, however is capable of capturing local structure while maintaining meaningful distances in global structure space. It projects the gene expression in an individual sample against a reference panel, and then essentially reduces dimensionality of the individual projections. We did RCA consistently for comparing brain and other tissues in terms of their similarities of both the transcriptome and the epigenome.

Our comparative analysis for gene expression shows that the brain tends to separate from the other tissues in the first component, showing it has a more distinct expression pattern (Figure 3C). This difference is accentuated when focusing on the tissue cluster centers and the distributions surrounding them. Inter-tissue differences are much larger than intra-tissue ones. A different picture emerges when one looks at our comparison using chromatin data. It shows that the chromatin levels at enhancers are much less distinguishable between brain and other tissues (Figure 3D).

Our RCA analysis focuses on inter-tissue differences in well-annotated regions (i.e. genes, promoters and enhancers). In addition to the expression differences in protein-coding genes, a much transcriptional diversity is present [[dc: rework]]across tissues in intergenic and noncoding regions. For protein-coding regions, it has previously been demonstrated that testes and lung tend to have the largest transcriptional diversity in terms of the percentage of transcribed regions (Figure 3E). However, when we shift to non-coding and unannotated regions, we find that brain tissues (such as cortex and cerebellum) do, to some degree, stand out by exhibiting greater non-canonical transcription than most other tissues.

**VI. QTL analysis**

To understand how the genotype affects the transcriptome and epigenome in the adult brain, we used the PsychENCODE resource data to identify quantitative trait loci (QTLs) affecting gene expression and chromatin activity. In particular, we calculated: expression QTLs (eQTLs), chromatin QTLs (cQTLs), splicing QTLs (sQTLs) and cell fraction QTLs. For eQTLs, we adopted a standard approach, adhering closely to the established GTEX eQTL pipeline. We identified ~2.5 million eQTLs and 32944 e-genes including non-coding genes in DLPFC by using matched genotype and gene expression data of 1387 adult individuals. There are 1,341,182 unique SNPs in the eQTLs from the 6,189,301 potential SNPs within 1 Mb of gene TSS. Because of the effect of linkage disequilibrium, many of these eQTL SNPs for the same gene were correlated. We calculated the genotype correlation coefficient (r2) for all eQTL SNPs and determined pruned eQTLs using r2 <0.6. This conservative estimate is a substantially larger number of eQTLs and eGenes than previous brain eQTL studies such as CommonMind and reflects a large sample size and statistical power (Supplement). We believe this eQTL number is close to saturation, in terms of associating almost every variant with some expression modulating characteristic (Figure 4B). We also applied the same QTL calculation pipeline to splicing and identified 157,592 sQTLs.

For the cQTLs, the situation is more complicated. There are no established standard methods for calculating these on a large scale, though previous efforts have detected QTLs associated with various chromatin activities [[dc: rework]]on non-brain context \cite{25799442, 26300125}.

To properly identify them, we focused on a reference set of enhancers to define the region associated with chromatin activity and then looked at how this activity varies in these enhancers across 292 individuals (Figure 4A, see Methods). Overall, we were able to identify ~46k cQTLs in addition to the 6200 cQTLs identified using individuals from CommonMind \cite{ https://doi.org/10.1101/141986}.

Next, we were interested to see if any SNVs were associated with changes in the relative fractions of various cell types across individuals; i.e., cell fraction QTLs (fQTLs). In total, we identified the 3720 distinct SNVs constituting 4186 different fQTLs between different cell types (all fQTLs were conservatively identified using Bonferroni-corrected p-values to 0.05). Of these the fractions of microglial and a particular type of excitatory neuron (Ex8) were associated with the most SNVs. Next, after factoring out these cell type differences, we also identified 260,280 SNVs significantly associated with the gene expression changes across individual tissues - these trans-eQTLs represent SNP-expression associations unexplained by variation in cell types. [[dc: ok to keep these for now – but don't forget to resolve orphans and widows prior to submitting]].

To further dissect the genomic elements associated with various QTLs we identified, we looked at how they overlapped with each other and different genomic annotations. For example, we observed a significant enrichment of eQTLs on ePromoters, TFBS and TSS (xx%, Figure 4C)(more significant than for the other QTLs), and also found cQTLs are mostly enriched on enhancers. As expected, there were lots of overlaps among eQTLs,cQTLs, sQTLs and fQTLs (Figure 4D). The percentage of overlapped snps of eQTL with other QTLs was the highest among all QTLs. Only 31% of fQTL SNPs overlapped with other QTLs, the lowest among all QTLs.

**VII. Gene regulatory networks in adult brain**

In this section, we provided an integrative analysis at the gene regulation level for the data and genomic elements described above and predicted a gene regulatory network revealing how the genotype and regulators control target gene expression in adult brain. To this end, we first process a reference Hi-C dataset for adult brain, [[dc: language too strong here? Maybe run by Hyejung – not sure if this is the script they use in the HiC community]] which provides direct physical evidence for potential interactions between enhancers and promoters (Figure 5A). Specifically, we generated and processed Hi-C data for the same reference adult brain that was used to identify the brain active enhancers \cite{27760116} (Supplement). In total, we identified 2,735 topologically associated domains (TADs) which set potential physical boundaries of enhancer-promoter interactions and then 149,097 putative enhancer-promoter interactions in adult DLPFC. This HiC dataset is substantially different than the fetal brain HiC data set (see suppl) highlighting the importance of stage.

As expected, we found that ~75% of enhancer-promoter interactions occur in the same TADs (Figures Sxxx), suggesting that a high percentage of cis-regulations from enhancers to target genes are through interactions to targets’ promoters within TADs; e.g., we found ~3.2 million enhancer-target linkages in total; [[===dc: do we have any means of knowing how many would generally be intuitively reasonable?]]. Also, as expected, the genes that have more potentially regulatory enhancers interacting with their promoters tend to express higher (Figure 5B).

We next integrated the Hi-C dataset with eQTLs to assess how much of the common variation-associated gene regulation is mediated by chromatin interactions. Interestingly, 30.7% of e-genes show evidence of chromatin interactions, accounting for 204,008 eQTLs and xxx sQTLs (Figure 5C). To our surprise, e/sQTLs supported by Hi-C evidence showed stronger associations not only to e/sQTLs without genomic annotations, but also to exonic and promoter e/sQTLs.

As a second step to build a full gene regulatory network, we integrated the enhancer-target linkages identified by TADs with other regulatory elements and relationships such as transcription factors (TFs) (Methods). In particular, we found TF binding motifs using ENCODE data and imputed TF-target gene relationships if TFs have enriched binding motifs on the target gene’s promoters and enhancers, providing a reference wiring network for TF-target regulation in brain. Using these “wiring” relationships, we inferred the final gene regulatory network linkages, which include the active regulatory links relating QTLs, enhancers, and transcription factors to target gene expression (Methods). This network also has a few particular characteristics such as scale-free and hierarchical structures, which have been revealed by previous network analyses (Figure 5D). Given a target gene, we associated coefficients with each of these wiring linkages predicting the target gene expression from the activities of their regulatory elements. We model them as simple linear relationships but regularize to minimize the number of connections using an elastic net model (Methods). Overall, we found this model could successfully predict expression of >xx% genes with the minimum mean square errors < \*\*xxx. For example, the expression of gene, \*\*\*xx can be predicted by its TFs expression with accuracy = \*\*\*xxx based on our model. In total, the gene regulatory network comprises ~43k TF-gene linkages (Elastic net coefficient’s absolute value>0.2) and ~1.8million eQTL-gene (FDR<0.05) linkages (Figure 5D).

**VIII. GWAS**

Using the regulatory map built with Hi-C and eQTLs, we are further interested to predict potential target genes of GWAS variants with psychiatric disorders (Figure 4E). First, we found significant associations between eQTLs and GWAS disease traits. In particular, we calculated the enrichment in cis-eQTL SNPs of GWAS SNPs of three brain related disorders (schizophrenia, bipolar disorders and parkinson’s disease) and non-brain related disorders (CAD, asthma and type 2 diabetes). As expected, enrichment for GWAS SNPs of brain disorders on cis-eQTLs SNPs are more significant than the ones of non-brain disorder GWAS SNPs. Schizophrenia GWAS SNPs have the highest enrichment. Collectively, these QTLs annotate a larger fraction of GWAS SNPs involving the brain (e.g., 21% in schizophrenia, 18% in bipolar) than previously observed, providing important leads on which genes are affected in disease.

To further predict the target genes of GWAS variants we exploited the combined Hi-C, eQTL-eGene and enhancer-gene linkages. For example, to identify putative target genes of newly identified 142 schizophrenia GWAS loci \cite{27869829}, In total, we identified 488 putative schizophrenia-associated genes, hereby referred as SCZ genes, and 99 genes that show evidence both at the level of Hi-C and eQTLs, providing a high-confidence gene list (Figure 5xx). This is a huge increase from the previously annotated 22 genes across 19 loci based on CMC adult brain eQTLs \cite{27869829, 27668389}. The majority of SCZ genes (288 genes, ~59%) were not in linkage disequilibrium (LD, r2>0.6) with index SNPs (Figure 5E), consistent with the previous observations that regulatory relationships often do not follow linear genome organization.

[[dc: delete]] We looked at the characteristics of the SCZ genes. First, as expected, these genes shared many of the characteristics of known SCZ genes. In particular, they are enriched in the loss-of-function mutation intolerant genes \cite{27869829}, translational regulators, cholinergic receptors, calcium channels, and synaptic genes attributing to schizophrenia and also differentially expressed genes associated with schizophrenia (Figure Sxx). Second, we found that these genes can reveal additional biological mechanisms and functions that GWAS is unable to identify. They were enriched for genes and co-expression modules dysregulated in DLPFC of schizophrenia-affected individuals \cite{27668389}, suggesting that common variation-mediated gene regulation contributes to the gene dysregulation in schizophrenia (Figure Sxxx). This also hints that there likely exists shared genetic etiology between common and structural variation since the SCZ genes are often affected by recurrent CNVs in schizophrenia. We further integrated this gene list with the single-cell profiles and found, interestingly, they were enriched in a variety of different neural types. (Figure 5E).

**IX. Integrative modeling to relate genotype to molecular and high-level phenotypes in the adult brain**

The interaction between genotype and phenotype involves multiple levels; in this section, we perform another level of integrative analysis by embedding our gene regulatory network derived above into a larger multi-level model. For this purpose, we introduce an interpretable deep-learning framework, a Deep Structured Phenotype Network (DSPN, Figure 6, Supplement Sec. 7). This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from the QTLs, Hi-C gene regulatory connections incorporated into our elastic net regression. As shown (Figure 6a), traditional classification methods such as logistic regression predict phenotype directly from genotype, without inferring intermediates such as the transcriptome. We build the DSPN via a series of intermediate models which add layers of structure to this traditional approach, including a layer for intermediate molecular phenotypes such as gene expression and chromatin state, multiple layers for functional modules and other intermediate phenotypes which may be inferred as hidden nodes in the network, and a layer for high-level phenotypes such as psychiatric disorders. Finally, we use special forms of connectivity (enforcing sparsity and adding lateral intra-level connections) to integrate our knowledge of QTLs, regulatory network structure, and co-expression modules from earlier sections of the paper (Figure 6b). By using a generative architecture, we ensure that the model is able to impute intermediate phenotypes when needed, as well as providing a predictive model for high-level traits and phenotypes.

Using the full model with genome and transcriptome data provided, we show that adding the extra layers of structure in the DSPN allows us to achieve substantially better prediction of disease and other high-level traits than without -- ie just using a traditional genotype to phenotype model (Figure 6d). In particular, we achieve a prediction of 73.6% with the extra layers for Schizophrenia vs 63% without (Figure 6d, DSPN vs LR-trans). Further, comparison with a simple logistic predictor from the genome alone shows that the transcriptome carries significant further trait relevant information, which the DSPN is able to optimally extract [[===dc: this improvement over a logistic regression-based predictor could be better highlighted in intro]]. For instance, a logistic predictor is able to gain a 2.8 times improvement when using the transcriptome versus the genome for Schizophrenia prediction (+13% vs. +4.6% from 50% chance), while the DSPN is able to gain a 5 times improvement (+23% vs. 4.6%); this may reflect the need to incorporate non-linear interactions between intermediate phenotypes at multiple layers as in the DSPN. Moreover, the model also, allows us to perform joint inference / imputation of the intermediate phenotypes (i.e. the transcriptome and epigenome) and high-level traits from the genotype alone using the DSPN. Using imputation, we achieve up to 66.7% for disease trait prediction, which is better than direct prediction using an additive logistic model, which achieves 50-56.7%, hence generating a ~3.5X gain in accuracy above chance averaged across conditions (+13.2% vs. +3.8%, see Figure 6d, DSPN-Imput vs LR-gen). These results demonstrate the usefulness of even a limited amount of functional genomics information for unraveling gene-disease relationships, and that the structure learnt from such data can be used to make more accurate predictions of high-level traits even when absent.

We transform the results above to the liability scale in order to compare with heritability estimated on this scale using GCTA. Using the PsychENCODE cohort, we estimate that common SNPs and eSNPs explain x% and x% of liability for Schizophrenia respectively, which is comparable to previous estimates. The imputation-based DSPN model explains a comparable level of variance to the eSNPs (4.5%, Fig. 6d), although we note that the DSPN may be capturing epistatic interactions not modeled in SNP-based heritability (Figure Sxx). The full DSPN model estimates that the transcriptome-based liability for the DLPFC is ~32.8%. Although we expect that a large portion of this will overlap with the common SNP based liability (which has previously been estimated as 25.6%) and genetically determined non-linear interactions, it may also include environmental and trait-influenced contributions (Supp Fig. A1), meaning that it is an upper-bound on the genetically determined liability modeled by the DSPN. Similar estimates of the liability explained for Bipolar and ASD by the DSPN (imputation and full models) are given (Figure 6d).

We examined the specific connections learnt [[dc: change to “learned”?]] by the DSPN between intermediate and high-level phenotypes for potentially relevant biological interactions. We included known co-expression modules as submodules in the model, and examined which of these the DSPN prioritized as well as new sets of genes associated with the DSPN latent nodes that were uncovered at each hidden layer using a common prioritization scheme (Supplement Sec. 7.5 and Supp Fig. A2). For instance, in Schizophrenia, we found that the highest prioritized module in the DSPN was associated with Dopaminergic and Glutamatergic synapse and calcium signaling pathways, with other modules associated with Oligodendrocyte marker genes, and the Complement cascade pathways, which confirms and extends previous smaller scale analyses \cite{JK’s refs}. Further, we found that excitatory neuron marker gene were enriched in the highest prioritized module for age and the gene NRGN occurred in many of the top prioritized modules/submodules, in agreement with the earlier analyses \cite{}. We provide a full summary of the functional enrichment analysis for all disease and high-level traits in supplement. We further show that the modules prioritized by the DSPN are strongly enriched for GWAS variants (Supplement).

**IX. Discussion (359)**

We integrated PsychENCODE datasets with other resources, and developed a comprehensive resource consisting of various functional genomic elements for the adult brain including data from 1931 individuals [[dc: again – this 1931 number will need to be changed throughout in order to reflec the fact that we only study adult samples.]]. This resource serves as an important step for linking variants to functional genomic elements in the brain. Overall, our study has identified a very large-scale set of eQTLs and eGenes for adult brain, several folds more than previous studies, almost achieving saturation for protein coding genes. Therefore, we suspect that larger population studies will not significantly expand on these. However, there exist other aspects of brain QTLs that can be extended in the future. For example, increasing individual sample size of chromatin data size and quality such as identifying enhancers using STARR-seq will help discover more accurate cQTLs.

Another area of future development is single-cell analysis. Current techniques suffer from low capture efficiency, and so it remains challenging to reliably quantify low-abundant transcripts/genes \cite{26949524, 25053837}. In this study, we found that 25 basic and known cells could explain large expression variations across tissues. However, there still exist the gene expression heterogeneities among these cell types (Figure Sxx\_magic [[dc:?]]), implying subdivision into potential additional types in the future. Moreover, given the challenge that high expression variability for certain cell types such as axons and dendrites from their scRNA-seq data \cite{25053837}, recent single-nucleus RNA sequencing techniques have emerged to address it and also complement the gene expression measurement at the whole cell level that may be impacted by various aspects such as intercellular interactions \cite{28729663, 29227469}. Also, our model performance can be potentially improved if we integrate more functional types such as the more accurate QTLs and cell fractions, imaging [[===dc: here, I think we’re alluding to the “iQTLs”? If so, we should provide a ref (if it has already been published)]] and medical data. Finally, this model provides a useful compression on the functional genomic datasets; e.g., XXX(TBD) KB for the model representation vs. XXX(TBD) TB for the resource data.

**X. Figures**

**Figure 1. Comprehensive data resource of functional genomics in adult brain**. As shown on data cubic, this resource includes multi-dimensional functional genomic data for adult brain generated from PsychENCODE and other related consortia such as GTEx, Epigenomics Roadmap, ENCODE. The resource data has various types including genotypes (N=xx), RNA-seq (N=xx), ChIP-seq (N=xx), ATAC-seq (N=xx), Hi-C (N=xx), single-cell transcriptomic data (N=xx), and in total ~6k samples from 1931 individuals. These individuals represent major brain phenotypes, gender and psychiatric disorders: Health (n=1445), SCZ (n=270), BP (n=160), ASD (n=65), AFF (n=8), Male (n=1244), Female (n=700), Developmental stages, Age, etc. (details in Supplement). In addition, these datasets cover major brain regions relating to brain disorders: DLPFC (n=xxx), Temporal cortex (n=xx), Cerebellum (n=xx).

**Figure 2. Deconvolution analysis of Bulk and single cell transcriptomics reveals cell fraction changes across tissue phenotypes and disorders**. **(A)** The genes have significantly higher expression variability across single cells than tissue samples. Left: dopamine gene, DRD3; Right: Age gene, xxx. **(B)** Top: the bulk tissue gene expression matrix (B, genes by individuals) can be decomposed by non-negative matrix factorization (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~=V\*H. 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~=C\*W. Three major cell types are neuronal cells (blue), non-neuronal cells (red), 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=xxx). For example, NMF-15 highly correlates with Ex3 cell type (r=xxx). **(D)** The estimate cell fractions contribute >85% bulk tissue expression variations; i.e., 1-||B-C\*W||/||B||>0.85. **(E)** The cell fractions change across brain phenotypes and disorders. The neuronal cell types (e.g., Ex3 and In6) have significantly higher fractions in Male than Female samples (p<xxx, and p<xxx). Also, they and Oligodendrocytes have lower fractions in ASD than others. In particular, Ex3 cell fractions increase with aging (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 are used to identify active enhancers. Enhancers are located in open chromatin region (high ATAC-seq signal), with strong H3K27ac signal and lack H3K4me3 signal. Enhancer activity varies 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 achieves a saturation to more than 70K with 20 samples. **(C)** The coefficients (PC1 vs. PC2) of RCA analysis for gene expression data of PsychENCODE samples (dark green), Other external brain samples (light green) including GTEx and Other tissue samples (magenta). **(D)** The center (cross) and ranges of different tissue clusters (dashed ellipse) on RCA scatterplot of (C). **(E)** The tissue clusters of RCA coefficients (PC1 vs. PC2) for chromatin data of active brain enhancers. The brain cluster (green ellipse) includes the PsychENCODE samples and other tissue clusters (magenta ellipses) consist 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) vs. on cumulative tissue samples (y-axis) for select major tissue types including Cerebellum (lightblue), Cortex (xxx), Lung (xxx), Ovary (xxx), Skin (xx) and Testes (xx).

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

**Figure 5. Data integration and modeling predicts gene regulatory network, revealing additional GWAS genes for psychiatric disorders**. **(A)** The full Hi-C data for adult brain found a variety of TADs (contact heatmap) which provide the genomic regions for potential enhancer-promoter interactions. The example highlights the gene XXX where its promoter, enhancers, Hi-C, eQTL, and TADs all point to the same regulatory relationship. **(B)** The number of genes (left y-axis) and the normalized gene expression levels (right y-axis) vs. the number of enhancers that interact with the gene promoters. **(C)** eQTLs and sQTLs that are supported by Hi-C evidence show more significant P-values than those that are not. **(D)** A full gene regulatory network of TFs (green), eQTLs (red), enhancers (blue) and target genes (xxx). The TF-target relationships are predicted using their gene expression data across individual tissues and Elastic net regression. **(E)** Schematic showing how to identify putative target genes of schizophrenia GWAS loci using Hi-C and eQTL. Hi-C identified genes and eQTL identified genes are highly overlapping. The SCZ genes show higher gene expression levels on neuronal cell types than others. **(F)** Evidence shows that GWAS SNPs overlap eQTLs and Hi-C point to the same gene; e.g., TSNARE1.

**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 Deep Structured Phenotype Network (DSPN) models. The hierarchical layers from genotype to phenotype, representing various multi-scale functional elements in adult brain, include genotype (yellow), gene expression, enhancer activity and other intermediate phenotypes (green), function and module (purple) and high-level phenotype (red). The model also allows the gene regulatory linkages (green) on the same layer. **(B)** shows the DSPN structure in further detail, with different types of regulatory linkages highlighted. **(C)** shows examples of associations found: the DSPN prioritized functional modules associated with Dopaminergic and Glutamatergic synapse, and calcium signaling pathways, as well as ones associated with Oligodendrocyte markers and the complement cascade pathway for Schizophrenia. **(D)** summarizes the performance of different models, comparing performance as layers of structure are added to the model; for different predictors (genotype/transcriptome/all intermediate phenotypes); and with or without imputation. Performance accuracy on a balanced sample is shown first, with variance explained on the liability scale shown in brackets.