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

**I. Introduction (616)**

Disorders of the brain affect nearly a fifth of the world’s population (ref).  Unlike cardiac disease, where lifestyle and pharmacological modification of environmental risk factors has had a profound effect on disease morbidity and mortality (ref), or cancer, which is now understood to be a direct disorder of the genome (ref), until recently, little progress has been made in our fundamental understanding of the molecular cause of the brain disorders. Recent progress has come is the form of genetic association signals from large GWAS studies of the psychiatric and neurological disorders and currently hundreds of genomic locations that alter the disease risk are known (ref). Unfortunately, for most of these locations, we have little to no understanding of the molecular mechanisms that cause those alterations..[[\*\*Too long, shorten 50%]]

To this end, a number of genomic studies have been created to focus on discovering the genomic functions for adult brain. On one hand, a variety of genomic elements and variants have been found to associate with brain and psychiatric disorders; e.g., the Psychiatric Genomics Consortia (PGC) that identified 108 GWAS loci associated with schizophrenia. On the other hand, large consortia have identified the reference sets of genomic elements across the entire body; e.g., the eQTLs and eGenes from GTEx, and the 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 has specifically tailored its work towards comprehensively identifying the functional elmeents in the brain.

To address this, recent technologies have started to detect the specific molecular activities for brain. Particularly, single-cell sequencing techniques show great promise to study the transcriptome. Also, recent HiC and ATAC-seq studies found the specific chromatin structure and activity of the regulatory elements such as brain active enhancers. However, all of these studies have focused on individual aspects, and not fully been integrated to comprehensively understand the brain functional genomics.

stopped

Past work such as CommonMind attempted an integrative analysis to identify the brain genomic elements. Though promising, larger sample size, more comprehensive data is further essential to obtain comprehensive view of brain functional genomics [refs]. To this endeavor, the PsychENCODE Consortium (PEC) has generated and assembled a large-scale dataset on the adult human brain, including genotyping, RNA-seq, ChIP-seq, ATAC-seq, HiC and single-cell data on the high quality healthy and diseased brain tissue samples of thousands of adult individuals with different phenotypes. We have thus built a central, publically available comprehensive resource (http://adult.psychencode.org/) for adult brain functional genomics, including all the raw and uniformly processed data at both tissue and single cell levels from PEC and other related projects, including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, recent brain single cells [refs] with up to X,XXX samples. Using the resource, our analyses identified the functional genomic elements for the adult brain [[++QTLs]]. We also combined these elements and built an integrated deep-learning model to impute missing data [[relate to]] brain phenotypes and psychiatric disorders.

## II. Comprehensive resource for adult brain functional genomics (250)

We built this comprehensive resource to have a coherent data structure. Broadly, it organizes a large amount of data for brain functional genomics pyramidally, with a large base of raw data (controlled access such as individual genotyping and raw next generation sequencing data of transcriptomics and epigenomics), to a middle of uniformly processed and shareable results (eg as open chromatin peaks and gene expression quantifications), all capped by a small integrative model, based on inputed regulatory networks and QTLs. As shown in Fig 1, In terms of the data corpus for building this large-scale comprehensive resource, we included all the datasets from PsychENCODE related to the adult brain and merged them with other relevant data from projects including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, and recent brain single cell studies. In total, this resource has XXXX data samples of 1931 individual adult brains from multiple cohorts, covering high variability among brain phenotypes and psychiatric disorders. The major data types include genotyping, RNA-seq, ChIP-seq, ATAC-seq, HiC and single-cell data. (This 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 tissues, which we used to develope an anchoring annotation for the whole dataset [[\*\*see suppl for ref brain descrip.]] .

**III. Bulk and single cell transcriptome analysis and deconvolution explain gene expression and cell fraction changes (810s)**

We are interested to identify the genomic elements that have specific transcriptional activities in adult brain at the tissue level. In particular, we used the ENCODE standard RNA-seq pipeline to uniformly process the RNA-seq data of all available samples from psychencode[[\*\*uni abbrev]] and GTEx. Using these data, we found more interpreted functional elements such as sets of differentially expressed and co-expressed genes characterizing various brain regions, phenotypes and disorders [cap1], and reported them in our resource. Moreover, we constructed a gene co-expression network using the samples across brain and other tissues and clustered it into a number of gene co-expression modules, many of which reveal the expression patterns specific for brain samples.

brain tissues have been found to comprise a variety of cell types including neuronal and non-neuronal cells such as astrocytes [refs]. One issue with the changes of gene expression over a population in our brain tissue samples is whether the changes are driven by gene expression in a particular cell type or different quantities of various cell-types. To address this, we integrated the single cell transcriptome data to discover how the gene expression from various cell types contribute to the bulk gene expression using two strategies.

First, we used the standard pipeline to uniformly process single cell RNA-seq data in PEC in conjunction with the number of other single-cell studies on the brain to put together a list of cell types in the brain; i.e., 16 neuronal types, five non-neuronal types and xxx additional fetal types from PsychENCODE (Supplement). This forms a matrix of 25 reference cell types signatures we call C in the below. These signatures are mostly concordant with what has been previously published with some minor modifications in terms of cell clusters based on their gene expression similarities (Figure Sxxx). Across these cell types, we found a number of genes varies much more substantially than they do amongst individual tissues; e.g., the dopamine receptor genes (DRD) that associate with SCZ (Figure xxx). This implies that the gene expression variation of cell types can give rise to substantial changes in bulk gene expression at the tissue level.

Second, we did an unsupervised analysis for the bulk tissue expression data and tried to find its main components that potentially relate to the single cell types. In particular, we decomposed the bulk gene expression matrix (B) from our resource using non-negative matrix factorization (NMF, see Methods), and compared whether the top principal components of NMF (NMF-PCs capturing most data covariance[[\*\*is this the right word]]) and the 25 reference gene expression signatures of single cells are consistent. As shown in Figure XX, we can see a number of NMF-PCs highly correlate with the gene expression signatures of neuronal, non-neuronal and fetal cell types as above; e.g., the NMF-PCs shown in Figure xxx. This shows that our unsupervised analysis derived the main components from the bulk tissue data, roughly matching the single cell data, and partially validates these 25 cell types.

After this analysis, as shown in Figure xx, we de-convolved the bulk tissue expression matrix B using the single cell data matrix C to estimate the cell fractions W, by solving the equation “B=WC” (See methods). We found that the multiplication of estimated cell fractions and single cell expression data can explain large variation of expression at the population level (i.e., across tissue samples). That is, ||W\*C||2/||B||2 > 85%, [[\*\*\*is this accurate??]]where ||.|| is the Frobenius norm of matrix (Methods), which shows that over 80% bulk gene expression variation across samples can be accounted for a variation in single cell types. Moreover, we found that our estimated fractions of NEU+/- cells match the experimental measurements for reference brain samples (r=xxx, Figure xxx).

Furthermore, we found that the cell fractions of individual tissues (i.e., deconvolution coefficients from W) vary, and a number of cell population changes highly associate with different phenotypes and psychiatric disorders (Figure xxx). For example, the excitatory and inhibitory neurons (EX3 and In6) have significantly different fractions between healthy Male and Female. The EX3 cell fractions also decrease significantly in ASD samples (p<xxx) while the non-neuronal cells increasing (e.g., oligodendrocytes). Another interesting association we found was the cell fraction changes with Age. In particular, the fractions of neuronal type(s) (EX 3 and 4) are significantly correlated with Age (r = xxx), but non-neuronal type, Oligodendrocytes anti-correlate. The cell fraction changes also potentially drive the differentially expressed genes in Age at the tissue level (Figure xxx). [[[\*\*\* linkage unclear, is it a eg]]For example, we identified a group of genes that differentially express across ages (Figure xxx). In particular, the gene involved in early growth response is down-regulated at elder samples whereas the gene with ceruloplasmin is down-regulated around the middle ages. Finally, we report the individual cell populations along with significantly associated relationships between particular cell type fractions and phenotypes (Supplement).[[[\*\*\*\*move sent earlier???]]

**IV. Active enhancers in adult brain (195)**

In addition to the transcriptome data, the uniformly processed chromatin data in the resource gave rise to uniform quantifications, peak calling lists and single tracks for adult brain epigenomics. Then, we used these data and derived further simplified epigenomic representations for the adult brain.

First, we developed a consistent set of brain active enhancers. We processed the H3K27ac and H3K4me3 ChIP-seq and ATAC-seq data of the reference brain using the standard ENCODE ChIP-seq processing pipeline.  [[\*\*\* need to rewrite again to 4 sentenc]] We then identified an overall set of brain enhancers based on these experimental data of the same reference sample using the ENCODE 3 candidates regulatory element (cRE) pipeline, where the ATAC-seq peaks indicates the open chromatin in the brain, and the histone marks together with the distance to genes transcription start site (TSS) identifies the enhancer regions (Moore et al, in review). Finally, we intersect these brain enhancers with H3K27ac peaks to find brain active enhancers consistently across all the PEC and Roadmap data, including ~88,800 active enhancers in dorsal lateral prefrontal cortex (Supplement). We have also developed reference sets in additional brain regions including CBC and ACC. We also developed reference enhancer sets for the other tissues.

[[Numbers about the Enhancer variation across Roughly 80k enhancers but across the 50 samples on avg 1K additional in each]][[\*\*\* MTG to insert… see tiffs]]

**V. Consistently comparative analysis reveals the brain related transcriptomic and epigenomic activity (499)**

One key aspect of our analysis is that we, as consistently as possible, processed the transcriptomic and epigenomic data across PEC, ENCODE, GTEx and Roadmap together. This allows us to compare the brain to other organs in a consistent fashion to see if brain has unique gene expression and chromatin activities. This comparison couldn’t be achieved without such a large-scale uniform data processing. Moreover, we attempted several methods for an appropriate comparison. Principal component analysis (PCA) and t-SNE are two popular techniques. We found that the former captured only global structures, ignoring most of the local structure.[[\*\*\*but it was undulying influenced by outliers]] On the other hand, the latter keeps local structures intact but “shatters” global structures. For example, t-SNE tends to separate samples from the same tissue so that the cluster distances on t-SNE space are not proportional to real gene expression dissimilarities and does not give one a sense of the overall effect. Therefore, we found another technique that is capable of capturing local structures while maintaining meaningful distances in the global structure. Reference Component Analysis (RCA) projects the gene expression in individual sample against a reference panel, and then essentially reduces dimensionality of individual projections. In fact, we did RCA consistently for comparing brain and other tissues in terms of their similarities of both transcriptome (RNA-seq gene expression) and epigenome (ChIP-seq signals on our consistent set of enhancers).

Our comparative analysis for gene expression shows that brain tends to separate from the other tissues in the first component, showing it has a more distinct expression pattern and that all the brain tissue samples from the different projects grouped together (due to our uniformly processing). This difference is even accentuated when one looks not at all the individual but simply looks at the tissue cluster centers and the distribution about them. The difference between brain and other tissues is much larger than the one within any of the given tissues. A different picture emerges when one looks at our comparison using chromatin data; i.e., ChIP-seq signals on our consistent set of brain active enhancers. It shows that the chromatin levels are much less distinguishable between brain and other tissues (Figure xxx).

Our RCA analysis focuses on inter-tissue differences in well annotated regions, ie genes, promotors and enhancers. In addition to the expression differences in protein-coding genes, a tremendous amount of transcriptional diversity is present across tissues in intergenic and noncoding regions. Thus, we looked at the overall level of transcriptional diversity across tissues. On protein-coding regions, it has been previously known that testes and lung tend to have the largest transcriptional diversity in terms of the percentage of transcribed regions (Figure SYYY sat’d for genes). However, when we shift to non-coding and unannotated regions, we find that brain tissues such as cortex and cerebellum do stand out to some degree in having more transcriptions than most other tissues. This transcriptional diversity tends to increase with the number of samples (Figure xxx sat’d).

**VI. QTL analysis (625)**

To understand how the genotype affects the transcriptomic and epigenetic activities in adult brain, we used the psychencode resource data to identify quantitative trait loci (QTLs) affecting gene expression and chromatin activity. In particular, we calculated the association of SNPs with normalized gene expression and chromatin states (Methods[[[\*\* std ref]]) to find the quantitative trait loci associating with gene expression and epigenomic activities in adult brain, including several major categories: expression QTLs (eQTLs), chromatin QTLs (cQTLs), splicing QTLs (sQTLs) and even cell fraction QTLs. For the eQTLs, we adopted a standard approach, adhered closely to the established GTEX eQTL pipeline. We identified ~2M of eQTLs and ~17000 e-genes in DLPFC region. This is a conservative larger number of eQTLs than previous brain eQTL studies and reflects the very large sample size and statistical power we have. We believe it's moving close to saturating in terms of associating almost every variant with some expression modulating characteristic. We also applied the same QTL calculation pipeline to calculate sQTLs and identified ~10M sQTLs.[[[\*check number 10M?]]

For the cQTLs, the situation is more complicated. There are no established standard methods of calculating these on a large scale. To properly identify cQTLs, we focused on a reference set of enhancers to define the region associated with the activity of the chromatin and then look at how this activity varies in these enhancers across individuals and then we correlated this with nearby variants. (See methods). Overall, we were able to identify ~2000 cQTLs.

Furthermore, we are interested to see if any genotype is also associated with the single cell fractions. In particular, we used our QTL pipeline and identified 443 distinct SNVs whose genotypes are significantly associated with differential cell fractions across individuals; i.e., cell fraction QTLs (fQTLs). In total, the 443 distinct SNVs constitute 508 different fQTLs between different cell types. Significant fQTLs are those with associated Bonferroni-corrected p-values of no more than 0.05. Different cell types exhibit a great deal of heterogeneity in terms of their abundance within the set of high-confidence fQTLs. For instance, we identified 45, 15, and 33 significant fQTLs associated with the endothelial cells, astrocytes, and microglia, respectively, but there were no significant fQTLs that were found to be associated with oligodendrocytes. Moreover, we also identified xxx SNPs significantly associated with the gene expression changes across individual tissues unexplained by our single cell deconvolution; i.e., Y-WX (Methods).

Given the QTLs we identified, we overlap and annotate them with a variety of different genomic annotations and look at the degree that they overlapped. The distributions of detailed QTL annotations on genomic regions are shown in Figure xxx. As expected, there's a very large amount of overlap between the cQTLs, sQTLs, and eQTLs, and with ~50% of the cQTLs essentially being a subset of the eQTLs. [[[\*\*\* sQTLs too]] We examined the enrichment of most significant eQTLs per gene enhancers across XX human tissues and cell lines.[[\*\*\* they are enriched in briain?]]]   We calculate the enrichment of cis-QTLs on GWAS SNPs of brain related disorders (schizophrenia, bipolar disorders and parkinson’s disease) and non-brain related disorders (CAD, asthma and type 2 diabetes). Cis-QTLs have more significant enrichment for GWAS SNPs of brain disorders than the ones of non-brain disorders.

**VII. Gene regulatory networks in adult brain (523)**

In this section, we provided an integrative analysis at the gene regulation level for the data and genomic elements in the resource and predicted a gene regulatory network revealing how the genotype and regulators to control target gene expression in adult brain.

To this end, our first step is to process a full Hi-C data for adult brain, which provides direct physical evidence for potential interactions between enhancers and promoters in the format of topologically associated domains (TADs) (Figure 5A). Specifically, we generated and processed the Hi-C data for the same reference adult brain that was used to identify the brain active enhancers, using the protocol of XXX (Supplement). In total, we identified xxx TADs in adult brain. Overall, these TADs show a number of established properties, such as the gene expression tends to increase with increasing numbers of interactive enhancers (Figure 5xx). More importantly, we found that >xx% enhancer-promoter interactions happen in the same TADs in adult brain (Figure 5xx), suggesting that TADs potentially provide at large cis-regulatory relationships between enhancers and target genes.

Therefore, the second step to build the gene regulatory network is that we integrated the TADs with other regulatory relationships such as the enhancers, transcription factors (TFs), miRNAs, eQTLs to target genes in the resource (Methods). In particular, we used Hi-C data to find all possible enhancer-target gene relationships if enhancers and targets’ promoters are in the same TADs. We then found the TF binding motifs using ENCODE data and inferred the TF-target gene relationships if TFs have enriched binding motifs on the target gene’s promoters and enhancers. In total, we included xxx enhancer-gene, xxx TF-gene, xxx eQTL-gene and xxx miRNA-gene regulatory linkages, providing a reference wiring network on gene regulation in brain.

Finally, using these “wiring” regulatory relationships, we inferred the gene regulatory network that include the active regulatory relationships on how QTLs, enhancers, and transcription factors relate to target gene expression (Methods). In particular, given a target gene, we then associate coefficients with each of these wiring linkages to try to predict our target gene expression from the activities of their regulatory elements. We model them as simple linear relationships but regularize them to minimize the number of connections using the elastic net model (Methods). Overall, we found this model could predict expression successfully as shown from Figure xxx.[[[\*\*\* how do we show??]] We repeated this for all genes and found how various subgroups of QTLs affect gene expression; e.g., a significantly number of predictive QTLs break the TFBSs on the enhancers or promoters (xx%, Figure xxx). We thus constructed a gene regulatory network consisting of the QTLs, enhancers, TFs and target genes with high predictive relationships (Methods), revealing the biological mechanisms on how QTLs regulate the target gene expression in the adult brain. This network also has a few particular characteristics such as scale-free and hierarchical structures, which have been revealed by previous regulatory network analyses (Figure Sxx).

[[\*\*\*Overall our network can connect many previously unannotated brain associated GWAS SNPs to potential genes. XXX can be connected via eQTLs. YYY via HiC & ZZZ in total. (Fig QQQ).]]

[[[\* move to net sect]]]

**VIII. Integrative modeling to explain the molecular mechanisms for genotype-phenotype relationships in adult brain (821)**

The interaction between genotype and phenotype is a complex process, involving multiple intermediate stages including gene regulatory network. Thus, in this section, we perform another level of integrative analysis for the resource by embedding our gene regulatory network from the previous section into a larger model; i.e., we introduce an interpretable deep-learning framework, a Deep Structured Phenotype Network (DSPN), which provides insight into how the genomic variants link to the regulatory network, then to functional modules, and eventually predict phenotypes such as schizophrenia (Figure xxx). This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from the QTLs and gene regulatory connections predicted from our elastic net regression. In particular, it integrates all high dimensional functional data types in this resource including genomics, transcriptomics, epigenetics and regulation, and genotype-phenotype relationships, and also allows us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only. This is because it is not only feed information from the bottom of the network to the top i.e. variants all the way up to phenotypes, but also propagate information throughout the network, predicting for instance the transcriptome from genomic variants or directly predicting phenotype from the transcriptome. We also make the model downloadable as a set of simplified files encompassing the elastic net model described earlier plus additional DSPN connections between layers such as a groups of gene modules to phenotypes.

As shown in Figure xxx, traditional classification methods such as logistic regression predict the phenotype directly from genotype, missing the intermediate information such as transcriptome (Figure xx). We build the DSPN via a series of intermediate models which add layers of structure to a logistic model, including a layer for intermediate molecular phenotypes such as gene expression and chromatin state, multiple layers for functional modules and other mid-level phenotypes which may be inferred as hidden nodes in the network, and a layer for high-level phenotypes such as brain traits. 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 (Supplement)[[[\* repetitive para, condense with previous]]

Using the model, w

Moreover, the model also enables practical imputation of a subset of the transcriptome and epigenome, with an accuracy of ~70% (Figure xxx). After imputation, we can use the model to improve prediction of biological variables and psychiatric diseases by the addition of transcriptomic data to genotype, as compared to genotype alone. In particular, we can predict schizophrenia with an accuracy of 61% using our model and an imputed transcriptome compared to 56% with genotype alone. This result demonstrates the usefulness of even a limited amount of functional genomics information for unraveling gene-disease relationships.

We transform the results above to the liability scale in order to compare with heritability estimated on this scale using GCTA (Figure xxx). Using the PEC 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%), suggesting this model is near optimal (although there may be further epistatic interactions the model can capture). The full DSPN model estimates that the transcriptome-based liability for the PFC is ~19.2%. Although we expect that a large portion of this will overlap with the common SNP based liability, it may also include environmental and epistatic contributions (see Supplemental Figure), precluding direct comparison. Similar estimates of the liability explained for Bipolar and Autism by the DSPN (imputation and full models) are given (Figure xxx).

[[[\* moved down]]We examined the connections learnt by the DSPN between intermediate and high-level phenotypes for potential mechanisms to see if they are biologically meaningful. For example, the module xxx is connected to genes enriched in the dopaminergic and glutamatergic synapse (GSEA enrichment score > xxx, Figure xx), and the module yyy is connecting to Age, and represents the neuronal cell fractions (Figure xxx). Furthermore, we used this model to recapitulate the pathways comprising the cross-layer nodes and predictive edges for particular phenotypes. For example, as highlighted in Figure xxx, the schizophrenia (SCZ) trait is activated by two modules on the layer of hidden nodes corresponding to glutamatergic signaling and excitatory synapse, respectively. The modules are connected by a set of genes including GRIN1, which are regulated by corresponding QTLs (e.g., rs1146020) and enhancers (e.g., GH09H137166) as shown in the blowup gene regulatory mechanism. In addition, we also found some suggestive connections for SCZ such as module(s) corresponding to dopamine-related pathways and complement pathways (Figure xxx). These modules are connected to the C4 family genes, regulated by eQTLs and enhancers (p<1e-4). The complete functional annotations of modules are available in supplement.

**IX. Discussion (558)**

We integrated PsychENCODE datasets and other projects from 1931 individuals, and developed a comprehensive resource consisting of various functional genomic elements for the adult brain. This resource serves as an important step for gaining biological insights from genomic functions in neuroscience. Overall, our study has identified a very large-scale set of eQTLs and eGenes for adult brain (Figure xx), almost achieving saturation. Therefore, we suspect that the future larger population studies would not be very helpful to this context. However, there exist other aspects of brain QTLs that can be extended in the future, in addition to eQTLs. The first would be chromatin QTLs for adult brain, which is currently much less than eQTLs in the resource. Increasing sample size such as large population potentially helps identify more cQTLs, which also can be further interrelated to eQTLs and other regulatory variants from our deep learning model. Moreover, the enhancers that this study used for cQTLs are defined from the current techniques such as ATAC-seq and ChIP-seq signals, especially from K27AC. The future and new state of the art methods available such as STARR-seq provide more accurate definitions on enhancers, and thus should be further used to better identify such as chromatin associated variants.

Another aspect that might move forward is single cell analysis. The current single cell techniques suffer from the low capture efficiency, so remain challenging to reliably quantify the low-abundant transcripts/genes and interrogate the biological variations [refs]. In this study, we thus integrated recent single cell data including thousands of neuronal and non-neuronal cells along with almost 1000 PEC single cells mainly consisting of fetal cells and found that these basic and known cells could explain large expression variations across tissues. However, increasing single cell data and more advanced techniques in the future will identify considerably large number of novel cell types, which might contribute to unexplained variations. Using these additional single cell data, our deconvolution analysis expects to estimate more complete cell populations and accurate fQTLs to brain tissues.[[\*\*\* but mention fundemental problem w/ brain rna]]

More accurate cQTLs and fQTLs can be input into our deelp learning model, which might also improve the model performance. Our model represents a state of the art method at the moment to reveal genotype-phenotype at the population level but might improve with the development of machine learning and additional data types such as image and medicine. Furthermore, while providing better prediction, some model connections are deliberately set to be interpreted simplifications, such as gene regulatory networks, to make the model more interpretable and easier to use. Thus, another major goal of the model is to provide a useful compression of large functional genomic datasets for brain; e.g., XXX KB of model files vs. XXX TB of total resource data.