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

**I. Introduction (543)**

Disorders of the brain affect nearly 20% 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 disorder of the genomic (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 of review). Unfortunately, for most of these locations, we have little to no understanding of which base pairs alterations constitute the functional genomic alteration, which transcripts and networks are altered, and what are the molecular mechanisms that cause those alterations. It is presumed that these changes in transcription modify the proteome, which leads to changes in brain structure and function, and these changes interact with environmental factors to change the probability of developing a brain disorder.

To this end, a variety of genomic elements have been found to associate with brain and psychiatric disorders, including genomic variants and genes found by many studies. For example, 108 GWAS loci and 693 differentially expressed genes associated with schizophrenia identified by Psychiatric Genomics and CommonMind consortia. Also, other consortia such as GTEx, ENCODE and Epigenomics Roadmap have generated large-scale RNA-seq and ChIP-seq data for dozens of brain tissues and cell lines to systematically identify brain specific genes, transcripts and regulatory elements (N=xxx). Moreover, recent studies show the specific chromatin structure and activity of the regulatory elements such as brain active enhancers [ref], and single cell techniques can detect gene expression and epigenetic patterns for neuronal and non-neuronal cells from brain tissues [ref]. However, these results still suggested that thousands of samples would be required to achieve statistical power of 0.8 for detecting a complete set of brain-related genomic elements [refs]. Also, individual molecules do not independently affect brain, and instead interact with each other in a network. Thus, effort is also needed to model and analyze the molecular interactions and mechanisms that drive the brain phenotypes and psychiatric disorders.

In fact, understanding the mechanisms on how these genomic elements affect various brain functions and phenotypes is still a key challenge in neuroscience. To address it, the PsychENCODE Consortium (PEC) has generated and assembled a robust 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/pec/) 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 and activities for adult brain on the genome scale. We also combined these elements and built an integrated deep-learning model to impute missing data and reveal the mechanisms about how they interact to drive the brain phenotypes and psychiatric disorders.

## II. Comprehensive resource for adult brain functional genomics (334)

We built this comprehensive resource to have a coherent data structure. Broadly, it organizes a large amount of data for brain functional genomics in a pyramid (or hierarchical?) shape (Figure 1). The bottom includes the largest scale data with often controlled access such as individual genotyping and raw next generation sequencing data of transcriptomics and epigenomics. It is followed by the uniformly processed and summarized data from the bottom such as open chromatin peaks and gene expression quantifications. Derived from these data, the middle part then includes the brain related genomic elements and interactions such as QTLs, enhancers and gene regulatory networks. Finally, at the top, the resource contains an intuitive and interpretive model revealing how the genomic elements interact to affect brain functions and phenotypes.

In terms of the [[bottom level]] 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 the data from other relevant 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. In particular, we used the annotations from reference brain project to define the brain related genomic elements, and the ENCODE standard pipelines to uniformly process all raw next generation sequencing data for both bulk and single cell data to find their activities.

[[cut]]

Moreover, we further analyzed these data and systematically identified the functional genomic elements for adult brain; i.e., the derived data including the brain-active enhancers, differentially expressed genes and transcripts, and QTLs associated with various phenotypes. In addition, our analyses and model revealed the interactions among these brain genomic elements such as imputed gene regulatory networks in which enhancer to gene linkages are identified by HiC data, and the single cell fractions of individual tissues for both neuronal and non-neuronal cells.

**III. Bulk and single cell transcriptome analysis and deconvolution explain gene expression and cell fraction changes (714)**

Given the large-scale bulk transcriptomic data in resource, 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 available samples from PEC and GTEx to quantify the expression levels for the protein coding genes, transcripts, noncoding RNA and novel transcribed regions of brain and other tissues. 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.[[consolidate?]] 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.

The 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 in our brain tissue samples is whether the changes are driven by gene expression in a particular cell type or different cell-type populations. 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 create a list of basic and primary cell types in the brain; i.e., 16 neuronal types, five non-neuronal types and xxx additional fetal types from PsychENCODE (Supplement). These 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 and so forth; 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 used an unsupervised analysis for the bulk tissue expression data and try to find its main components that potentially relate to the single cell types. In particular, we decomposed the decomposed the bulk gene expression matrix from our resource using non-negative matrix factorization (NMF, see Methods), and compared if top principal components of NMF (NMF-PCs capturing most data covariance) and the gene expression of single cells are consistent. As shown in Figure XX, we can see a number of NMF-PCs highly correlate with the biomarker 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 suggests that these cell types do make sense and contribute bulk tissue gene expression.

After this analysis, we further devolved the bulk tissue expression matrix Y [[B??]]using the single cell data matrix X to estimate the cell fractions W, by solving the equation “Y=WX” (See methods). We found that the multiplication of estimated cell fractions and single cell expression data can explain [[the pop. Variation]] of expression across tissue samples; i.e., [[expl \_F]]||WX||\_F/||Y||\_F>80% showing 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). 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 types (EX3 and In6) have significantly different fractions between healthy Male and Female. The fraction(s) of neuronal type(s) (Inhibitory X) is significantly anti-correlated with Age (r = xxx), and Inhibitory X cells have functions of XXX involving the differentially expressed genes in Age from our resource (Figure xxx). [[consolidate]]The excitatory neuronal cell populations (e.g., EX3) decrease significantly in ASD samples (p<xxx) while the non-neuronal cells increasing (e.g., oligodendrocytes). Finally, we report the individual cell populations along with significantly associated relationships between particular cell type fractions and phenotypes (Supplement).

**VI. Active enhancers in adult brain (215)**

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 data sets in adult brain.

First, we developed a consistent set of brain active enhancers. In particular, we processed the H3K27ac and H3K4me3 ChIP-seq and ATAC-seq data of the reference brain using the standard ENCODE ChIP-seq processing pipeline. 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.

**V. Consistently comparative analysis reveals the brain related transcriptomic and epigenomic activity (463)**

One key aspect of our analysis is that we, as consistently as possible, processed the transcriptomic and epigenomic data across PEC, 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 and t-SNE are two popular techniques, but we found the former tended to be overly influenced by data outliers, and t-SNE tends to just uniformly separate all the clusters and not give one a sense of the overall effect. [[can we be more objective??]] Therefore, we found another very useful technique to be Reference Component Analysis (RCA), which 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 of RCA, 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 enhancers. It shows that the chromatin levels are indistinguishable between brain and other tissues (Figure xxx). Thus, the gene expression difference, in a sense, in brain cannot be simply attributed to that chromatin, but potentially be driven by more complex gene regulatory mechanisms involving enhancers.[[too strong]] In addition, the expression differences that we're looking at are those confined to known canonical regions such as protein-coding genes, however people have previously remarked about the tremendous amount of transcriptional [[discuss]]diversity in intergenic and noncoding regions. Thus, we tried to get a sense of this looking at the overall level of transcript diversity across the entire genome. In terms of protein-coding genes, it has been previously known that testes tends to have the most transcriptional diversity (Figure SYYY sat’d for genes). However, when we shift to non-coding and unannotated regions, we find that brain does stand out to some degree in having more transcriptional diversity than most other tissues. This transcriptional diversity tends to increase with the number of samples (Figure xxx sat’d).

**VI. QTL analysis (598)**

To understand how the genotype affects the transcriptomic and epigenetic activities in adult brain, we used the resource data to identify more interpreted association relationship data such as the 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) 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 and emphasized the scale of the database. We adhered closely to the established GTEX eQTL pipelines. We identified ~2M of eQTLs and ~17000 number of 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 great 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. For the cQTLs, the situation is more complicated. There are no established standard methods of calculating cQTLs on a large scale. In order to properly identify cQTLs, we have to both define the region associated with the activity of the chromatin and then look at how this activity varies. [[really???]] We did joint K27 peak calling over the hundreds of brains in PsychENCODE with H3K27ac marker. From the consensus regions, we calculated an average signal value of the chromatin marker. And then we correlated this with nearby variants. (See methods). Overall, we were able to identify ~2000 chromatin QTLs. 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 xxx SNPs whose genotypes are significantly associated with both neuronal and non-neuronal cell fractions across individuals; i.e., cell fraction QTLs (fQTLs). For example, fQTL xxx is for Ex3 fraction. This suggests that these fQTLs potentially can be used to predict the cell fractions in adult brain. 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). These SNPs are likely causing certain gene expression changes driven by unknown cell types in adult brain.

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 and eQTLs, and with most of the cQTLs essentially being a subset of the eQTLs. [[contradict??]] We found that XX% of cQTLs are overlapped with eQTLs.We examined the enrichment of most significant eQTLs per gene in Roadmap Epigenomics Consortium and ENCODE enhancers across XX human tissues and cell lines. Collectively, these QTLs annotate a larger fraction of GWAS SNPs involving the brain (e.g., 6% in schizophrenia, 10% in bipolar) than previously observed,[[for 2m eqtl snps??]] providing leads on which genes are affected in disease. We also 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 related disorders than the ones of non-brain related disorders.

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

Gene regulation is a key mechanism that genotype affects phenotype. This comprehensive resource thus also enables us to identify the gene regulatory relationships among the brain genomic elements and predict the gene regulatory networks revealing how the genotype changes the expression and chromatin to regulate gene expression. To this end, we first used the HiC data to identify the chromatin interactomics. Specifically, we generated and processed the Hi-C data for adult brain including three reference brain samples and identified the xxx detectable promotor-enhancer interactions for XXXX genes. Using this full Hi-C data for adult brain, we identified xxx Topologically Associating Domains (TADs) of adult brain. [[more]]These TADs provide the regions at which the enhancers interact with target gene promoters in adult brain and enable the systems identification of potential cis-regulatory enhancers of the genes. [more from HJ&DH]. [[more]]Also, the co-expressed genes are highly likely co-regulated by similar mechanisms such as genes in same TAD. Thus, we also constructed the gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods]. [[trans]??]]

To construct the gene regulatory network, we also integrated and imputated the regulatory relationships in brain such as the enhancers, transcription factors (TFs), miRNAs and target genes in this resource (Methods). For example, we used known enhancer-target genes [JEME] and 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 regulatory regions such as promoters and enhancers. We also used Hi-C data to filter the enhancers that are not in the TAD regions for given target genes. In total, we included xxx enhancer-gene, xxx TF-gene, and xxx miRNA-gene regulatory linkages, providing a reference wiring network on gene regulation in brain. It should be noted that activations of these regulatory connections are highly attributed to the genotypes of QTLs, leading to various phenotypes. Thus, using these “wiring” regulatory relationships, we inferred the gene regulatory networks that identify the regulatory relationships on how QTLs, enhancers, and transcription factors relate to target gene expression (Methods). In particular, given a target gene, we found its related regulatory elements from the resource including the eQTLs, the enhancers that control its gene expression plus their cQTLs, and predicted the transcription factors (TFs) that have enriched binding sites on these enhancers and its promoter. We then used RNA-seq and ChIP-seq data based on the Elastic Net model with regularization that combines the L1 and L2 penalties of the lasso and ridge regressions to predict the regression coefficients of genotypes of various QTLs, the chromatin stages of enhancers, splicing patterns and TFs gene expression to the target gene expression, and identified the highly predictive relationships (i.e., large coefficients). [[be more intuitive]] 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 networks 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.

**VIII. Integrative modeling to explain the molecular mechanisms for genotype-phenotype relationships in adult brain (613)**

The interaction between genotype and phenotype is a complex process, involving multiple intermediate stages including gene regulatory network. We thus introduce an interpretable deep-learning framework, Deep Structured Phenotype Networks (DSPN), which provides insight into how the brain genomic variants affect gene expression and regulation, and eventually predict phenotypes (Figure xxx). This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from the QTLs and regulatory networks estimated in our resource. [[how related to elastic]] On the resource website, we provide a list of DSPN pathways for each phenotype and disease. We also make the model downloadable as a set of simplified files summarizing represented genotype-phenotype pathways. In particular, this model 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. The model is trained as a deep generative model to represent the conditional distribution of all variables given the genotype. Unlike a feed-forward network, this architecture allows information to flow in top-down, bottom-up and lateral directions during inference.

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)

We examined the connections learnt by the DSPN between intermediate and high-level phenotypes for potential mechanisms. 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 potentially additional molecular mechanisms 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).[[supp??]]

Moreover, the model also enables practical imputation of a subset of the transcriptome and epigenome, with an accuracy of ~70% (Figure xxx). We 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 bipolar disease and schizophrenia with much higher accuracy from the transcriptome than from genotype alone; i.e., three times improvements (+18% vs. +6%) from the random prediction 50% for schizophrenia, Figure XXX). The imputed transcriptome also clearly adds predictive value, as 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.

**IX. Discussion (506)**

We integrated the high-dimensional brain genomic datasets of PsychENCODE and other projects from 1931 individuals, and developed this comprehensive resource consisting of various functional genomic elements for the adult brain. This resource serves as an important step in gaining biological insights from genomic functions and mechanisms in neuroscience. In particular, our comparative analyses found that these genomic elements significantly relate with the psychiatric disorders and other brain phenotypes including developmental stages [cap2]. The neuroscientists can use this resource as a reference to compare with their data, generate hypotheses and help design experimental validations. In addition, this resource is publicly available online and can be extendable and scalable to integrate additional data types and phenotypes in brain such as individual’s fMRI image features measuring functional neuro-connectivity to identify the associated genotypes such as image-QTLs (iQTLs) [xx]. Also, it can incorporate with the neurodegenerative diseases like Alzheimer and Parkinson.

Moreover, by combining the resource data, we built an integrative deep learning model, DSPN to reveal the interactions and mechanisms among various high-dimensional functional genomic elements from a number of directions between genotype and phenotype. In particular, this model also incorporates the derived data types into its hierarchical structure such as imputed gene regulatory networks and QTLs and provides the additional statistical powers to better predict phenotype. It is also available online as a general-purpose tool and enables quantitatively imputing missing transcriptional and epigenetic information for samples with genotypes only. Also, the model can be used to prediction the outcomes of in-silico perturbations; e.g., knocking down GRIN1 potentially breaks the excitatory and glutamatergic signaling pathways to likely affect schizophrenia. 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 compression of large functional genomic datasets for brain; e.g., XXX KB of model files vs. XXX TB of total resource data, beyond a purely predictive network from genotype to phenotype.

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]. However, it is still worthwhile using the biomarker genes with strong expression signals in single cell to deconvolve the tissue gene expression data to find the cell fractions for individual tissues and relate to the individual phenotypes. With increasing amount of single cell data in near future, we could deconvolve the tissue data in the resource to find potential new cell types and obtain more complete cell populations. Furthermore, the limited amount of RNA molecules in single cell makes it even harder to capture the weak signals, which makes the data sensitive to technical noise. Thus, given that the RNA decaying issues in single cell RNA-seq, we could also relate this resource to recent in situ transcriptomic data such as the spatial gene expression by optogenetic techniques, and find the consistent expressed genes driving the brain phenotypes at the cellular and tissue levels.