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


## Abstract

Understanding how genomic variation influences associated phenotypes in the brain remains a key challenge in neuroscience, one where the potential of functional genomic approaches has not yet been fully realized. To this end, we developed a comprehensive, population-level resource that includes ~2,000 samples processed by the PsychENCODE consortium for healthy controls and neuropsychiatric disorders. Available online, the resource comprises genotyping, RNA-seq, ChIP-seq, and single-cell data, in addition to analytic summaries of quantitative trait loci (>5,000,000 expression QTLs and >5,000 chromatin QTLs), brain-active enhancers, differentially expressed genes and transcripts, and novel non-coding RNAs. Leveraging and comparing this resource with other data, we show that the brain has distinct expression and epigenetic profiles as evident from spectral analysis and more non-coding transcription from most other tissues. Also, using single cell data, we deconvoled the tissue-level gene expression of this resource to find the populations of different cell types corresponding to particular phenotypes. Finally, we developed and built an integrative epigenome- and transcriptome-wide association model (eTWAS) to predict the brain phenotypes using high-dimensional functional genomics data with genotype-phenotype associations in this resource to highlight key brain genes and modules and relate the mechanisms on how variants in these affect gene expression. This model allows us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only.  This model also shows that the integrated data has significantly improved the prediction accuracy over individual genomic data types and relates these predictions to well characterized functions and pathways in the brain. Therefore, developing this PsychENCODE resource and integrated model to a population-level scale serves as an important step in gaining meaningful biological insights from functional genomics studies in neuroscience.


## Introduction

The brain is the most complex organ in adult human. A variety of individual molecules have been found to associate with brain phenotypes including mental diseases. For example, GWAS has identified XXX SNPs significantly associated with SCZ. Also, a number of genes have been reported to have specific activities for mental disease. For example, the differentially expressed genes were found for SCZ (n=xxx), BP (n=xxx), ASD (n=xxx). Therefore, recent studies like GTEx, ENCODE and Epigenomics Roadmap have generated large-scale RNA-seq and ChIP-seq data for brain tissues and cell lines, trying to systematically detect the brain specific genes, transcripts and regulatory elements. However, these studies only focused on the normal healthy brains, so their data is unable to find the specific genomic elements for additional phenotypes

especially for mental health. The CommonMind Consortium and others have provided the gene expression and genotyping data for both healthy and disease brains such as SCZ for xxx samples [refs]. Given that the complexity of brain samples of mental diseases, this sample size appears too limited to discover a complete set of genomic elements for mental diseases or other phenotypes. Moreover, the adult brain phenotypes are highly likely driven by interactions among various molecules, rather than individual molecules. Thus, effort is needed to model and analyze the molecular interactions that drive the brain phenotypes.

Also, understanding the molecular 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 integrates a group of projects to produce a public resource of multi-dimensional genomic data from thousands of high quality healthy and diseased human post-mortem brains (6). Particularly, it has generated and assembled a robust large-scale dataset on the adult human brain to address this challenge, including genotyping, RNA-seq, ChIP-seq and single-cell transcriptomic data on ~2000 brain tissue samples with different phenotypes. The rich data generated by the PsychENCODE Consortium are a preeminent resource for studying regulatory mechanisms in the human brain [1]. One of its unique aspects is the coverage of major psychiatric diseases, such as autism spectrum disorder (ASD) and schizophrenia (SCZ). PsychENCODE datasets have been assembled by many investigators over several years, and they are housed in a central depository (www.synapse.org) and shared with the public. Integration of these multi-dimensional and large-scale datasets potentially benefits understanding the molecular mechanisms for brain phenotypes, which however still remains the bioinformatics challenge.

In this paper, we integrated the PsychENCODE datasets over all ~2000 samples, compared them against various brain phenotypes and merged with other brain genomic data from the ENCODE, GTEx, Brainspan and Epigenomics Roadmap projects to develop a comprehensive and online available resource for the adult brain. This resource comprises all possible functional genomic elements including the brain-active enhancers, transcripts, expression models, imputed regulatory networks, eQTLs and cQTLs for various phenotypes. We then analyzed this resource and found the specific genomic and transcriptomic activities on genome wide in brain including gene expression, non-coding transcription and enhancers. Finally, we developed and built an integrative model to reveal how the interactions among genomic variants, gene expression, enhancers and phenotypes, trying to explain the molecular mechanisms from genotypes to brain phenotypes.

## Comprehensive resource for adult brain functional genomics

To systematically understand the molecular functions and mechanisms how genomic variants affect associated phenotypes in the brain, we need to find the related molecules that have specific activities for the brain phenotypes. Therefore, the PsychENCODE consortium has generated and assembled a robust large-scale dataset on the adult human brain, including genotyping, RNA-seq, ChIP-seq and single-cell transcriptomic data on ~2000 individual brain tissues with different phenotypes including mental diseases (Assay summary in Methods). We

uniformly processed and integrated this dataset with complementary genomic information from other large consortia, particular from ENCODE, GTEx and Epigenomics Roadmap to develop a comprehensive resource for the brain functional molecules across genomic, transcriptomic, epigenomic and regulatomic levels (Methods). We also compared the resource data against various phenotypes, and identified the brain specific elements. For example, this resource includes the regulatory variants such as QTLs, brain active enhancers, differentially expressed genes and transcripts, novel transcribed regions and non-coding RNAs, and putative genome-wide regulatory networks. It is also publicly accessible and available on the PyschENCODE website (xxxx), and can be used as interactive web tool. As shown in Figure 1, the resource comprises the following major data types:

*QTLs* - we merged genotype and gene expression data of Brain DFC region from a number of studies relating to PyschENCODE. We calculated the association of imputed SNPs with normalized gene expression and chromatin states (Methods) to find expression QTLs and chromatin QTLs respectively using an additive linear model from QTLtools. This linear model was also adjusted by covariates like PEER factors of gene expression, genotype PCs and disease diagnosis. Among these SNPs, we identified a great number of the regulatory variants significantly associated with brain transcriptional and epigenomic activity: >5 million expression QTLs (eQTLs) and >5 thousand chromatin QTLs (cQTLs) for histone modification signals. The number of eQTLs in this resource is significantly greater than previous studies, approaching the saturation of human mutations (Figure xxx). We also showed that the eQTLs number can be predicted from the sample size using a fitted curve (Figure xxx).

*Epigenomics* - we used ChIP-seq data in PsychENCODE to discover xxx open chromatin regions covering xxx enhancers (K27). We then related them with ENCODE and Epigenomics Roadmap data, and summarize a list of xxx PsychENCODE brain enhancers, mainly active on DLPFC and CBC (Supplement).

*Transcriptomics* - we uniformly processed the RNA-seq data from a number of PsychENCODE-relate studies, ENCODE and GTEx, and found the xxx genes and transcripts that express in brain samples,and xxx eGenes associated with eQTLs (Methods). In addition, we discovered xxx non-coding RNAs and novel transcribed regions in brain. Also, we compared them against different phenotypes and derived the phenotype-specific genes and transcripts.

*Regulatomics* - we also integrated and imputated the regulatory relationships in brain such as the enhancers, transcription factors (TFs), miRNAs and target genes [refs] in this resource (Methods). 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. The activations of various wires may change across phenotypes. In addition, we link the QTLs that overlap the enhancers and promoters in the resource to reveal the potential regulatory activities such as QTLs break TF binding sites.

*Phenotypes* - the PsychENCODE data covers a number of phenotypes on mental health. They are normal control (n=xxx), SCZ (n=xxx), BP (n=xxx), ASD (n=xxx), Male (n=xxx), Female

(n=xxx), Age (distribution), etc. (Supplement). This resource also links the functional genomic elements and particular phenotypes such as differentially expressed genes for SCZ using the analysis and modeling that we would discuss in next sections.

The establishment of this comprehensive resource enables the modeling and analysis for the biological processes that drive the brain phenotypes to eventually understand the molecular mechanisms between genotypes and phenotypes. Therefore, we analyzed and modeled the data from this resource to further reveal the brain specific genomic and transcriptomic elements, and the biological mechanisms explaining how these brain elements affect the phenotypes in the adult brain.

## System identification of brain specific genomic and transcriptomics activity via comparative analysis

This comprehensive resource allows us to discover the specific functional genomic elements that relate the brain phenotypes. Thus, we leveraged this resource against various phenotypes and compared with other tissue types to reveal the unique brain genomic activities, particularly relating to transcriptomic and regulatory binding activities. In particular, we first performed the spectral analysis for comparing the similarities of gene expression other tissue samples from GTEx (Figure xxx). It shows that the brain samples, though from different studies are clustered together in a major cluster, significantly separated from the other major cluster consisting of non-brain samples. This suggests that there exist the brain has unique and distinct gene expression programs, involved by the brain elements in our resource that make brain very different from other tissues. In addition, this major brain cluster can be further subdivided into several clusters, each of which mainly comprises the samples from same brain region; e.g., the cortex and cerebellum clusters in Figure xxx. Additionally, we found that the brain has more transcriptional activities at the non-coding and novel transcribed regions than most other tissues (Figure xxx), which implies that the non-coding transcription is highly likely another factor to make the brain tissues unique.

To systematically find the specific expressed functional elements in brain, we identified the differentially expressed genes and non-coding RNAs for various phenotypes including mental disease, gender, regions (Methods and Table XXX) for the resource. For example, XXX genes have been found to differentially express between SCZ and healthy samples. We also checked the enriched pathways and functions among the SCZ genes, and indeed found that many are relating to SCZ. Moreover, we also found that these brain dex genes are significantly less/greater than DEX genes for other tissues in GTEX (p<xxx), which suggesting that the brain expression uniqueness is highly driven by a small/large set of genes. We report the DEX genes for all phenotypes in our resource along with their enriched functions and pathways in supplement. Also, the brain specific gene expression is likely driven by a group of genes, rather than individual genes, so we constructed the gene co-expression network using all PsychENCODE and GTEx samples, and clustered it into gene co-expression modules using WGCNA [Methods]. The genes clustered in a same module are highly likely co-regulated by similar mechanisms. Our co-expression analysis indeed found several modules whose

eigengenes show very different expression levels between brain and non-brain samples (Figure xxx, Supplement), which suggests that there exist brain specific regulatory mechanisms drive these brain co-expression modules.

Therefore, we are further interested to compare the regulatory regions between brain and other tissues to see any brain specific regulatory activities. We performed the spectral analysis to compare the similarities of epigenetic profiles of PsychENCODE samples with Epigenomics Roadmap data. It is also interesting to somewhat similar patterns with the gene expression comparison; e.g., the brain samples can also cluster together in terms of active enhancer similarity (Figure xxx). This result suggests that the brain has specific and distinct epigenomic activities as well, involving the brain active enhancers from our resource. Furthermore, our resource includes a great number of regulatory variants significantly associated with brain transcriptional and epigenomic activity: >5 million expression QTL for gene expression and >5 thousand chromatin QTL for histone modification signals. We also compared them with existing QTLs databases. We found that these variants cover a larger fraction of disease-associated brain GWAS SNPs than any previous analyses, suggesting potential molecular targets for these associations (xx% for SCZ, xx% for BP, ASD,). We also evaluated the overlap of eQTLs with cQTLs and found that XX% of cQTLs are overlapped with eQTLs. The SNPs in cis-eQTL list(Cis-eSNPs) were enriched within XXXX, and depleted XXXXXX (Fig. X). We examined the enrichment of most significant eQTLs per gene in Roadmap Epigenomics Consortium and ENCODE enhancers across XX human tissues and cell lines. Cis-eQTL were enriched for enhancer sequences present in brain tissues and the strongest enrichment is observed in DLPFC enhancers. 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.

## Single cell analysis and deconvolution explain gene expression changes across adult phenotypes **[[should we move up??]]**

One issue with the changes of gene expression in our brain tissue samples is whether the changes are driven by a particular cell type or different cell-type populations. To address this tissue, we integrated the single cell gene expression data to discover the expression changes of brain tissue genes across various cell types including both neuronal and non-neuronal. Furthermore, deconvolved the gene expression data of individual tissues over both novel and known cell types to find the cell populations for individuals, and relate to the individual phenotypes. We found that the gene expression changes across adult brain phenotypes at the tissue level can more easily be explained by the changes of cell populations.

First, we integrated the single cell RNA-seq data for ~900 cells from PsychENCOCE, ~3000 neuronal cells with 8 excitatory and 8 inhibitory types from Lake's 2016 paper, and ~400 cells including 5 non-neuronal types, astrocytes, endothelial, microglia, oligodendrocytes and OPC. We then compared these single cells based on the (biomarker) gene expression similarity using

tSNE, and found that the same-type cells generally can be clustered together (Figure xxx). In particular, xx% PsychENCODE cells have been found to cluster together with known cell types (xx% neuronal, xx% non-neuronal, details in supplement). In addition, xx% PsychENCODE cells form their own clusters, away from known cell types, suggesting that the potential novel cell types found by PyschENCODE for brain tissues.

We further checked the expression changes across these single cells for the brain genes in the resource, and found that a group of brain genes show the expression dynamic changes among cells. For example, the SCZ gene, XXX is (or ww% of SCZ genes) significantly more highly expressed in YYY and ZZZ neuronal cells than others (Figure xxx), suggesting that YYY and ZZZ drive the SCZ gene expression changes at the tissue level [ref]. In contrast, we also found that a number of brain genes don't show expression changes across cell types, which implies that their expression changes at the tissue level are potentially explained by the cell populations. Therefore, we deconvolved the tissue-level gene expression data of all 2000 samples using single-cell data to find the populations of different cell types corresponding to different phenotypes (Y=WX, Methods) (using all genes or brain/biomarker genes). The single cells used in deconvolution cover all 16 neuronal types, five non-neuronal types and xxx additional PsychENCODE types. We found the gene expression differences at the tissue level can be largely explained by cell population changes (p<xxx, xx% covariances) (Figure xxx covariance). In addition, we used the heatmaps (Figure xxx) to display the cell populations of individuals across different phenotypes. We found that there exist a number of cell population changes that highly associate with brain phenotypes. For example, the fraction(s) of neuronal type(s) (Inhibitory X) is significantly anti-correlated with Age (r = xxx). The non-neuronal cell populations increase significantly in SCZ (or Male) samples (p<xxx) while the neuronal cells decreasing. Finally, we report the individual cell populations along with significantly associated relationships between particular cell type fractions and phenotypes (Supplement).

## Integrative modeling to explain the molecular mechanisms for genotype-phenotype relationships in adult brain

Finally, we built an integrative model to understand how the brain genomic variants affect gene expression and regulation, and eventually predict the phenotypes (Figure xxx). This model integrated all high dimensional functional data types in this resource including genomics, transcriptomics, epigenetics and regulatomics, and genotype-phenotype relationships. This model also allowed us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only.

We first inferred the gene regulatory networks that identify the regulatory connectivities 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 [JEME] 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 to predict the target gene expression from genotypes of eQTLs and cQTLs, the

chromatin stages of enhancers, and TFs gene expression using the resource samples, and identified the highly predictive relationships (i.e., large coefficients). We repeated this for all genes and found that 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 (coeff. > xxx, Methods), revealing the biological mechanisms on how QTLs regulate the target gene expression in the adult brain.

The interactions between genotypes and phenotypes is a very complex process experiencing multiple intermediate stages including gene expression, signaling, modulation and so on. Thus, to understand the entire processes how genotypes and phenotypes affect to each other, we then built a Deep Boltzmann Machine-based eTWAS model that directly embeds regulatory network information to predict genotype-phenotype associations. Specifically, eTWAS uses the undirected edges rather than feed-forward directed edges in that the phenotypes potentially impact back to the intermediate stages like gene expression. As shown in Figure xxx, the eTWAS consists of four layers: 1) genotypes such as QTLs; 2) gene expression and enhancers; 3) intermediate modules and 4) phenotypes such as brain traits, and provides the additively predictive relationships between layer nodes. We also allows the nodes on Layer 2 have connections based on the gene regulatory network. In particular, many intermediate-layer modules (i.e., strongly predictive features on Layer 3) that correspond to known gene sets associated with well-characterized pathways and functions in the brain; e.g., the module xxx is connecting to the genes enriched with ZZZ pathways (p<xxxx). Also, some modules are used to capture the information on single cell populations; e.g., the module yyy is connecting to Age, and represents the neuronal cell populations. We show that this integrated model has significantly improved the prediction accuracy over individual genomic data types. For example, its AUC/MSE for classifying SCZ and health samples is xxx beating other classification methods using gene expression only (Table 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 Autism is activated by two modules, x, and y corresponding to dopamine-related pathways and neuronal cell fractions, respectively. Each module is connected by a set of genes, which are regulated by corresponding QTLs and enhancers as shown in blowup gene regulatory mechanism. For each phenotype, we also provide a list of such eTWAS pathways on resource websites. At the other hand, the model can be used to make in-silico predictions for the perturbation outcomes. For example, we can knockdown the genes connecting to the module x to deactivate Autism. In addition, this model also allows us to quantitatively impute the missing transcriptional and epigenetic information by inputting given genotype data only. We also make the model available as a set of distributive software from the resource.

## Discussion

We integrated the genomic, transcriptomic and regulatomic PsychENCODE datasets from ~2000 samples and developed this comprehensive resource consisting of various functional

genomic elements for the adult brain. In particular, we compared it with other tissues such as GTEx data and identified the genotypes and QTLs, the specific expressed genes, transcripts and noncoding RNAs, active chromatin regions, the regulatory networks that significantly relate with different brain phenotypes at both cellular and tissue levels. For example, the QTLs allow one to potentially interpret most of the known brain-associated GWAS SNPs in terms of perturbations to specific genes. Thus, the neuroscientist 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. For example, it can integrate the clinical data like fMRI images measuring neuroconnectivties, and identify the functional genomic elements for the neurodegenerative diseases like Alzheimer or developmental stages.

Moreover, we built an integrative epigenome- and transcriptome-wide association model (eTWAS). This model allows us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only. More importantly, it integrates high-dimensional functional genomics data with genotype-phenotype associations to highlight key brain genes and modules and relate how variants in these regulate gene expression. This integrative model is also available online as a general purpose platform. The users can apply it to impute missing data , predict the genotype-phenotype relationships, and reveal potentially novel gene regulatory mechanisms and modules for additional phenotypes.

With increasing amount of single cell data in near future, we could deconvolve the resource data at tissue level to find potential new cell types and obtain more complete cell populations. [[Xs - bullets why tissue is still relevant w/ single cell , dendritic rna]] . Given that the RNA decaying issues in single cell RNA-seq, we could also relate this resource to the in situ transcriptomic data such as optogenetic techniques measuring the spatial gene expression, and find the consistent expressed gene for the brain phenotypes at the tissue level.

# References


1. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, Shah HR *et al*: **Gene expression elucidates functional impact of polygenic risk for schizophrenia**. *Nat Neurosci* 2016, **19**(11):1442-1453.
2. Consortium GT: **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans**. *Science* 2015, **348**(6235):648-660.
3. Psych EC, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S *et al*: **The PsychENCODE project**. *Nat Neurosci* 2015, **18**(12):1707-1712.
4. Neale BM, Sklar P: **Genetic analysis of schizophrenia and bipolar disorder reveals polygenicity but also suggests new directions for molecular interrogation**. *Curr Opin Neurobiol* 2015, **30**:131-138.
5. Schizophrenia Working Group of the Psychiatric Genomics C: **Biological insights from 108 schizophrenia-associated genetic loci**. *Nature* 2014, **511**(7510):421-427.
6. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida *et al*: **Genetic effects on gene expression across human tissues**. *Nature* 2017, **550**(7675):204-213.
7. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A *et al*: **Population Variation and Genetic Control of Modular Chromatin Architecture in Humans**. *Cell* 2015, **162**(5):1039-1050.
8. Roshyara NR, Horn K, Kirsten H, Ahnert P, Scholz M: **Comparing performance of modern genotype imputation methods in different ethnicities**. *Sci Rep* 2016, **6**:34386.
9. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K *et al*: **A reference panel of 64,976 haplotypes for genotype imputation**. *Nat Genet* 2016, **48**(10):1279-1283.
10. Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D *et al*: **Chromosome conformation elucidates regulatory relationships in developing human brain**. *Nature* 2016, **538**(7626):523-527.
11. Geschwind DH, Flint J: **Genetics and genomics of psychiatric disease**. *Science* 2015, **349**(6255):1489-1494.
12. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O: **Fast and efficient QTL mapper for thousands of molecular phenotypes**. *Bioinformatics* 2016, **32**(10):1479-1485.
13. What constitutes the prefrontal cortex? Science 2017, DOI: 10.1126/science.aan8868