#### Supplement link

Please edit <https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing>

# **\*\*\*\*\*\* Notes on organization \*\*\*\*\*\***

goo.gl/f228aK

I) Introduction

II) Overall structure of Resource [Fig 1 - PE]

- Raw data => derived =>

- ref tissue project

- use it for neu-/neu+ fraction

Phenotypes, genotypes

III) Bulk & Single cell Transcriptome ^ Data [Fig 4 - XS] (including aging)

IV) Population data on epigenetics & enhancers

- (how do we make consistent enhancers over sample)

V) Consistent comparison of Transcriptome & Epigenetic across tissues w./ RCA [Fig 3 - FN, rca]

VI) QTL analsyis [Many different QTLs (fQTL, sQTLs, cQTLs, etc) [Fig 5 - SL, fql]

VII) Reference Networks & Connections -- for Brain , incl. HiC

- (Hic - networks)

- Co-expression modules

- [Fig 2 - NEW ]

- Number of hic connections & the gene expression of the links

- (elastic network )

VIII) Integrative Model [Fig 6]

IX) Discussion

# 

# 

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

## **Abstract** (246 words)

Understanding how genomic variation influences brain phenotypes remains a key challenge in neuroscience, one where the potential of functional genomic approaches has not yet been fully realized. To this end, the psychENCODE consortium developed a comprehensive, population-level resource that includes thousands of samples processed 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 deconvolved 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.

cap4: 5829, brainspan: 4486 words

<http://www.sciencemag.org/authors/science-information-authors>

Article: ~4500 words, ~5 pages incl. Refs, etc

(up to 6)

**Research Articles** (up to ~4500 words, including references, notes and captions–corresponds to ~5 printed pages in the journal) are expected to present a major advance. Research Articles include an abstract, an introduction, up to six figures or tables, sections with brief subheadings, and about 40 references. Materials and Methods should be included in [**supplementary materials**](http://www.sciencemag.org/authors/instructions-preparing-initial-manuscript), which should also include information needed to support the paper's conclusions.

*Science* also accepts a few Research Articles for online presentation. These are expected to present significant research results that cannot be fully presented in the print format and merit the extra length and attention provided by this presentation. The cover letter should indicate why the additional length is merited. These can be up to 8000 words and include methods, additional figures and potentially videos, as part of the main article. Additional [**supplementary materials**](http://www.sciencemag.org/authors/instructions-preparing-initial-manuscript) which include information needed to support the paper's conclusions, are allowed. The full text will be included in all digital versions of *Science*, and a structured abstract will be included in the print version. A pdf of the full article can be downloaded.

*Science* also accepts a few Research Articles for online presentation: up to ~8k words

**Introduction (curr words: 844 , perhaps should be 550 words)**

**[[MG: Dw should assimilate Jim's comments]]**

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 functions (ref), until recently, little progress has been made in our fundamental understanding of the molecular cause of the brain disorders. This 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, or list disorders in text below, depending on space). 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 by many GWAS studies [refs] to associate with psychiatric behaviors such as ones in mental diseases. [[JK: Add in details of other GWASs we have in the paper, once we know which ones they are.]] For example, the Psychiatric Genomics Consortium (PGC) identified a set of genomic variants including SNPs and CNVs associated with psychiatric disorders; e.g., 108 GWAS loci associated with [[confirm with Jim]] schizophrenia (SCZ) , which explained ~20%[[ask jim]] liability across major disorders \cite{23933821}. In addition to genotype, a number of genes have been reported to have specific transcriptional activities in mental diseases; e.g., the specific gene expression in mental diseases \cite{xx}. In another context, recent large 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 (N=xxx) in order to systematically identify brain specific genes, transcripts and regulatory elements [[JK: Maybe more details here, such as samples size]]. However, these studies were limited to healthy brains, so their data is unable to be used to find genomic elements for mental health. For neuropsychiatric-specific analysis, the CommonMind Consortium[[ask Jim for help on how to refer]] and others have generated gene expression and genotyping data for both healthy and schizophrenia samples (N=279 vs. 258), identifying ~693 differentially expressed genes in schizophrenia. However, their results still suggested that thousands of samples would be required to achieve statistical power of 0.8 for detecting differential expression of eQTL-associated genes [refs]. Moreover, recent studies show that specific chromatin activity of the regulatory elements such as enhancers has been found to potentially control gene expression in brain [ref], and that single cell techniques can detect gene expression and epigenetic patterns for neuronal and non-neuronal cell types from brain tissues [ref]. Given the complexity of adult brain, we need a variety of additional samples to gain the statistical power necessary for discovering a complete set of genomic elements for neuropsychiatric disorders and other phenotypes. In addition, individual molecules do not independently affect brain, and instead interact with each other in a network. Thus, effort is needed to model and analyze the molecular interactions that drive the phenotypes of adult brain including neuropsychiatric disorders.

In fact, 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 (PEC ref) (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 [[not 1931]]transcriptomic data on the brain tissue samples of 1931 adult individuals with different phenotypes and these data are housed in a central, publically available depository (xxxx). In addition, for these analyses, we have supplemented the PEC data with the primary data at both tissue and single cell levels from other related genomic resources, such as: ENCODE, CommonMind, GTEx, Epigenomics Roadmap, recent neuronal and non-neuronal single cells [refs], and uniformly processed all the data together and performed integrated analyses with up to X,XXX samples. Using single cell data, we also calculate the fractions of neuronal and non-neuronal cell types in normal and disease states for individual tissue samples. We provide all the PEC data and integrative analyses in an online resource which contains all [[too strong]]possible functional genomic elements for adult brain including the brain-active enhancers, transcripts, expression models, imputed regulatory networks, eQTLs and cQTLs for various phenotypes, and [we combine] an integrated deep-learning model, Deep Structured Phenotype Networks (DSPN) for predicting and imputing brain phenotypes. We then use this resource to discover the properties of brain gene expression, non-coding transcription and enhancers, and to build this model, to describe how interactions between genomic variants, gene expression, enhancers might work together molecularly to alter disease risk.

## **\*\*II\*\*(Fig-1) Comprehensive resource for adult brain functional genomics**

Intro + Next generation sequencing data for brain functional genomics (291) [[2par target, base on MG's rough text. Use edits]]

\* Overall, this resource is structured in a pyramid shape, with the largest scale and raw data at the bottom level and the lightest and most interpretive data at the top level.

First step in our analysis is to construct a coherent data resource. Broadly, this data resource is shaped like a pyramid with a large amount of data, often controlled access at the bottom, followed by progressive summarization of the data, first into uniform peak calls and quantifications and then into specific genomic elements, such as answers and regulatory networks. Finally, at the top, our resource, of course, contains an intuitive model, which we discussed at the end. In terms of the building the data corpus, we took all the psych encode data related to adult brains and merged it with relevant data from outside projects including Road Map, G Text, Encode. We also added a number of relevant single cell data sets, as described in the methods. All the raw data was processed using encode standard pipeline, both for the bulk and for the single cell data and for both chip and RNA seek.

\* the raw data is show(Figure 1)

PEC has carefully desinged the study to diseases, cohorts, assay

Combined adult data w/ relevant orther [data cube]]

\* used the reference brain project for annotations

- using the hic for hard data ETG linkages

- special reference

& use the cohorts for variablility

(206 words)The PsychENCODE consortium has generated and assembled a large-scale dataset of genotypes, RNA-seq, ChIP-seq, ATAC-seq, Hi-C and single-cell transcriptomic data from adult brains of 1931 individuals, with and without several mental illnesses (Figure 1, Assay summary in Methods).

\*\*\*IV\*\*\*ENHANCERS

\*\*\* Go later in deerived sections

- enhancers use are reference

- (how do we make consistent enhancers over sample)

- we use the reference to call enhancers consistently to encode ... then we intersect these w k27 peaks on the samples to find enahncers [enhancer variability... we use the consistent for the cqTL later]]

The processing of the data gave rise to uniform quantifications, uniform peak call lists and single tracks. Then, we took these and derived further simplified data sets.

First, we developed a consistent set of brain active enhancers. These were based essentially on K-27 peak calls and across all the psych encode data sets that we then merged in a consistent fashion with the encode and roadmap enhancer calls, which are based on K-27 and DNAs, as described in the methods. This gave us an overall set of enhancers across all the psych encode data, Road Map and You Text. We found, in total, 88,000 active enhancers in dorsal lateral prefrontal cortex. We also derived enhancer sets in other brain regions.

*Epigenomics* - we used the ENCODE standard ChIP-seq pipeline and uniformly processed the ChIP-seq data of available tissue samples in PsychENCODE and Roadmap Epigenomics, and neu+ and neu- cell samples for the signal track qualifications and peak identifications.

[[tissues]]

We have 88K enhancers in DLPC

We have a reference set in another another 2 region

We developed reference sets for tissues ACC

Then, we derived a number of other key components to our resource include the networks derived [inaudible 00:03:15] processing, the single cell data types and the [inaudible 00:03:21] described in subsequent sections.

To harmonize and integrate the datasets across multiple consortia,[[cut]] we processed these datasets using standard bioinformatic pipelines in common use (Methods). For instance, we adopted the ENCODE processing pipelines for the bulk and single cell RNA-seq and ChIP-seq data. Likewise, we used the GTEx eQTL pipeline and associated parameters, to allow comparison to previously published eQTL maps. All these uniformly processed datasets are available in our XXXX resource (URL here). Finally, we also compared the resource data against various phenotypes, and identified the brain specific data (derived data type). 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 (http://adult.psychencode.org/pec/).[[cut]]

At the bottom, we have the large scale raw data and the phenotype information for 1931 individuals, much of which is private and under controlled access. Based on this, we have then uniformly processed raw datasets from PyschENCODE and other consortia (ENCODE, CommonMind, GTEx, Epigenomics Roadmap, etc), including RNA-seq expression quantifications, ChIP-seq signal track qualifications and peak identifications using ENCODE standard pipelines, and private imputed genotypes. The processed functional genomic data is much easier to interpret but still rather large scale. In details, they include the following major data types:

[ [genotypes?]]

*[[too detailed]]*

*Phenotypes* - the PsychENCODE data covers a number of phenotypes on mental health. They are normal control (n=1104), SCZ (n=558), BP (n=217), ASD (n=44), AFF (n=8), Male (n=1246), Female (n=685), Age (distribution), etc. (Supplement).

*Transcriptomics* - we also used the ENCODE standard RNA-seq pipeline to uniformly process the RNA-seq data of available samples from a number of PsychENCODE-relate studies, ENCODE and GTEx to quantify the expression levels for the protein coding genes, transcripts, noncoding RNA and novel transcribed regions.

*Chromatin interactomics* – we generated and processed the Hi-C data for adult brain including three reference brains, and identified the xxx regions on which the enhancers and promoters interact. Using this full Hi-C data for adult brain, we identified xxx Topologically Associating Domains (TADs) of adult brain. 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]

System identification of the specific transcriptomic and epigenomic elements in adult brain (438)

Given the large-scale transcriptomic and epigenomic data in resource, we further integrated them and identified the genomic elements that have specific activities in adult brain. We used the uniformly processed data and compared against various phenotypes[[delete]] to have even more interpreted functional elements such as sets of differentially expressed genes characterizing various brain regions and phenotypes, sets of aggregated brain enhancers from merging the the K27 peaks on the ENCODE regulatory elements. And then above these individual elements, we even identified more interpreted association relationship data such as the QTLs affecting gene expression and enhancers, and imputed the regulatory networks consisting of QTLs, transcriptional factors (TFs), enhancers and genes. This includes:

*Brain active enhancers* - we identified the brain enhancers from the uniformly processed ChIP-seq data and related them with the regulatory elements in ENCODE and Epigenomics Roadmap , and summarized a list of PsychENCODE brain enhancers which are activated on major brain regions such as ~88,800 enhancers in pre-frontal cortex including xxx ones in adult brain TADs (Supplement).[[other tissue]]

*Differentially expressed genes, transcripts and brain splicing patterns [cut to 1 line]* - we compared expression changes in uniformly processed RNA-seq data from brain samples across PsychENCODE-related studies, ENCODE, and GTEx, and found xxx expressed genes and ~79,000 transcripts in pre-frontal cortex, ~11k eGenes associated with eQTLs (Methods), and xxx non-coding RNAs and novel transcribed regions.[[cut]] We also derived phenotype-specific genes and transcripts. In addition, we calculated the alternative splicing patterns at the transcript level; i.e., the percentage of the transcript abundance over its gene abundance, and found the brain-specific spliced transcripts. Our resource contains differentially expressed and spliced genes and transcripts across a number of biological variables, including neuropsychiatric disorders and developmental stages.

*Gene co-expression modules* - 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 Sxxx, Supplement), which suggests that there exist brain specific regulatory mechanisms drive these brain co-expression modules.[[move to sup]]

We should emphasize that our comparative analysis is consistent for finding various brain elements including brain enhancers, genes and transcripts. More specifically, we compared them against a same set of brain and non-brain tissues; e.g., the RNA-seq gene expression data from GTEx and the ChIP-seq binding signal data from Epigenomics Roadmap for brain pre-frontal cortex vs. other non-brain tissues including liver, lung, blood, etc.

## \*\*\*VI\*\*\* QTL

## System identification of the QTLs and gene regulatory networks associated with adult brain transcriptomics and epigenomics (955)

For the EQTLs, we adopted a fairly standard approach and emphasized the scale of the database. We adhered closely to the established GTEX pipelines. An overall calculated XXX number of EQTLs and YYY number of e-genes in DLPSE. This is a conservative larger number of EQTLs than have been previously been calculated [inaudible 00:00:47] 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.

\*\*\* more on cQTLs

cQTLs ? b/c enhancer move ... how do

- we std and quantified over the indiv.

Fig.

This doesn't capture all the variation

For the chromatin QTLs or CQTLs, the situation is more problematic. There aren't as established methods of calculating these on a large scale. In particular, to calculate these properly, we have to both define the region associated with the activity of the chromatin and then look at how this activity varies. To do this, we did joint K27 peak falling over the hundreds of brains in psyche end code with HVK27AC mar. From the consensus [inaudible 00:02:05] region, we calculated an average value of the chromatin mark. And then we correlated this with a variety of nearby sniffs. See methods.

Overall, from doing this, we were able to calculate YYY a thousand chromatin QTLs.

\*\*\* decoration

Then we took all our IQTLs, EQTLs and CQTLs and decorated them, overlapped them and decorated them with a variety of different genomic annotations and look at the degree that they overlapped. As expected, there's a very large amount of overlap between the CQTLs and EQTLs, and with the CQTLs essentially being a subset of the EQTLs.

[[nuke]]To understand how the genotype affects the transcriptomic and epigenetic activities in adult brain, we first used the resource data as above to identify more interpreted association relationship data such as the quantitative trait loci (QTLs) affecting gene expression and chromatin activity. In particular, we merged genotype and gene expression and chromatin 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 the quantitative trait loci associating with gene expression and epigenomic activities in adult brain, including three major categories: expression QTLs (eQTLs), chromatin QTLs (cQTLs), splicing QTLs (sQTLs) and even cell fractions (fQTLs, more details from the single-cell analysis as below). We used the GTEx standard pipeline for discovering eQTLs to find the associations, which is based on an additive linear model from QTLtools. Given the complex relationships between genotype and phenotype, potentially driven by batch effects and biases (e.g., merging different chromatin datasets), this linear model was also adjusted by covariates like PEER factors of gene expression, genotype PCs and disease diagnosis.[[more on cQTL]] Among these SNPs, we identified a great number of the regulatory variants significantly associated with brain transcriptional and epigenomic activity: >1 million expression QTLs (eQTLs) with ~11k eGenes, >5 thousand chromatin QTLs (cQTLs) for histone modification signals, and xxx splicing QTLs for alternative splicing patterns. The distributions of detailed QTL annotations on genomic regions are shown in Figure xxx.

Given a great number of QTLs we identified, we are further interested to see how they relate to the known variants for brain. In particular, we compared them with existing QTLs databases and subdivided our QTLs into different functional categories, mainly including the disease GWAS SNPs, the SNPs breaking the TF binding sites, etc (Table/Figure xxx). Collectively, these QTLs annotate a larger fraction of GWAS SNPs involving the brain (e.g., 6% in schizophrenia, 10% in bipolar) than previously observed, providing leads on which genes are affected in disease. [[too rep]][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. In addition, we link the QTLs that overlap the enhancers and promoters in the resource to reveal the potential regulatory activities. We thus classified the QTLs into subgroups in terms of their gene regulatory characteristics including the regulatory QTLs (rQTLs) that break TF binding sites on promoters and/or enhancers, and the modular QTLs (mQTLs) that highly associate with a set of co-expressed genes. Finally, we found that the eQTLs/eGenes number can be predicted from the sample size using a fitted curve (Figure xxx).

*Gene regulatory networks* - 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). For example, we 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 wires 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 [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 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). 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.

In summary, the establishment of this comprehensive resource enables the modeling and analysis for the biological processes in adult brain and helps understand the molecular mechanisms between genotypes and phenotypes. Therefore, we later analyzed and modeled the data from this resource to further reveal the brain specific genomic and transcriptomic activities, and the biological mechanisms explaining how the brain specific elements affect the phenotypes and diseases in the adult brain.

\*\*\*\*V\*\*\*\*) Consistent comparison of Transcriptome & Epigenetic across tissues w./ RCA [Fig 3 - FN, rca]

## **Comparative analysis reveals the brain related transcriptomic and epigenomic activity (813)**

\*\*\* use the blue text

One of the key aspects of our analysis is that we, as consistently as possible, processed the RNA and chromatin data across [sican 00:00:19] code, encode, road map and Gtex together. This allows us to compare the brain to other organs in a consistent fashion, consistently both for RNA and chromatin. This not being able to be done without such a large scale uniform processing. We show this in figure xxx. To do the comparison we had to find an appropriate way of doing the comparison. Principal component analysis and t-SNE are two popular techniques, but we found the former tended to be overly influenced by outliers in the data which are often, and t-SNE tends to just uniformly separate all the clusters and not give one a sense of the overall effect. We found a very useful technique to be RCA, which compares the gene expression in each sample against a reference panel, and then essentially does dimensionality reduction against that. We did RCA consistently for both the chromatin using the genes and ... Sorry for RNA using the genes and chromatin using our consistent set of enhancers.

The genes analysis 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 tissues from the different projects grouped together. This difference is accentuated when one looks not at all the individual [inaudible 00:02:14] but just simply looks at the cluster centers and the distribution about them. The difference between brain and other tissues is much different than the variability within any of the given tissues.

A different picture emerges when one looks at chromatin. Here we see that the chromatin levels are indistinguishable between brain and other tissues. Thus the expression difference, in a sense, in brain cannot be simply attributed to that chromatin. [[trans]]The expression differences that we're looking at are those confined to well known canonical regions such as genes, however people have previously remarked about the tremendous amount of transcriptional diversity in intergenic regions.

\*\* sat'd figure - unannotated transcription

As a supp figure we going to have FN's unnannotated done for genes

We tried to get a sense of this looking at the overall level of transcript diversity across the entire genome as shown in figure YYY. In terms of genes, it has been previously known that testes tends to have the most transcriptional diversity and not genes, however when we shift to non coding 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.

Finally, one can to some degree relate the transcriptional diversity in the brain to the specific difference in expression of different genes in the brain relative to other tissues and amongst different brain regions, and this is evident in our differential expression resource on the website.

We leveraged this resource to compare the human brain with other tissues. To reveal potential brain specific genomic activities, particularly relating to transcriptomic and epigenomic activities, we performed a consistent spectral analysis and compared the similarities of RNA-seq gene expression and ChIP-seq binding signals on enhancers and found that the brain has more distinct expression patterns compared to most other tissues, including a greater amount of non-coding transcription. However, the differences in epigenetics are relatively smaller.

For gene expression, we compared the adult brain samples from our resource with the other tissue samples from GTEx, using uniformly reprocessed RNA-seq data. We tested three well established dimensionality reduction methods to identify structures of gene expression. Principal Component Analysis (PCA) was able to capture some, but not all structure of human tissues. On the other hand, tSNE is too sensitive to batch effects and exposed structures that have not originated from biological differences. We finally tested Reference Component Analysis (RCA), that projects the gene expression into a reference panel of tissues and genes and shows highlights intermediate structures in the data. Using the reference component RCA, we show 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 from their leading reduced dimension (left vs. right clusters in Figure xxx). This suggests that the brain has unique and distinctive gene expression programs, which are involved by the brain elements including brain expressed genes, transcripts and non-coding RNAs in our resource. In addition, the samples of PsychENCODE that include psychiatric disorders have larger variations than other tissue clusters (Figure xxx). The cluster radiuses were estimated by fitting the two main principal components into a multivariate normal model and finding a 0.95 confidence interval (Methods). This suggests that the psychiatric diseases still have larger variations of gene expression, and different gene regulatory programs from the normal, though even more distant from other organs. Thus, we then want to check all unified transcriptional activities on the genome scale in brain including potentially novel transcribed non-annotated regions. Specifically, to understand where the human brain sits in regards of its the transcription diversity compared to other tissues, we estimated the proportion of genome that is transcriptionally active across hundreds of samples. We first found that transcript diversity[[how measure]] is mostly saturated at the scale of hundreds of individuals (Figure xxx). The saturation is observed for both the annotated and non-annotated portions of the genome. The human brain does not stand as a highly diverse in protein coding regions. For example, the tissues such as the testis is highly diverse [Ref]; however, we found that the brain has more transcriptional activity at the non-annotated 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.

As shown above, the brain samples have different chromatin and gene expression activities from other organs, implying that the brain also has specific gene regulatory activities. Therefore, we are further interested to compare the enhancers between brain and other tissues to see any brain epigenomic activities. In particular, we integrated the H3K27Ac ChIP-seq signal data of enhancers in the resource and performed dimensional reduction analysis consistent to the for gene expression RCA to compare the similarities of epigenetic profiles of PsychENCODE samples with Epigenomic Roadmap data. It is also interesting to find dissimilar patterns with the gene expression comparison; e.g., while the brain samples separates from other tissues when using genes expression data, the active enhancers[[active genes]] are not able to separate brain from other tissues (Figure xxx). This result suggests that the brain has less specific and distinct epigenomic activities, involving the brain active enhancers from our resource. Thus, there may exist more complex regulatory mechanisms among the brain enhancers with low signal variability than other tissues to drive the brain distinct gene expression. One important mechanism is that the brain active enhancers or gene expression patterns are intermediate phenotypes, potentially driven by particular large set of brain regulatory variants such as our QTLs as previously described.

Our comparative analysis reveals that the brain is different from other organs in gene expression. Thus, we are then interested to identify the functional genomic elements in brain that give rise to the uniqueness of brain. To systematically find the specific expressed functional elements in brain, we identified the differentially expressed genes for phenotypes such as gender (Methods and Figure XX) for the resource. For example, we identified a group of genes that differentially express across different 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 DEX genes for all phenotypes in our resource along with their enriched functions and pathways in supplement[[how fit? Put all age together]].

## **\*\*\*III\*\*\* Single cell analysis and deconvolution explain gene expression changes across adult phenotypes (822)**

## **[[MG: this comes after the resource, use my blue text as base and merge Jim's. Add up top the bulk transcriptome -- we processed all of gtex & encode w/ pipeline - differential expereinctial (captstone1)]]**

\*\* bulk stuff --- rnaseq (goes up in the new outline)

We also derived uniform quantifications for RNA seek, differential gene expression between the different brain regions and between brain and other tissues and we clustered the gene expression patterns across the tissues to develop modules of brain active genes.

\*\* use the

For the single cell processing, we adopted a two prong strategy. On one hand, we used the [inaudible 00:00:18] code single-cell data in conjunction with the number of other single-cell studies on the brain to create a list of basic cell types in the brain. These are mostly concordant with what's been previously published with some minor modifications and are shown in figure "XXX."

As is obvious across these cell types a number of genes varies much more substantially than they do amongst individuals and so forth. This implies that easily the variation of cell types can give rise to substantial changes in expression.

The [[2nd]] second thing we did was an unsupervised analysis of our bulk tissue data trying to find main components in the data that perhaps relate to fugitive cell types. Then we interrelated these two things to see if they were consistent. This is shown in figure "YYY" where we can see that the brain cell types [inaudible 00:01:38] is derived by others. And from our single-cell data, seem to match in a rough fashion against an unsupervised analysis just of the main components in the prefrontal cortex sequencing. This suggests that these cell types do make sense and are a good way to parse the data.

After this, we then deconvolved the tissue expression in terms of the main brain cell types. Here, we essentially solved the equation "Y-WX" finding the fractions of cell types, multiplying the cell types expression that best explained the gene expression.

After doing this, we found that a lot of the overall variation in gene expression that we found over the population could be accounted for by a variation in cell types. In fact, up to 80% of the variation in gene expression could be accounted for in terms of cell types.

Next we showed that a lot of the main phenotypic differences that we have in our study, for instance, between men and women, between people with neuro-psychiatric diseases and not can be seen in terms of differences in cell types. For instance, there being a different number of [inaudible 00:03:19] neurons in males and females, and also different cell fractions [inaudible 00:03:24] neurons in people with schizophrenia versus not.

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 tissue, we integrated the single cell gene expression data to discover how the gene expression from various cell types including both neuronal and non-neuronal contribute to the gene expression at the tissue level. In particular, we used the biomarker genes with strong expression signals in single cell to deconvolve the gene expression data of individual tissues over both novel and known cell types to find the cell fractions for individuals, and relate to the individual phenotypes.[[goes to method]] We found that the gene expression changes across individual tissue samples can be largely explained by the single cell gene expression, and the changes of single cell fractions are also associated with the individual phenotypes.

Specifically, we integrated and used the same pipeline to uniformly process the single cell RNA-seq data for the neuronal and non-neuronal cell types from PsychENCODE and recent publications [lake&quaker]. In total, we included 23 single cell types (Supplement) and found that the same-type cells generally can be clustered together (Figure Sxxx) using our uniformly processed data. We also include these single cell data as well as their cell-type biomarker genes in the resource. Moreover, we found that a group of psychiatric disorder related genes indeed show the expression dynamic changes among cells. For example, the dopamine receptor genes (DRD) that associate with SCZ, are significantly more highly expressed in neuronal cells than others (Figure Sxxx), and their expression levels across cells vary significantly larger than tissue samples, suggesting that the cell fraction changes potentially equalize the tissue expression variability. Therefore, we are further interested to see if the brain gene expression at the tissue level in our resource is contributed by the above cell types and affected by the cell fractions[[delete]].

To this end, we decomposed the gene expression data across individuals at the tissue level from our resource using non-negative matrix factorization (NMF, see Methods). Indeed, we found that three groups of top principal components of NMF (NMF-PCs) capturing the most covariance of brain gene expression across individual tissues, highly correlate with the biomarker gene expression signatures of neuronal, non-neuronal and fetal cell types as above, respectively. For example, the NMF-PCs shown in Figure xxx. This suggests that the large portion of tissue’s gene expression changes is a linear combination of these cell types’ gene expression. Thus, we want to further identify the cell fractions showing how individual single cells contribute the tissue’s gene expression, using the deconvolution.

Therefore, we deconvolved the tissue-level gene expression data of all 1931 individuals’ tissue samples using single-cell gene expression data of 450 biomarker genes to find the fraction of different cell types corresponding, and compare cell fractions across different phenotypes (Supplement). The single cells used in deconvolution cover all 16 neuronal types, five non-neuronal types and xxx additional fetal types from PsychENCODE single cell data [ref: brainspan]. It is very interesting that the linear combinations of single cell expression of 23 cell types, where combinational coefficients, can explain >80%[[show heatmap?]] of the gene expression variations across 1931 individual tissues (Figure xx). The coefficients of cell types for linear combination are estimated from our deconvolution analysis (Methods in supplement), and proportional to the cell fractions of individuals. In addition, we found that the cell fractions of individuals (i.e., deconvolution coefficients) vary, and a number of cell population changes highly associate with different phenotypes and disorders (Figure xxx). For example, 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). The excitatory neuronal cell populations (e.g., EX1) increase significantly in ASD samples (p<xxx) while the non-neuronal cells decreasing (e.g., oligodendrocytes). Finally, we report the individual cell populations along with significantly associated relationships between particular cell type fractions and phenotypes (Supplement).

Furthermore, we are interested to see if any genotype is also associated with two single cell features: (1) the cell fractions and (2) the gene expression changes that can’t be explained by the cell fractions. In particular, we used our QTL pipeline and identified xxx SNPs whose genotypes are significantly associated with yyy neuronal cell fractions across individuals, (or zzz non-neuronal cell types); i.e., cell fraction QTLs (fQTLs). This suggests that these fQTLs potentially can be used to predict the yyy cell fractions in adult brain. Moreover, we identified xxx SNPs significantly associated with the gene expression changes across individual tissues unexplained by our single cell deconvolution[[move fqtl to QTL sec]]; i.e., Y-WX (Methods). These SNPs are likely causing certain gene expression changes driven by unknown cell types in adult brain.

## 

\*\*\*\*\*VII)\*\*\*\*\* Reference Networks & Connections -- for Brain , incl. HiC

- (Hic - networks)

- Co-expression modules

- [Fig 2 - NEW ]

- Number of hic connections & the gene expression of the links

We full data set ,... we processed ... we have connections ... we also used jeme

- (elastic network )

## 

## **Integrative modeling to explain the molecular mechanisms for genotype-phenotype relationships in adult brain (737)**

The interaction between genotype and phenotype is a very complex process, involving multiple intermediate stages including gene expression, signaling, modulation and so on. Thus, to understand this and merge all these stages in one model, we 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; i.e., the DSPN pathways from genotype to phenotype (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. 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 regulatomics, 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 architecture, the undirected form of the Boltzmann machine allows information to flow in top-down, bottom-up and lateral directions during inference, so that intermediate and high-level phenotypes may be jointly inferred while respecting their mutual dependencies. This allows us for instance to impute transcriptome and epigenome data when it is missing. In particular, our inference is performed using a mean-field approximation, and training is performed using a Persistent Markov Chain Monte Carlo algorithm which is able to ensemble multi-dimensional datasets (Supplement).

\*\*\* we need to put refer the new schematic [[put fig]]

Simple logistic to the

As shown in Figure xxx, the DSPN consists of four layers: 1) genotypes such as QTLs; 2) molecules and genomic elements, including genes and enhancers; 3) functional modules and other mid-level phenotypes at a series of intermediate layers; i.e., the hidden nodes of deep learning modeling; 4) high-level phenotypes such as brain traits. In addition, we enforce the DSPN to have sparse connectivity (Supplement). Specifically, we built each layer of our model as follows. We first used the imputed gene regulatory networks that identify the regulatory connectivities on how QTLs, enhancers, and transcription factors relate to target gene expression (Supplement). We then connected the nodes on the molecular layer[elastic net?] of our model to follow the inferred gene regulatory network structures; i.e., embedding 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 connected to genes enriched in the dopaminergic and glutamatergic synapse (GSEA enrichment score > xxx, Figure xx). 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 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 discovered[do not use discover] 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).

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.

## **Discussion (Currently at 723 words => 500 words)**

**[[MG: Dw should assimilate Jim's comments]]**

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. Developing this 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. 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 add the individual’s fMRI image features measuring functional neuro-connectivity, and use our model to identify the genotypes that associated with image features such as image-QTLs (iQTLs) [xx]. Also, our resource can incorporate with the neurodegenerative diseases like Alzheimer or developmental stages.

Moreover, we built an integrative epigenome- and transcriptome-wide association model (eTWAS), built on the Deep Boltzmann Machine (RBM) and integrates the high dimensional functional genomic and phenotypic data at multiple layers, using the hierarchical structures in deep learning. The model reveals the relationships among various data types from a number of directions for genotype to 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 the genotype to phenotype. 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. Also, the model can be used to make in-silico predictions for the perturbation outcomes. For example, we can identify the module X that have the extremely highest connection weights to Austin, and thus knocking down the genes connecting to the module highly likely will deactivate Autism. Furthermore, while the model does provide better predictive performance, some of these correlations are deliberately set to be interpreted simplifications, such as the known enhancers, or gene regulatory network structure, to make the model more interpretable and easier to use. Thus, another major goal of the model is to provide a compression of larger amount of 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.

Though single cell remains challenging to reliably quantify the low-abundant transcripts/genes and interrogate the biological variations using single-cell sequencing technology, it is still worthwhile using the biomarker genes with strong expression signals in single cell to deconvolve 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. 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. The current single-cell sequencing technology suffers from the low capture efficiency [PMCID: PMC4758375, PMCID: PMC4132710]. Due to this reason, the single-cell sequencing will only measure a small fraction of cellular transcriptome as the final sequencing library only contains a subset of input materials. 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 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 **Supplement** Please edit<https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing> |

|  |
| --- |
| [References](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) [1. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, Shah HR](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[et al](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)*[:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Gene expression elucidates functional impact of polygenic risk for schizophrenia](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Nat Neurosci](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2016,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[19](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(11):1442-1453.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [2. Consortium GT:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Science](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2015,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[348](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(6235):648-660.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [3. Psych EC, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[et al](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)*[:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[The PsychENCODE project](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Nat Neurosci](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2015,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[18](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(12):1707-1712.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [4. Neale BM, Sklar P:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Genetic analysis of schizophrenia and bipolar disorder reveals polygenicity but also suggests new directions for molecular interrogation](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Curr Opin Neurobiol](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2015,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[30](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[:131-138.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [5. Schizophrenia Working Group of the Psychiatric Genomics C:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Biological insights from 108 schizophrenia-associated genetic loci](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Nature](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2014,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[511](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(7510):421-427.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [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](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[et al](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)*[:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Genetic effects on gene expression across human tissues](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Nature](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2017,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[550](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(7675):204-213.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [7. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[et al](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)*[:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Population Variation and Genetic Control of Modular Chromatin Architecture in Humans](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Cell](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2015,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[162](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(5):1039-1050.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [8. Roshyara NR, Horn K, Kirsten H, Ahnert P, Scholz M:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Comparing performance of modern genotype imputation methods in different ethnicities](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Sci Rep](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2016,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[6](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[:34386.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [9. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[et al](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)*[:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[A reference panel of 64,976 haplotypes for genotype imputation](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Nat Genet](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2016,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[48](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(10):1279-1283.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [10. Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[et al](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)*[:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Chromosome conformation elucidates regulatory relationships in developing human brain](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Nature](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2016,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[538](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(7626):523-527.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [11. Geschwind DH, Flint J:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Genetics and genomics of psychiatric disease](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Science](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2015,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[349](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(6255):1489-1494.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [12. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[Fast and efficient QTL mapper for thousands of molecular phenotypes](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) *[Bioinformatics](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)* [2016,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) **[32](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**[(10):1479-1485.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)  [13. What constitutes the prefrontal cortex? Science 2017, DOI: 10.1126/science.aan8868](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) [Supplement](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) [Please edit <https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing>](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) |

### 

**[############](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)**

[old\_abstract](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[The PsychENCODE consortium has developed a comprehensive dataset on the adult human brain, including genotyping, RNA-seq, ChIP-seq and single-cell analysis on many individuals. We integrated this dataset, compared it against various brain phenotypes and merged it with complementary genomic information from the ENCODE, GTEx and the Epigenomics Roadmap projects to develop a comprehensive resource for the brain comprising brain-active enhancers, transcripts, expression models, imputed regulatory networks, eQTLs and cQTLs. Overall, this involves ~2000 adult brains samples. We make the derived resources downloadable and available on the PyschENCODE website (xxxx). We then used this resource to identify both cross-tissue conserved and brain specific genomic elements using comparative analysis with other tissue data from GTEx and Epigenomics Roadmap and associate the brain-specific ones with adult brain phenotypes. This shows the brain has distinct expression and epigenetic profiles as evident from spectral analysis and more non-coding transcription from most other tissues. Moreover, we developed and built an integrated model to predict the brain phenotypes using all the functional genomics data in this resource from QTLs to variants breaking TFBSes on enhancers to differentially expressed genes and non-coding RNAs. This model 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. In addition, the model allows us to impute the functional genomics data not present in our dataset.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Mark’s transcripts on Nov 25 2017](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[The core of the PsychENCODE dataset, obviously, is a large amount of functional genomics and genotype information related to the human brain. However, to make the dataset maximally useful, we interconnected it with a number of other related genomics resources to both make it larger and also [inaudible 00:00:29]. These other resources include, of course, ENCODE, CommonMind, GTEx, Roadmap, and so forth.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[To interrelate all these datasets, we had to process them all to a common standard. We adopted the ENCODE standards for PsychENCODE and then had to reprocess them over the other main datasets such as Roadmap and GTEx to this standard. After we were done, we could uniformly relate the PsychENCODE brain data to related data in other organs from GTEx and Roadmap.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Some thoughts on the overall structure of the model and the data for the resource section of the introduction.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Overall, this resource is structured in kind of a pyramid shape, with the largest scale and most unwieldy data at the bottom and the lightest and most interpretive bits at the top.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[At the bottom, of course, we have the large scale bands from this project and the phenotype data. Much of this data, of course, is private and under controlled access. Then, above this we have the uniformly processed data from this project, singles tract, rnaC quantifications, ChIP-seq single tracts, quantifications, and peaks both from this project and also uniformly processed from other projects. Much of this data is much easier but it's still rather large in scale. The large scale imputed genotypes obviously are still private.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Above this we have even more interpreted things, sets of dex genes characterizing various brain regions and phenotypes, sets of aggregated enhanced recalls from merging the encode regulatory elements of the K27 and K27 peaks. And then above this even more interpreted stuff, imputed regulatory networks based on the enhanced [promotocalls 00:01:40] and the motif catalogs and then of course eQTLs and cQTLs and some notion of which of them are perhaps the strongest of these variants.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Finally, at the top we have our model. The model is meant to play very well with this data and on one hand be a multi-tiered deep-learning model that can be used in different directions but to also incorporate explicitly a lot of interpretive data. So in particular the model incorporates the structure of the imputed regulatory network and the cQTLs and the enhanced recalls directly with of course the quantifications. The idea of course is that someone can download them all and be able to quickly impute transcriptomes or get a sense of the variant positions that have the largest overall effect on relevant gene expression quantifications.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[The structure of the model is deliberately set up as a hybrid model, that on one hand incorporates contemporary deep-learning ideas to model this large amount of data with a multi-tier architecture. But it also incorporates a hybrid structure that explicitly incorporates much of the imputed EQTLs and CQTLs. The idea behind the RBM architecture is that the model can be used in a number of directions. On one hand, it can be used to better predict phenotype and genotype, adding in some additional predictability from all the expression and chip data. On the other hand, it can be run the other way, using known or elaborated genotype/phenotype associations and better pinpointing them to specific gene expression changes, or modules of dysregulated genes. This latter use, of course, enables one to better localize a known genotype to phenotype relation to specific molecular events that may be associated [inaudible 00:01:31] with a particular use in relation to mental illness.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[The model is made available as a set of simplified files, where one can explicitly see the correlations being used at various stages. Furthermore, while the model does provide better predictive performance, some of these correlations are deliberately set to be interpreted simplifications, such as the known enhancers, or regulatory network structure, to make the model more interpretable and easier to use. The main goal of the model is to be a compression of larger amount of data, rather than a purely predictive construct](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Notes from brain meeting](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Title:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Integrative modeling, analysis and resource reveal the functional genomics in the adult brain](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*compresensive genomic resource & integrative model for the brain](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Integrative analysis and resource for the functional genomics in the adult brain](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\* complex or composite phenotypes](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\* aging](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\* neuronal v non-neuronal, modules](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\* across whole](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Title:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Comprehensive resource and integrative model for functional genomics of the adult brain](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Abstract:](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[The PsychENCODE consortium has developed a comprehensive data set on the adult human brain, including genotyping, RNA-seq, Chip-seq and single-cell analysis on many individuals. We integrate much of the data, compared it against various brain phenotypes and merged it with complementary genomic information from the ENCODE, GTEx and Roadmap projects to develop a resource for the brain comprising brain-active enhancers, transcripts, expression models, imputed regulatory networks, eQTLs and cQTLs. Overall, this involves ~2000 adult brains samples. We make the derived resources downloadable and available on the PyschENCODE website (xxxx). We then used this resource to identify both cross-tissue conserved and brain specific genomic elements using comparative analysis with other tissue data from GTEx and Roadmap and associate the brain-specific ones with adult brain phenotypes. This shows the brain has distinct expression profiles as evident from spectral analysis and more non-coding transcription from most other tissues. Moreover, we developed and built an integrated model to predict the brain phenotypes using all the functional genomics data in this resource from QTLs to variants breaking TFBSes on enhancers to differentially expressed genes and non-coding RNAs. This model shows that the integrated data has significantly improved the prediction accuracy over individual genomic data types and relates these predictions to well characterized pathways in the brain. In addition, the model allows us to impute the functional genomics data not present in our dataset.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Abstract from Mark](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[We develop a comprehensive resource for brain functional genomics through integrating the adult data in the psyche end code dataset comprising almost 2,000 brains. This data resource comprises brain active enhancers EQTLs, CQTLs, and transcripts. We integrate our brain resource with that some other genomic resources, such as GTEx and Roadmap the highlight brain specific transcripts and enhancers and QTLs. Finally, we develop an integrated quantitative model relating transcription binding genotype and phenotype. This allows us to impute the functional genomics data not present in our dataset and also show how the data integration can make a prediction onto disease and phenotype.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[We analyze and integrate the adult component of the PsychENCODE data resource, comprehensively determining active transcription and binding in the adult brain and relating it to genotype. This enables us to develop a resource consisting of active enhancers, transcripts, and ENT qtl's in the adult brain. We integrate this data resource from other genomic's resources such as GTEX and our roadmap, using to characterize genomic aspects of brains that are most unique and those that are similar to other tissues.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Finally, we build an integrated model of all the functional genomics data, qtl's, and phenotypes of the PsychEncodes which allows us to compute much of the functional genomics data from our model and also to make integrated predictions of phenotype that are more accurate than from an individual data type alone.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[[old 20171116] The PsychENCODE consortium has developed a comprehensive dataset on the adult human brain, including genotyping, RNA-seq, ChIP-seq and single-cell analysis on many individuals. We integrated this dataset, compared it against various brain phenotypes and merged it with complementary genomic information from the ENCODE, GTEx and the Epigenomics Roadmap projects to develop a comprehensive resource for the brain comprising brain-active enhancers, transcripts, expression models, imputed regulatory networks, eQTLs and cQTLs. Overall, this involves ~2000 adult brains samples. We make the derived resources downloadable and available on the PyschENCODE website (xxxx). We then used this resource to identify both cross-tissue conserved and brain specific genomic elements using comparative analysis with other tissue data from GTEx and Epigenomics Roadmap and associate the brain-specific ones with adult brain phenotypes. This shows the brain has distinct expression and epigenetic profiles as evident from spectral analysis and more non-coding transcription from most other tissues. Moreover, we developed and built an integrated model to predict the brain phenotypes using all the functional genomics data in this resource from QTLs to variants breaking TFBSes on enhancers to differentially expressed genes and non-coding RNAs. This model 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. In addition, the model allows us to impute the functional genomics data not present in our dataset.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[The PsychENCODE consortium has developed a comprehensive dataset on the adult human brain, including genotyping, RNA-seq, ChIP-seq and single-cell analysis on many individuals. We integrated this dataset, associated it against various brain phenotypes and compared it with complementary genomic information from the ENCODE, GTEx and the Epigenomics Roadmap projects to develop a comprehensive resource for the brain comprising brain-relevant QTLs, brain active enhancers, differentially expressed genes and transcripts, novel non-coding RNAs. In particular, it contains regulatory variants significantly associated with brain transcriptional and epigenomic activity in >2000 individuals: >5 million expression QTL for gene expression and >5 thousand chromatin QTL for histone modification signals. We make the derived resources downloadable and available on the PyschENCODE website (xxxx). Moreover, using single-cell data, we deconvoled the tissue-level gene expression of this resource to find the populations of different neuronal and non-neuronal cell types and relate them to various phenotypes. We then used this resource to identify brain specific genomic elements using comparative analysis with other tissue data from GTEx and Epigenomics Roadmap, for various adult brain phenotypes. 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. 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 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 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.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

### [\*\*\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

### [Figures](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

### [\*\*\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Figure 1 data](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Figure 2 brain specific genomic aspects](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

* [PCA, RCA](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Chromatin (brain clusters by enhancers)](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [DEX genes (brain, disease, region,...)](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [non-coding/TAR](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Table 1 summary table of brain specific genomic elements](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Figure 3 QTLs](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

* [eQTLs](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [cQTLs](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Overlap with enhancer, promoters, TFBSs](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Figure 4 integrative model](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

* [SNPs (QTLs) to enhancers(chrom.) to gene expression](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Genes to modules to traits](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Single cell](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Website: (e.g.,](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing) [[https://www.encodeproject.org/comparative/](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)](https://www.encodeproject.org/comparative/)[)](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\* old single cell analysis\*\*\*\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

* [Novel and known cell types, biomarker genes, and signatures including neuronal (Lake 2016), non-neuronal (PNAS 2015) and Nenad’s 900 cells](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Deconvolution (or decomposition) to find cell populations, and associate population changes with phenotypes (gene expression, etc…)](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[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 some degree, this issue can be addressed using single cell gene expression data. Therefore, we integrated the single cell transcriptome data from PsychENCODE and others and discovered the potentially novel cell types along with biomarker genes that do not match existing neuronal and non-neuronal cell types. We further 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 show that the gene expression differences across brain tissues can more easily be explained by the changes of cell populations.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[We integrated single-cell data from PsychENCODE and other studies to determine whether particular cell types drive gene expression changes across tissues. We “de-convolved” the tissue-level gene expression data using single-cell data to find the populations of different cell types corresponding to different phenotypes. We found many gene expression differences were more easily explained by cell population changes.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\*\* old modeling\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

* [Integrating genotype, transcriptomics, epigenetics, regulatomics](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Gene regulatory networks explaining how QTLs affect gene expression](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [eTWAS embedding GRN to better predict genotype-phenotype](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Intermediate modules are enriched with bio function and pathways](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Imputate gene expression/enhancers using genotypes only](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Additional bullets](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Finally, we built an integrative model to integrate all the functional genomic data types in this resource and understand how the brain genomic variants affect gene expression and regulation, and eventually predict the phenotypes (Figure C). This model also allows us to quantitatively impute the missing transcriptional and epigenetic information given genotype data only. In particular, we first inferred the gene regulatory networks consisting of QTLs, enhancers, transcription factors and target genes using the genotype, RNA-seq and ChIP-seq data. This gene regulatory network explained the regulatory mechanisms on how QTLs affect gene expression. We then built a Restricted Boltzmann Machine (RBM) based on this gene regulatory network to predict the brain genotype-phenotype relationships. Specifically, this RBM 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 predictive relationships between layer nodes. We show that this integrated model has significantly improved the prediction accuracy over individual genomic data types and relates these predictions to well characterized functions and pathways (e.g., intermediate modules in RBM) in the brain. We also make the model available as a set of distributive software from the resource.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[We built an integrative epigenome- and transcriptome-wide association model (eTWAS). This integrated our high-dimensional functional genomics data with genotype and phenotype data to relate how variants affect gene expression and regulation to brain phenotypes (Figure C). This model allowed us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only. We first inferred gene regulatory networks that show how QTLs, enhancers, and transcription factors relate to target gene expression. We then built a Deep Boltzmann Machine-based eTWAS model (available online) that directly embeds regulatory network information to predict genotype-phenotype associations with significantly improved prediction accuracy over individual data types. This model identified intermediate-layer modules (i.e., strongly predictive features) that correspond to known gene sets associated with well-characterized pathways in the brain.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\* old introduction\*\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Gene expression elucidates functional impact of](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[polygenic risk for schizophrenia](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[[https://www.nature.com/articles/nn.4399](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)](https://www.nature.com/articles/nn.4399)

[[https://www.nature.com/articles/nn.4156](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)](https://www.nature.com/articles/nn.4156)

[The PsychENCODE Consortium is a group of projects that “aims to produce a public resource of multi-dimensional genomic data using tissue and cell-type specific samples from approximately 1,000 phenotypically well-characterized high quality healthy and diseased human post-mortem brains, as well as functionally characterize disease-associated regulatory elements and variants in model systems”(6). 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. These data are complemented by a number of other large-scale genomic resources, such as ENCODE, GTEx, Roadmap, BrainSpan, and CommonMind, which provide valuable contexts for additional human organs and tissues. As part of the activities of this Consortium, we integrated these datasets and generated a high power eQTL map of adult frontal cortex by combining ~2,000 samples from BrainSpan, GTEx, CommonMind, PsychENCODE, and other available sources.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[[[GTEX, encode, roadmap,cmc - how these position us for the brain ?]]](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[[[the problem - for psych disease - we have g-p but not mechanism ]]](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\*\*\*\* old resource\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

* [Figure 1](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Base resource & derived [cqts]](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Assay summary (RNA-seq, ChIP-seq, genotype, …)](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Phenotype summary (2k samples, disease, gender, …)](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Functional genomics summary( SNPs, QTLs, regions, enhancers, genes, transcripts, ncRNAs, imputated networks,…)](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Resource summary (website, accessibility, app?...)](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [QTL analysis: eQTLs and cQTLs, and compare with GWAS](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[Understanding the molecular mechanisms that genomic variants change associated phenotypes in brain disorders is still a key challenge. The PsychENCODE consortium 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 many individuals with different phenotypes including normal, mental diseases. We integrate 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 genomics (~2000 samples in total) and compare it against various phenotypes. This resource comprises 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. We make the derived resource downloadable and available on the PyschENCODE website (xxxx).](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[\*\*\*\*\*\*\* old brain activites\*\*\*\*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

* [Figure 2](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Brain samples separate from other tissues by spectral analysis for gene expression and epigenetics](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [Functional genomic elements driving brain clusters: DEX genes, non-coding RNAs/TARs, enhancers, regulations, ...](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)
* [JW's stats](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[We then use this resource and compare against various phenotypes to reveal the unique brain genomic activities, particularly relating to transcriptomic and regulatory binding activities. In particular, we performed the spectral analysis for comparing the similarities of gene expression and epigenetic data with other tissue samples from GTEx and Epigenomics Roadmap (Figure A). Our analysis revealed that the samples can be clustered together from their tissue types using either protein-coding gene and non-coding RNA expression or epigenetic data; e.g., brain and other tissue clusters. It shows that brain samples separated more significantly from other tissues in terms of gene expression. Additionally, we found that the brain has more non-coding transcriptional activity than most other tissues. These results suggest that the brain has specific and distinct expression and epigenetic profiles. These brains related clusters and activities that the brain has specific and distant expression and epigenetic profiles and transcription, involved by the brain elements in the resource. Furthermore, we identified 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 (Figure B). These variants cover a larger fraction of disease-associated brain GWAS SNPs than any previous analyses, suggesting potential molecular targets for these associations.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)