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

**Abstract (214)**

Understanding how a variety of genomic variants influence brain phenotypes and disorders remains a key challenge. To this end, the PsychENCODE consortium has generated large-scale sequencing datasets on the brain at the population level over 1866 adults with various phenotypes and diseases, including genotype, transcriptomics, chromatin and single-cell. By uniformly processing and consistently analyzing the datasets, we developed a comprehensive resource on functional genomics of adult brain. We deconvolved the bulk tissue expression using single cell data and found the cell fraction changes can explain >85% tissue covariance and also associate with psychiatric phenotypes. Moreover, we found that the brain differs from other tissues more notably in gene expression than chromatin activities. We used the chromatin and HiC data for reference brain samples to generate ~80k active enhancers on adult prefrontal cortex, and link them to genes, transcription factors and identified ~2M eQTLs and xx cQTLs. In addition, we built a full gene regulatory network linking all possible genomic elements, and showed that it can discover 112 more genes associated with GWAS variants with psychiatric disorders. Finally, we developed a deep-learning model embedding the regulatory network to connect genotype and phenotypes on multiple layers, achieving ~4.6X improvement in disease-trait prediction above an additive genetic model. This model enables highlighting intermediate genes and functional modules implying potential mechanisms, and also imputation of missing transcriptome and epigenome from genotype data only.

**Introduction (423)**

Disorders of the brain affect nearly a fifth of the world’s population \cite{19507169}. Decades of research has led to little progress in our fundamental understanding of the molecular causes of psychiatric disorders. This contrasts with cardiac disease, for which lifestyle and pharmacological modification of environmental risk factors has had a profound effect on disease morbidity , or cancer, which is now understood to be a direct disorder of the genome \cite{9603539, 24071849, 16461820, 19360079, Jim}. Though GWAS studies have identified many genomic variants associated with psychiatric disease risk, a detailed understanding of the precise molecular mechanisms behind these associations remains elusive \cite{19339359, Jim}.

To this end, a number of genomic studies have recently focused on discovering genomic variants relating to psychiatric diseases. As such, a variety of common and rare variants and linked genes have been found to be associated with brain and psychiatric disorders \cite{26404826}. For instance, the Psychiatric Genomics Consortium (PGC) identified 142 GWAS loci associated with schizophrenia \cite{29483656}. Many variants for psychiatric disorders have been found to lie in non-coding regions \cite{26404826}, suggesting potential roles in gene regulation. Other consortia have annotated many of these non-coding elements -- e.g., expression quantitative trait loci (eQTLs) and eGenes in GTEx, and enhancers from ENCODE and Epigenomics Roadmap that are associated with various cells and tissues. Though some of these elements relate to the brain, none of these groups have specifically tailored their efforts toward comprehensively identifying functional elements in the brain. Pioneering work from the CommonMind consortium \cite{27668389} has attempted to do this but has not reached maximal scale. Moreover, a number of new technologies have recently been developed to better measure aspects of gene regulation. For example, HiC studies have been used to identify specific chromatin structural and regulatory elements, such as active enhancers in the fetal brain \cite{ 27760116}. Single-cell sequencing techniques also offer great promise for studying the transcriptome \cite{27339989, 26060301}. However, each of the studies that leverage such technologies have generally focused on individual aspects of brain functional genomics. Therefore, these data have not yet been fully integrated at scale.

To this end, we have built a central, publically available resource for adult brain functional genomics, including all the raw and uniformly processed data at both tissue and single cell levels from PsychENCODE and other related major projects \cite{27339989, 26060301} with nearly ~12k data samples in total. By leveraging this resource, our analyses identified various functional genomic elements and QTLs specific to the adult brain, including novel psychiatric GWAS and gene linkages. We also combined these elements and built an integrated deep-learning model which enables us to leverage the richly structured data of the resource to learn non-linear interactions between genotype and molecular phenotypes at multiple layers, as well as predict high-level traits.

**Resource construction (205)**

We designed this resource to provide a coherent and comprehensive structure to the data (adult.psychencode.org) \cite{Supplement}. Broadly, it organizes a large amount of data for brain functional genomics pyramidally, with a large base of raw data files (much of it restricted-access, such as individual genotyping and raw next-generation sequencing data of transcriptomics and epigenomics), a middle layer of uniformly processed and easily shareable results (such as open-chromatin "peaks" and gene-expression quantifications), and a compact cap at the top, consisting of an integrative model (based on imputed regulatory networks and QTLs). As shown in Figure 1, to build the base layer, we included all the datasets from PsychENCODE related to the adult and merged these with other relevant data from ENCODE, CommonMind, GTEx, Epigenomics Roadmap, and recent brain single cell studies. In total, this resource constitutes up to ~5.5k data samples derived from 1931 individual adult brains from multiple cohorts of PsychENCODE and ~5k brain related samples from other consortia, which covers a large representation of brain phenotypes and psychiatric disorders. Furthermore, the PsychENCODE project developed a specific "reference brain" project utilizing many assays on the same set of brain tissues, which we used to develop an anchoring annotation for the entire resource \cite{Supplement}.

**Bulk & single-cell transcriptome analysis (701)**

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

Brain tissues have been found to comprise a variety of cell types, including neuronal and non-neuronal cells. Previous studies have suggested that gene-expression changes at the tissue level can be associated with changing proportions of cell types \cite{21614001, 29439242,18849986, 27409810}. However, they have not systematically revealed how different cell types can quantitatively contribute to the tissue-expression variation. Here, we address the question whether expression changes over our population of 1866 adults are driven by gene expression changes in a specific cell type or whether they result from changing proportions of various cell-types.

We used two complementary strategies. First, we used the standard pipeline to uniformly process single cell RNA-seq data in PsychENCODE, in conjunction with a number of other single-cell studies on the brain, in order to assemble a list of basic cell types for PsychENCODE in the brain. This list includes 16 previously identified neuronal types, 5 major non-neuronal types and a number of additional types in brain development \cite{Supplement}, and constitutes a matrix (C) of expression signatures, which are mostly concordant with what has been published (Figure S5 and Discussion). Across these, we found that a number of genes whose expression levels vary more substantially than they do across individuals in a population (e.g., the dopamine receptor gene DRD3, Figure 2A). This implies that small variations in cell types can readily give rise to substantial changes in bulk gene expression at the tissue level.

To explore this further, we performed an unsupervised analysis for bulk tissue expression data to identify the primary components as they relate to different single cell types. In particular, we decomposed the bulk gene expression matrix (B) from our resource using non-negative matrix factorization (NMF, Figure 2B, see Methods); i.e., B~=VH, and then determined whether the top components of the NMF (ie, NMF-TCs) that capture the majority of covariance are consistent with the single cell signatures. As shown in Figure 2B, we found a number of NMF-TCs highly correlated with neuronal, non-neuronal and developmental-related cell types (Figure 2C). This demonstrates that an unsupervised analysis derived from bulk data roughly matches the single cell data, partially corroborating our basic cell types.

We then want to understand how the differing cell types contribute to variation in bulk gene expression and relate to different brain phenotypes. In particular, as shown in Figure 2B, we de-convolved the expression matrix of tissue, B using the single cell signatures (C) to estimate the cell fractions W, solving the equation “B~=CW” (See methods). As a validation, we found that our estimated fractions of NEU+/- cells match the experimentally determined fractions for the reference brain samples (Median error = 0.04 , Figure S6). Moreover, we found that using the single-cell expression signatures can explain much of the population-level expression variation (i.e., across tissue samples of the same brain region taken from different individuals); i.e., 1-||B-CW||2/||B||2>85% (Methods).

Furthermore, we found cell fraction changes were associated with different phenotypes and psychiatric disorders (Figure 2E, SXXX). For example, particular excitatory and inhibitory neurons exhibit significantly different fractions between male and female samples (Ex3 and In6). The fraction of Ex3 is also significantly reduced in ASD samples (p=2.73e-11, ANOVA test), while non-neuronal cells (e.g., oligodendrocytes) are represented in greater abundance. Another interesting association was that cell fractions change with age. In particular, the fractions of neuronal types Ex3 and Ex4 significantly increase with age, but some non-neuronal types such as oligodendrocyte are found to decrease. Furthermore, these age-related cell changes are potentially associated with differentially expressed genes across age groups – e.g., the gene involved in early growth response is down-regulated in older age groups, whereas ceruloplasmin is down-regulated among middle-aged groups (Figure 2E).

[[{Questions on single cell types for Dan & Jim}]]

**Enhancers in adult brain (417)**

We use chromatin modification signals to identify active enhancers in the brain. The reference brain is valuable here as it allows integration of many assays, including ATAC-seq, ChIP-seq and Hi-C. Combining the data from the reference brain with DNase and ChIP-seq of the same brain region from ENCODE and Roadmap Epigenomics, we identified a set of ~80k brain PFC enhancers, using a consistent approach with that in ENCODE (See details in the supplement)

 - i.e. the identified enhancers are enriched in H3K27ac signal and depleted in H3K4me3 signal (Figure 3A).

Assessing the variability of enhancers across individuals and tissues is more difficult than evaluating the variability of gene expression. As shown in figure 3A, not only the chromatin signal level associated with the enhancers goes up and down across population, the boundaries of enhancers grow and shrink and sometimes disappear altogether -- or re-appear in different places in a few individuals. These 1D signal changes would fundamentally reflect change in 3D chromatin structure. To investigate the variability within population, we first aggregated the H3K27ac peaks from the individuals in the cohort. The H3K27ac ChIP-seq data from prefrontal cortex (PFC), temporal cortex (TC) and cerebellum (CBC) of 50 individuals was uniformly processed followed by peak calling using the same pipeline (see supplement). Aggregation among the population gives in total 37,761 H3K27ac in PFC, 42,683 in TC, and 26,631 in CBC. While these numbers are smaller than the number of enhancers, they actually cover a larger fraction of the genome, including enhancer and promoter regions, as the average width of H3K27ac peaks is larger than that of enhancers. Comparing aggregate set of these three brain regions, PFC shares a more similar chromatin signature with temporal cortex (~90% overlap between H3K27ac peaks) and is largely different from that of cerebellum (34% overlap), consistent with previous report (\cite{27863250}).

We also examined the overlap of the reference brain enhancers with the H3K27ac peaks in each of the individuals. As expected, not every active enhancers in the reference brain are active in every individual in the cohort. In fact, on average ~55K out of 80K active enhancers in the reference brain are active in another individual (Figure 3B). As expected, only a core set of enhancers is ubiquitously active in every single person in the cohort, with a larger fraction (~68%) of the enhancers identified in the reference brain active in more than half of the population. We estimate that ~30% of one’s PFC active enhancers is individual specific and/or related to environmental factors. This can be seen from figure SX that the cumulative number of active enhancers increases dramatically for the first 20 individuals sampled, but becomes saturated at the 30th sample. We hypothesize that pooling together the identified prefrontal cortex enhancers from 20~30 individuals is enough to cover all potential regulatory enhancers in human PFC, which we estimate to be around 120K.

**Consistent comparison of transcriptome and epigenome (309)**

A key aspect of our analysis is that we uniformly processed the transcriptomic and epigenomic data across PsychENCODE, ENCODE, GTEx and Roadmap. This allows us to compare the brain to other organs in a consistent fashion and also to compare the consistency of this over transcriptome and epigenome. We attempted several methods including PCA and tSNE for an appropriate comparison, and finally used Reference Component Analysis (RCA). PCA, though popular, tends to capture global structures, ignoring most of the local structure, but it can easily be influenced by outliers. On the other hand, t-SNE analysis preserves local structure but “shatters” global structure \cite{Supplement}. RCA, however is capable of capturing local structure while maintaining meaningful distances in global structure space. It projects the gene expression in an individual sample against a reference panel, and then essentially reduces dimensionality of the individual projections.

Our comparative analysis for gene expression shows that the brain tends to separate from the other tissues in the first component. Inter-tissue differences are much larger than intra-tissue ones. A different picture emerges when one looks at our comparison using chromatin data. It shows that the chromatin levels at enhancers are much less distinguishable between brain and other tissues (Figure 3E).

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

**QTL analysis (453)**

We used the PsychENCODE resource data to identify quantitative trait loci (QTLs) affecting gene expression and chromatin activity. In particular, we calculated: expression QTLs (eQTLs), chromatin QTLs (cQTLs), splicing isoform QTLs (isoQTLs) and cell fraction QTLs. For eQTLs, we adopted a standard approach, adhering closely to the established GTEX eQTL pipeline. We identified 2,542,908 eQTLs (2,097,741 eQTLs after LD pruning) and 32944 e-genes (including non-coding ones) in DLPFC. There are 1,341,182 (25%) unique snps involved in the eQTLs from the 5,297,875 potential snps (within 1 Mb of gene TSS). This conservative estimate is a substantially larger number of eQTLs and eGenes than previous brain eQTL studies such as CommonMind and reflects a large sample size and statistical power in psychencode \cite{Supplement}. We believe the eQTL number is close to saturation, in terms of associating almost every variant with some expression modulating characteristic. We also applied the same QTL calculation pipeline to splicing and identified 157,592 isoQTLs.

For the cQTLs, the situation is more complicated. There are no established standard methods for calculating these on a large scale, though previous efforts have detected QTLs associated with various chromatin activities on non-brain context \cite{25799442, 26300125}To properly identify them, we focused on a reference set of enhancers to define the region associated with chromatin activity and then looked at how this activity varies in these enhancers across 292 individuals(See methods). Overall, we were able to identify ~2000 cQTLs in addition to the 6200 cQTLs identified using individuals from the CMC cohort \cite{https://doi.org/10.1101/141986}.

Next, we were interested to see if any SNVs were associated with changes in the relative fractions of cell types across individuals; i.e., cell-fraction QTLs (fQTLs). In total, we identified the 3720 distinct SNPs constituting 4186 different fQTLs between different cell types. Of these the fractions of microglial and a particular type of excitatory neuron (Ex8) were associated with the most variants. Next, after factoring out these cell type differences, we also identified 260,280 SNPs significantly associated with the gene expression changes across individual tissues - these "residual trans-eQTLs" represent SNP-expression associations unexplained by variation in cell types.

To further dissect the associations between genomic elements and the QTLs, we intersected our QTL lists with each other and also with a comprehensive set of genomic annotations. (Figure 4). For the overlap among different QTLs, we originally expected that many of the other QTLs would be a subset of the very large number of eQTLs but in fact an appreciable number of isoQTLs, fQTLs and cQTLs were actually not overlapped with any eQTLs as shown in (Figure 4C). As expected, we found that eQTLs tended to enriched on promoter regions and cQTLs were mostly enriched on enhancer regions and fQTLs tended to occur broadly in many different elements (Figure 4D).

**Regulatory networks (421)**

In this section, we provided an integrative analysis at the gene regulatory level for the data and genomic elements described above. We created a putative regulatory network revealing how the genotype and regulators control target gene expression in adult brain. To this end, we first processed a Hi-C dataset for adult brain in the same reference samples used for enhancer identification, providing potential physical boundaries of interactions between enhancers and promoters (Figure 5A). \cite{27760116, Supplement}. In total, we identified 2,735 topologically associating domains (TADs) which set potential physical boundaries of enhancer-promoter interactions. This HiC dataset is substantially different than the fetal brain HiC data set highlighting the importance of the developmental stage in chromatin topology \cite{Supplement}.

As expected, we found that ~75% of enhancer-promoter interactions occur in the same TADs (Figures Sxx, 5B), suggesting that the majority of cis-regulation occurs within TADs. Also, as expected, the genes that have more potentially regulatory enhancers interacting with their promoters tend to express higher (Figure 5B). We next integrated the Hi-C dataset with eQTLs to assess how much of the common variation-associated gene regulation is mediated by chromatin interactions. Interestingly, 32% of eGenes show evidence of chromatin interactions, accounting for 239,837 eQTLs and 3,235 isoQTLs (Figure 5C). To our surprise, enhancer e/isoQTLs supported by Hi-C evidence showed stronger associations than exonic and promoter e/isoQTLs.

Distinguishing from HiC and eQTLs that reveal potential regulatory relationships in terms of chromatin interaction and genomic variation, we tried to predict another type of gene regulatory relationship based on input-output gene expression from TFs to target genes. In particular, we first found TFs having binding motifs on target genes’ promoters as well as potentially interacting enhancers \cite{Supplement}, providing a reference wiring network for at large potential TF-target regulatory linkages. Second, based on these “wiring” relationships, we used elastic-network regression to find the TFs-target relationships likely control target gene expression (i.e. associated with a high coefficient for predicting the target gene expression from TF activity). We model them as simple linear relationships but regularize to minimize the number of connections (Methods). Overall, we found this model could successfully predict expression of xxx genes with the minimum mean square errors < xxx (Figure Sxx). For example, the expression of schizophrenia risk gene, DGCR2 can be predicted by its TFs expression with MSE<xx based on our model. To complete building gene regulatory network, we used the imputed TF-gene linkages to filter potential enhancer-target gene linkages, finding a robust set of enhancer-target gene linkages.

Finally, we generated a full regulatory network, linking all possible regulatory elements such as enhancers, TFs, and target genes (Figure Sxx). In total, this network has scale-free and hierarchical structures (Figure Sxx), and consists of ~150K HiC-based linkages (enhancer-promoter), ~2M variant-based linkages (eQTLs-eGene) and xxx TF-target and xxx enhancer-target linkages based on input-output gene expression model.

**Associating GWAS variants with genes (285)**

Next, we tried to use the regulatory network to link GWAS variants to potential target genes. First, we found significant associations between our eQTLs and GWAS associated with brain diseases. In particular, we calculated the overlap enrichment of our brain cis-eQTL SNPs and GWAS SNPs for schizophrenia, bipolar disorder and parkinson’s disease and compared this to that for non-brain related disorders (CAD, asthma and type 2 diabetes). As expected, enrichment for GWAS SNPs for brain disorders are more significant, with schizophrenia GWAS SNPs having the highest enrichment (Figure 4E).

To further predict the target genes of GWAS variants we exploited all possible linkages from our full gene regulatory network including HiC, eQTL and input-output expression relationships. For example, for the newly identified 142 schizophrenia GWAS loci \cite{27869829}, in total, we identified a set of 488 putative schizophrenia-associated genes, hereby referred as "SCZ-genes", and 99 genes that show evidence both at the level of Hi-C and eQTLs [[&& actifity]], providing a high-confidence subset (Figure 5xx). This is a huge increase from the previously annotated 22 genes across 19 loci based on CMC brain eQTLs \cite{27869829, 27668389}. The majority of SCZ genes (288 genes, ~59%) were not in linkage disequilibrium (LD, r2>0.6) with index SNPs (Figure 5E), consistent with the previous observations that regulatory relationships often do not follow linear genome organization \cite{xxx}.

We then looked at the characteristics of the 488 SCZ genes. As expected, these genes shared many of the characteristics of known schizophrenia-associated genes. In particular, they are enriched in genes intolerant to loss-of-function mutations \cite{27869829}, translational regulators, cholinergic receptors, calcium channels, and synaptic genes and genes known to be differentially expressed in schizophrenia (Figure Sxx). Next, we further integrated this gene list with the single-cell profiles and found, interestingly, they were enriched in a variety of different neural types.

99 genes that show evidence both at the level of Hi-C and eQTLs [[&& actifity]]

**Integrative deep-learning model (830)**

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

Using the full model with genome and transcriptome data provided, we show that adding the extra layers of structure in the DSPN allows us to achieve substantially better prediction of disease and other observed traits than without -- ie just using a traditional genotype to phenotype model (Figure 6d). In particular, we achieve a cross-disorder average prediction of 67.4% with the extra layers vs 59.3% without (Figure 6d, DSPN vs LR-trans). Further, comparison with a simple logistic predictor from the genome alone shows that the transcriptome carries significant further trait information, which the DSPN is able to extract. For instance, a logistic predictor is able to gain a 2.4X cross-disorder improvement when using the transcriptome versus the genome alone (+9.3% vs. +3.8% from 50% random chance), while the DSPN is able to gain a 4.6X improvement (+17.4% vs. 3.8%); this may reflect the fact that the DSPN is able to incorporate non-linear interactions between intermediate phenotypes at multiple layers (in contrast to LR). Moreover, the DSPN also allows us to perform joint inference and imputation of the intermediate phenotypes (i.e. transcriptome and epigenome) and observed traits from the genotype alone. Using imputation, we achieved 56.6% cross-disorder accuracy for disease trait prediction averaged across all 3 conditions, which is better than the performance using direct prediction with a logistic model (53.8%) or ~1.7X gain in accuracy above 50% chance (+6.6% vs. +3.8%, see Figure 6d, DSPN-Imput vs LR-gen). These results demonstrate the usefulness of even a limited amount of functional genomics information for unraveling gene-disease relationships and show that the structure learned from such data can be used to make more accurate predictions of high-level traits even when absent.

We transform the results above to the liability scale in order to compare with narrow-sense heritability estimates (using GCTA, see supplement and Fig. 6d). Prior studies have estimated that common SNPs explain 25.6%, 20.5% and 19% of the variance for Schizophrenia, Bipolar and Autism respectively \cite{bioRxiv\_10.1101/048991}. These may be taken as limits which an additive predictive model using the common SNPs will approach given unlimited training data, while a non-linear predictor can potentially exceed these. Liability scores for the LR-gen and DSPN-Imput models are substantially below these levels, implying that incorporating further data will be beneficial. The variance explained by the full DSPN model however is of a similar order to that explained by common SNPs for all three conditions (16.3%, 30% and 14.4%), suggesting that improved imputation of intermediate phenotypes may capture most of the variance due to common SNP narrow-sense heritability. We note however that the full DSPN may also be capturing environmentally determined and trait-influenced variance (Supp Fig. A1), meaning these are upper-bounds on the genetically determined variance available through imputation in the current model.

We examined the specific connections learned by the DSPN between intermediate and high-level phenotypes for potentially relevant biological interactions. We included known co-expression modules in the model and examined which of these the DSPN prioritized, as well as new sets of genes associated with the DSPN latent nodes that were uncovered at each hidden layer, using a multilevel prioritization scheme (Supplement Sec. 7.5 and Supp Fig. A2).

Overall, we can show that the modules prioritized by the DSPN are strongly enriched for known GWAS variants (Supplement). We provide a full summary of the functional enrichment analysis for the prioritized modules associated in supplement (Supp section xx) and highlight some of the associations found using the Schizophrenia model (Figure 6c). Among the highest SCZ prioritized modules/gene-sets, we found enrichments for (i) glutamatergic-synapse pathway genes, (ii) calcium-signaling pathways and astrocyte-marker genes, and (iii) complement cascade pathway genes including C4A, C4B and CLU -- confirming and extending previous analyses \cite{26814963}. Further, we found enriched in the prioritized modules for aging the marker genes for Ex4 cell-type and the gene NRGN (in a module associated with synaptic and longevity functions), consistent with our differential expression analysis (Fig XXXX).

**Discussion (206)**

We integrated PsychENCODE datasets with other resources and developed a comprehensive resource of various functional genomic elements for the adult brain. Overall, our study has identified a large-scale set of eQTLs for adult brain, several folds more than previous studies, almost achieving saturation for protein coding genes. We suspect that larger population studies will not significantly expand on these. However, there exist other aspects of brain QTLs that can be extended in the future. For example, increasing individual sample size of chromatin data size and quality such as identifying enhancers using STARR-seq will help discover more accurate cQTLs.

Another area of future development is single-cell analysis. In this study, we found that a set of basic and known cells could explain large expression variations across tissues. However, there still exist the gene expression heterogeneities among these cell types (Figure Sxx), implying subdivision into potential additional types in the future. Also, current techniques suffer from low capture efficiency, and so it remains challenging to reliably quantify low-abundant transcripts/genes \cite{26949524, 25053837}, and particularly have problems for some specific cell types such as axons and dendrites; e.g., very noisy expression measurements \cite{25053837}. Thus, we may see the novel techniques such as single-nucleus RNA sequencing techniques that emerge to address these problems. Finally, our deep-learning approach is readily extensible to modeling complex genotype-phenotype relationships involving other kinds of intermediate phenotypes, and can naturally embed new types of QTL and phenotype-phenotype interactions.

 [[could add something here... perhaps about chromatin variability]]

**Figures**

**Figure 1. Comprehensive data resource of functional genomics in adult brain**. The functional genomics data generated by the PsychENCODE consortium (PEC) constitute a multidimensional exploration across tissue, developmental stage, disorder, species, assay, and sex. From this larger corpus of PEC samples, we focused on adult datasets, integrated with those from consortia such as GTEx, the Roadmap Epigenomics Consortium, ENCODE, the CMC and HBCC studies, and previously published single-cell transcriptomic data. The central data cube represents the results of this integration for the three dimensions of disorder, assay and tissue, where only the numbers of datasets used in the current analysis are depicted. Projections of the data onto each of these three parameters are shown in graph form for assay and disorder, and in schematic form for the primary brain regions of interest. **Assay:** The bars represent datasets across a subset of the assay types, including RNA-seq (N = 2040 PEC + 1632 uniformly processed GTEx samples), genotypes (N = 1753 PEC + 113 GTEx = 1866 individuals), scRNA-seq (N = 932 PEC + 3693 external datasets), and H3K27ac ChIP-seq (= 408 PEC + 5 uniformly processed Roadmap samples). **Disorder:** The number of individuals under the control category include the 113 from GTEx and 926 from PEC, while individuals from PEC provide data on the remaining disorders of schizophrenia (SCZ, N = 558), bipolar disorder (BPD, N = 217), autism spectrum disorder (ASD, N = 44), and affective disorder (AFF, N = 8), resulting in a total of 1866. **Tissue:** In this schematic, we focus on the datasets derived from three primary brain regions evaluated in our integrative study, viz., the prefrontal cortex (PFC, N = 3521), the temporal cortex (TC, N = 2153), and the cerebellum (CB, N = 348).

**Figure 2. Deconvolution analysis of Bulk and single cell transcriptomics reveals cell fraction changes across tissue phenotypes and disorders**. **(A)** The genes have significantly higher expression variability across single cells than tissue samples. Left: dopamine gene, DRD3; Right: Age gene, xxx. **(B)** Top: the bulk tissue gene expression matrix (B, genes by individuals) can be decomposed by non-negative matrix factorization (NMF) to the product of two matrices: NMF component matrix (V, genes by top NMF components; i.e., NMF-TCs) and component fraction matrix (H, top NMF components by individuals); i.e., B~=V\*H. Bottom: the bulk tissue gene expression matrix B can be also deconvolved by the single cell gene expression matrix (C, genes by cell types) to estimate the cell fractions across individuals (the matrix, W); i.e., B~=C\*W. Three major cell types are neuronal cells (blue), non-neuronal cells (red), developmental (dev) cells (green), as highlighted by columns groups in C (also row groups in W). **(C)** The heatmap shows the Pearson correlation coefficients of gene expression between the NMF TCs and single cell types for the biomarker genes (N=xxx). For example, NMF-15 highly correlates with Ex3 cell type (r=xxx). **(D)** The estimate cell fractions contribute >85% bulk tissue expression variations; i.e., 1-||B-C\*W||/||B||>0.85. **(E)** The cell fractions change across brain phenotypes and disorders. The neuronal cell types (e.g., Ex3 and In6) have significantly higher fractions in Male than Female samples (p<xxx, and p<xxx). Also, they and Oligodendrocytes have lower fractions in ASD than others. In particular, Ex3 cell fractions increase with aging (trend analysis p<6.3e-10).

**Figure 3. Comparative analysis for transcriptomics and epigenomics between brain and other tissues**. **(A)** Chromatin features of the reference brain are used to identify active enhancers. Enhancers are located in open chromatin region (high ATAC-seq signal), with strong H3K27ac signal and lack H3K4me3 signal. Enhancer activity varies among individuals, as indicated by the varying H3K27ac peak signal in the cohort. Each row corresponds to an individual, with shallow to dark blue indicating low to high signal peak value. **(B)** The number of brain active enhancers achieves a saturation to more than 70K with 20 samples. **(C)** The coefficients (PC1 vs. PC2) of RCA analysis for gene expression data of PsychENCODE samples (dark green), Other external brain samples (light green) including GTEx and Other tissue samples (magenta). **(D)** The center (cross) and ranges of different tissue clusters (dashed ellipse) on RCA scatterplot of (C). **(E)** The tissue clusters of RCA coefficients (PC1 vs. PC2) for chromatin data of active brain enhancers. The brain cluster (green ellipse) includes the PsychENCODE samples and other tissue clusters (magenta ellipses) consist of Epigeomics Roadmap samples. **(F)** The transcriptional diversity on coding (circle) and non-coding (triangle) regions among the tissue samples (inter-sample on x-axis) vs. on cumulative tissue samples (y-axis) for select major tissue types including Cerebellum (lightblue), Cortex (xxx), Lung (xxx), Ovary (xxx), Skin (xx) and Testes (xx).

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

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

**Figure 6. Deep-learning model predicts genotype-phenotype and reveals intermediate molecular mechanisms**. **(A)** The schematic outlines the model structures for Logistic Regression (LR), conditional Restricted Boltzmann Machine (cRBM), conditional Deep Boltzmann Machine (cDBM) and Deep Structured Phenotype Network (DSPN) models. Nodes are partitioned into four possible layers (L0-L3) and colored according to their status as (i) conditioning nodes visible during training and testing (light blue); visible nodes during training and possibly testing (dark blue); hidden nodes (grey). **(B)** shows the DSPN structure in further detail, with the biological interpretation of layers L0, L1 and L3 highlighted. The GRN structure learned previously is embedded in layers L0 and L1, with different types of regulatory linkages and functional elements shown. **(C)** shows examples of associations found: model traces are shown for three co-expression modules and associated higher-order modules prioritized by the DSPN Schizophrenia model, along with functional annotations enriched at each level. Genes, enhancers and SNPs associated with each module are shown. **(D)** summarizes the performance of different models, comparing performance as layers of structure are added to the model; for different predictors (genotype/transcriptome); and with or without imputation (colors highlight relevant models for each comparison). Performance accuracy on a balanced sample is shown first, with variance explained on the liability scale shown in brackets.

# **Supplement**

#### Supplement link: goo.gl/A6wmmN

#### Mar 5 ###

\* named version - "After-MG-edits-5Mar18"

PEC-Capstone4-draft--f228aK.gdocd.After-MG-edits-5Mar18.docx

in dropbox

\* global comments

- SNV v snps

- (Suppl) & (Meth) & (Fig) at end of sent.

- Headings to Sci. format

\* my comments (##MG##)[[ ]]

can we highlight more - color ?

\* careful wordcount of my version, removing comments

###################################

(##MG##)[[old HiC captions from Hyejung]] I hope the figure is self-explanatory, but here is a brief explanation.

A. Example Figure where enhancers, Hi-C, eQTL, and TADs all point to the same regulatory relationship.

B-C. eQTL analysis. B. eGenes supported by Hi-C vs. not. C. eQTLs that are supported by Hi-C evidence shows more significant P-values than those that are not.

D. Schematic showing how to identify putative target genes of schizophrenia GWAS loci using Hi-C and eQTL. Hi-C identified genes and eQTL identified genes are highly overlapping.

\*\*One thing to note: While we think the intersect (99 genes) will be the high-confidence gene list, they didn't show enrichment signals in most over-representation analysis as outlined in G (largely due to the small number of genes). Also, they didn't show significant gene ontology enrichment. Therefore, I used the union of Hi-C and eQTL gene lists for GO analysis. We should discuss whether we should leave the intersect list from G or just exclude this section.

E. Evidence showing eQTL and Hi-C point to the same gene, this case TSNARE1. I have other examples that I can show on Wednesday. Therefore, this is subject to change.

F. Gene ontology for the union list (see D for explanation).

G. Cell-type enrichment analysis of the intersect (green) and union (blue) lists. The union lists are enriched with neuronal genes and a specific subset of interneurons (CXCL14-CCK-RELN). If you want to make this consistent with the single cell data that you are using, I can run similar analyses. Let's discuss about this as well.

H. Brain regional enrichment analysis of the intersect and union lists, showing that both lists are enriched in prenatal cortex, while only union list is enriched in postnatal cortex.

I. The identified gene lists are enriched with genes affected by SCZ CNV, mutation-intolerant genes (EXAC pLI>0.9), genes in the ASD neuronal module (ASD M12), genes that are downregualted in SCZ, and genes in the modules affected in SCZ (SCZ M9c and M13c) - I think this might be most important one among SCZ figures.

 I also put some potential Supplementary Figures for

1. Loops located in the same TADs vs. not.

2. Comparing TWAS genes (Gusev et al., Nat Genet in press) vs. eQTL genes that we identified. I will explain the rationale behind this.

3. Comparing LD genes vs. eQTL genes that we identified: mainly showing many genes are not located in LD.

# **\*\*\*\*\*\* Notes on organization Jan 2018 \*\*\*\*\*\***

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

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

Article: ~4300 words, ~5 pages incl. Refs, etc (up to 6 figures )

\*\*\* sample word counts , avian genome model, modencode , funseq

**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

|  |
| --- |
|  **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 editgoo.gl/A6wmmN |

|  |
| --- |
| [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)  |

### [### brain call on Jan 28 2018####](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[goo.gl/f228aK](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[PEC-Capstone4-draft--f228aK.gdocd.After-call-28Jan.docx](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[1) write FN w/ todos: \*\* sat'd figure - unannotated transcription](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[As a supp figure we going to have FN's unnannotated done for genes](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

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

[2) tell MTG to get Jill et all the datasets asap - expl 1 brain DLPFC](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[3) where to put the aging - singlecell , transcriptome, model \*\*\*\*](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[4) XS todos : N+/- validation, age](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

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

[Main deadline - Wed. at 430 pm](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[But if you can get me some sections (say, intro & disc.) by Tue at 7 pm that'd be good](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

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

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

[Comments again the following Tue](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[3500 words in the results](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

[25% model](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

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

[Word counts of sections](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)

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

[(##MG##)[[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)