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

**Comprehensive resource and integrative model for functional genomics of the adult brain (~5800!!)[[keep doing]]**

**Abstract (212)**

Understanding how genomic variation influences brain phenotypes and disorders remains a key challenge. To this end, the PsychENCODE consortium has generated large-scale datasets on the adult human brain, including genotyping, RNA-seq, ChIP-seq, ATAC-seq, Hi-C and single-cell data on healthy and diseased tissues of thousands of individuals with different phenotypes. Using this, we developed a comprehensive resource on functional genomics of adult brain including ~2 million QTLs for expression and chromatin, and ~80k active enhancers. Leveraging the single cell data, we deconvolved tissue-level gene expression and found ~500 QTLs significantly associated with the cell fraction changes for various phenotypes **[[mention 85%]]**. Comparing this resource with others using spectral analysis, we show that the brain has unique expression and greater non-coding transcription than most other tissues, but not for chromatin. Moreover, we integrated the Hi-C and regulatory data to predict the gene regulatory network linking all possible functional genomic elements including QTLs, regulatory factors and target genes, identifying novel linkages between psychiatric GWAS SNPs and genes.**[[what is the key number?? 22 => 112 ]]** Finally, based on this, we developed a deep-learning model to predict genotype-phenotype associations with ~5X accuracy improvement achieved by integrating functional genomics data.**[[expl better]]** This model highlights intermediate genes and functional modules, revealing potential mechanisms, and also enables quantitative imputation of missing transcriptome and epigenome information from genotype data only.

**I. Introduction (550)**

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

To this end, a number of genomic studies have recently focused on discovering genomic elements relating to the phenotypes in adult brain. A variety including common and rare variants and genes have been found to be associated with brain and psychiatric disorders \cite{26404826}. For instance, the Psychiatric Genomics Consortia (PGC) identified 142 GWAS loci associated with schizophrenia \cite{25056061}.[[most variants for psychiatric disorders have been found to lie in non-coding regions \cite{26404826}}]]

Following a different path, other large consortia have also identified the reference sets of genomic elements across the entire body; e.g., eQTLs and eGenes in GTEx, and enhancers from ENCODE and Epigenomics Roadmap that are associated with various human cells and tissues. Though some of these elements relate to the brain, none of the consortia have specifically tailored their efforts toward comprehensively identifying the functional elements in the brain. **[[insert]]**There has been some pioneering work by CommonMind \cite{27668389} but this is just on rnaseq and has not reached maximal scale. **[[end-insert]]**

**[[cut next bit??]]** Moreover, in addition to eQTLs and eGenes, most variants for psychiatric disorders have been found to lie in non-coding regions \cite{26404826}. Thus, understanding the cell and tissue specific gene regulation by which these variants control brain gene expression is a critical next step to understand the molecular mechanisms of brain disorders.

**[[para doesn't fit]][[cut]]** To address this gap, recent technologies have started to detect the specific molecular activities within the brain, especially for the gene regulatory mechanisms that reveal how the genomic variants affect regulatory regions, control gene expression and drive the phenotypes and disorders. **[[end-cut]]**

Finally, there are a number of new technologies which have been developed but which have to be employed at scale and fully integrated… For example, recent HiC studies have been used to identify specific chromatin structural and regulatory elements, such as brain-active enhancers. Single-cell sequencing techniques also offer great promise for studying the transcriptome. However, each of the studies that leverage such technologies have generally focused on individual aspects of brain functional genomics; e.g., the Therefore, these data have not yet been fully integrated at scale.

**[[cut]]**In addition to the multi-facet data integration, larger sample sizes and more comprehensive assays are warranted to obtain a fuller view of brain-relevant functional genomics **[[refs]]**. **[[end-cut]]**

To this end, we have built a central, publically available comprehensive 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 including single-cell data [refs] with up to **XXX** samples. By leveraging this resource, our analyses identified various functional genomic elements and quantitative trait loci (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 to predict the molecular relationships from genotype to phenotype with high accuracy and impute missing data. The results obtained from this model are then studied in relation to specific brain phenotypes and psychiatric disorders.

## **II. Comprehensive resource for adult brain functional genomics (233)**

We designed this resource to provide a coherent and comprehensive structure to the data (http://adult.psychencode.org/). Broadly, it organizes a large amount of data for brain functional genomics pyramidally, with a large base of raw data files (much of it as restricted-access data, such as individual genotyping and raw next-generation sequencing data of transcriptomics and epigenomics), a middle layer of uniformly processed and shareable results (such as open chromatin peaks and gene expression quantifications, which are release without privacy restriction), 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 brain and merged these datasets with other relevant data from additional projects including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, and recent brain single cell studies. In total, this resource constitutes **XXX**X 3000 individual sets

data samples derived from 1931 individual adult brains from multiple cohorts, which covers a large representation of brain phenotypes and psychiatric disorders. The major data types include RNA-seq, ChIP-seq, ATAC-seq, HiC, single-cell data, and genotyping. (The later required large-scale imputation for all the PsychENCODE datasets, and we make full genotype sets available). Furthermore, the PsychENCODE project developed a specific "reference brain" project utilizing many assays on the same set of brain tissues, which we used to develop an anchoring annotation for the entire resource (Supplement).

**III. Bulk and single cell transcriptome analysis: deconvolution explains gene expression in terms of cell fraction changes (946)**

To identify the genomic elements that exhibit transcriptional activities specific to the adult brain, we used the ENCODE standard 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 (Supplement), sets of differentially expressed and co-expressed genes characterizing various brain regions, phenotypes and disorders, which are provided as part of our resource. In particular, the co-expressed genes are summarized as a list of gene co-expression modules for psychiatric disorders \cite{cap1} and for brain regions after clustering together with other tissues of GTEx (Supplement).

Brain tissues has been found to comprise a variety of cell types, including neuronal and non-neuronal cells such as astrocytes. Previous studies have suggested that differences in tissue gene expression is likely attributed to the gene expression variations across cell types in both healthy and diseased tissues \cite{18849986, 27409810} \cite{21614001, 29439242}. Therefore, one question with measuring gene expression changes over a population in our brain tissue samples is reliably determining whether the changes are driven by gene expression in a particular cell type or whether they result from changes in relative proportions of various cell-types. To address this, we integrated single cell transcriptome data to discover how the gene expression from various cell types contribute to bulk gene expression.

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 standard cell types in the brain (i.e., 16 neuronal types, 5 non-neuronal types and 4 additional fetal-related types from PsychENCODE; see Supplement). This list constitutes a matrix (C) of the gene expression signatures of 25 basic cell types, which are mostly concordant with what has been published, with some modifications (Figure S**xxx** and Discussion). (These basic single cell types are shared through the psychencode project ((capstones)))]]

**[[should we say that this is used in cap1 & devcap?]]** \*\*\*[[xS]]]Across these cell types, we found that a number of genes whose expression levels vary much more substantially than they do across individuals in a population (e.g., the dopamine receptor gene DRD3, Figure **xxx**). This implies that the variation of cell types can give rise to substantial changes in bulk gene expression at the tissue level.

To explore this further, we performed an unsupervised learning analysis for the 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, see Methods), and then determined whether the top components (TCs) of the NMF (ie, NMF-TCs) that capture the majority of covariance in the data and the 25 standard gene expression signatures of single cells are consistent. As shown in Figure XX, we found a number of NMF-TCs that are highly correlated with the gene expression signatures of neuronal, non-neuronal and development-related cell types. This demonstrates that an unsupervised analysis derived from the main components of the bulk tissue data roughly matches the single cell data, partially corroborating our standard 25 cell types. In addition, previous studies have identified cell type specific expression patterns from co-expression analysis \cite{18849986}. We found here that some of our NMF-TCs correlate with the eigengenes of gene co-expression modules \cite{cap1}, especially for the cell type modules, supporting again that they connect the cell type information from the bulk tissue data.

Second**[[is this second??]]**, previous studies found that the gene expression changes at the tissue level can be significantly associated with various cell types; e.g., the cell type specific gene co-expression modules \cite{18849986, 19829370}, but have not systematically and quantitatively revealed how different cell types contribute to the tissue gene expression changes. Thus, we next used the supervised method to estimate the cell fractions for tissue samples of individuals. In particular, as shown in Figure xx, we de-convolved the normalized bulk gene expression matrix of tissue, B using the single cell data matrix C to estimate the cell fractions W, by solving the equation “B~WC” (See methods). As a validation, we found **[[should we put here]]** **[[doesn't this get moved - doesn't fit here]]**Moreover, we found that our estimated fractions of NEU+/- cells match the experimentally determined fractions for the reference brain samples (r=**xxx**, Figure **xxx**), providing an additional source of corroboration.[**[end-move]]**

Moreover, We found that using the single cell expression signatures weighted by the derived cell fraction can explain much of the population-level expression variation (i.e., across tissue samples of the same brain region taken from different individuals). Specifically, we find 1-||B-WC||2/||B||2>0.85, where ||.|| is the Frobenius norm of matrix (Methods).

Furthermore, we found cell fraction changes were found to be highly associated with different phenotypes and psychiatric disorders (Figure **xxx**, Supplement for complete individual cell population estimates). For example, the excitatory and inhibitory neurons (Ex3 and In6) exhibit significantly different fractions between healthy male and female samples. The fraction of Ex3 cell types is also significantly reduced in ASD samples (p<**xxx**), while non-neuronal cells (e.g., oligodendrocytes) are represented in much greater abundance (Also reduced microglia fractions for Bipolar and increased astrocyte fractions for SCZ; see Supplement).

Another interesting association we found was that cell fractions change with age. In particular, the fractions of neuronal type(s) (Ex3 and Ex4) significantly increase with age (trend analysis p<6.3e-10 and 1.5e-6), but some non-neuronal types such as oligodendrocyte are found to decrease with age (p<2.1e-14). Furthermore, these age-related cell fraction changes are also potentially associated with differentially expressed genes across age groups (Figure **xxx**). For example, the gene involved in early growth response is down-regulated in older age groups, whereas the gene ceruloplasmin is down-regulated among middle-aged groups.

**IV. Active enhancers in adult brain (288)**

In addition to the transcriptome data, we uniformly processed chromatin data in the resource to give uniform quantifications, peak calling lists and single tracks for adult brain epigenomics.

In particular, we processed the ChIP-Seq (H3K27ac and H3K4me3) and ATAC-Seq data for the reference brain and identified a consistent set of ~80k brain active enhancers in prefrontal cortex, >90% of which overlap with the Epigenomics Roadmap annotation (Supplement). In addition, we have also developed reference sets of active enhancers in other brain regions including CBC (N=**xxx**) and ACC (N=yyy), overlapping DLPFC by ZZZ% (Supplement).

We then looked at the epigenetic signal variations across individuals at these enhancers. For this we checked the H3K27ac histone modification signal in DLPFC across 50 healthy individuals. We found that number of enhancers with active H3K27ac signal in an individual person varies largely from 20K to 70K (Mean = ~ 51K). \*\*\***[[discuss - too wide a range!]]** Only a small fraction (~7%) of these enhancers bear active H3K27ac signals across all samples, but the majority (~68%) are active in more than half of the population. This can be also observed from the cumulative numbers of active enhancers by increasing the sample size. For example, the cumulative number increase dramatically for the first 20 sample examined, but becomes nearly saturated at the 30th sample to more than 75K. Again, this suggests that enhancer activity varies across individuals, yet the majority of brain enhancers are active in most of the population. We also compared the distribution of the saturation curve on the normal samples (N=50) with the ASD samples (N=43) and found no significant differences in overall enhancer activity.

**V. Consistently comparative analysis reveals the brain related transcriptomic and epigenomic activity (474)**

One 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 in order to delineate gene expression and chromatin activities unique to the brain. We attempted several methods including PCA and tSNE for an appropriate comparison, and finally used the 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; \*\*\***[[thought we'd shorten & put in suppl.]]** e.g., it separates samples from the same tissue so that the cluster distances on t-SNE space are not proportional to real gene expression dissimilarities. It thus does not give a sense of overall effects.[**[end-cut]]** 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. We did RCA consistently for comparing brain and other tissues in terms of their similarities of both the transcriptome and the epigenome.

Our comparative analysis for gene expression shows that the brain tends to separate from the other tissues in the first component, showing it has a more distinct expression pattern and, and that all the brain tissue samples from the different projects tend to group together (which is a consequence of our uniformly processing). This difference is accentuated when focusing on the tissue cluster centers and the distributions surrounding them. Inter-tissue differences are much more accentuated than intra-tissue differences. A different picture emerges when one looks at our comparison using chromatin data (i.e., ChIP-seq signals on our consistent set of brain active enhancers). It shows that the chromatin levels are much less distinguishable between brain and other tissues (Figure **xxx**).

Our RCA analysis focuses on inter-tissue differences in well-annotated regions (i.e. genes, promoters and enhancers). In addition to the expression differences in protein-coding genes, a tremendous amount of transcriptional diversity is present across tissues in intergenic and noncoding regions. Thus, we looked at the overall level of transcriptional diversity across tissues. 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 (Figure SYYY sat’d for genes). However, when we shift to non-coding and unannotated regions, we find that brain tissues (such as cortex and cerebellum) do, to some degree, stand out by exhibiting greater transcription than most other tissues. \*\*\***[[last bit doesn't really relate much]]**This transcriptional diversity tends to increase with the number of samples (Figure **xxx** sat’d). Also, on the primate specific lncRNA regions \cite{ 24463510, 27919067}, the brain transcription even tops others**[[TBD, to suppl]]**.

**VI. QTL analysis (590)**

To understand how the genotype affects the transcriptome and epigenome in the adult brain, we used the PsychENCODE resource data to identify quantitative trait loci (QTLs) affecting gene expression and chromatin activity. In particular, we calculated the association of SNPs with normalized gene expression and chromatin state (Methods) to find quantitative trait loci associating with gene expression and epigenomic activity in adult brain, including several major categories: expression QTLs (eQTLs), chromatin QTLs (cQTLs), splicing QTLs (sQTLs) and cell fraction QTLs. For the eQTLs, we adopted a standard approach, adhering closely to the established GTEX eQTL pipeline. We identified 2,542,908 eQTLs **[[TBU by LD]]** and 32893 e-genes including non-coding genes in the DLPFC by using matched genotype and gene expression data of 1387 individuals. This conservative estimate is a substantially larger number of eQTLs and eGenes than previous brain eQTL studies such as CommonMind and reflects the very large sample size and statistical power we have; e.g., thousands in our resource vs. hundreds in CommonMind (Supplement). We believe this 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

 sQTLs.\*\*\***[[trans eqTL?? where??]]**

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 the activity of the chromatin and then looked at how this activity varies in these enhancers across individuals who have chromatin data available **[[UCLA\_ASD plus EpiDiff]]**, correlating this with nearby variants. (See methods). Overall, we were able to identify ~2000 cQTLs in addition to the 6200 cQTLs identified using individuals from CommonMind \cite{ https://doi.org/10.1101/141986}.

Next, we were interested to see if any SNVs were associated with changes in the fractions of various cell types. In particular, we used our QTL pipeline to identify distinct SNVs whose genotypes are significantly associated with differential cell fractions across individuals; i.e., cell fraction QTLs (fQTLs). In total, we identified the 3720 distinct SNVs constituting 4186 different fQTLs between different cell types. These were conservatively identified using a bonferronoi ....

Significant fQTLs are those with associated Bonferroni-corrected p-values of no more than 0.05 -- a very conservative cutoff.**[[move to suppl]]** Different cell types exhibit a great deal of heterogeneity in terms of their abundance within the set of high-confidence fQTLs. For instance, we identified 42, 58, and 195 significant fQTLs associated with the endothelial cells, astrocytes, and excitatory (class 4), respectively, but there were no significant fQTLs that were found to be associated with oligodendrocytes. **[[cut??]]** Moreover, we also identified **XXX** SNPs significantly associated with the gene expression changes across individual tissues after factoring out these cell type differences - these eQTLs represent SNP-expression associations unexplained by variation in cell types \*\*\***[[DC to update]]**]].

To further dissect the genomic elements associated with various QTLs we identified, we looked at how they overlapped and annotate them with a variety of different genomic annotations. The distributions of detailed QTL annotations across genomic regions are shown in Figure **xxx**. For example, we observed a significantly number of predictive QTLs break the TFBSs on the enhancers or promoters (xx%, Figure **xxx**), and also found **xxx** e-promoters on which eQTLs SNPs lies promotors of distal genes. **[[[more??]]**] As expected, there is a very large amount of overlap between the cQTLs, sQTLs, and eQTLs, and with ~50% of cQTLs also being eQTLs \*\*\***[[hyper overlap off ZZZ%]]**]. Also, fQTLs are very distinct from others, YYY of which overlap with trans eQTLs. \*\*\***[[can we say more?]]**

**VII. Gene regulatory networks in adult brain (540)**

In this section, we provided an integrative analysis at the gene regulation level for the data and genomic elements in the described above and predicted a gene regulatory network revealing how the genotype and regulators control target gene expression in adult brain. To this end, we first process a full, reference Hi-C dataset for adult brain, which provides direct physical evidence for potential interactions between enhancers and promoters (Figure 5A). Specifically, we generated and processed Hi-C data for the same reference adult brain that was used to identify the brain active enhancers, as previously described \cite{27760116} (Supplement). In total, we identified 2,735 topologically associated domains (TADs) which set potential physical boundaries of enhancer-promoter interactions and then 149,097 putative enhancer-promoter interactions in adult DLPFC. This HiC dataset is substantially different than the fetal brain HiC data set (see suppl) highlighting the importance of stage.

As expected, we found that ~75% of enhancer-promoter interactions occur in the same TADs (Figure 5xx), suggesting that TADs provide physical boundaries for cis-regulatory relationships between enhancers and target genes.\*\*\***[[how?]]** Also, as expected, the genes that have more enhancers **[[potentially??]]** interacting with their promoters tend to express higher (Figure 5xx).

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, 30.7% of e-genes show evidence of chromatin interactions, accounting for 204,008 eQTLs (Figure **xxx**). To our surprise, eQTLs supported by Hi-C evidence showed stronger associations not only to eQTLs without genomic annotations, but also to exonic and promoter eQTLs, highlighting the importance of incorporating chromatin interactions in deciphering regulatory relationships (Figure 5xx).

As a second step to build a full gene regulatory network, we integrated the TADs with other regulatory elements and relationships such as transcription factors (TFs) (Methods).\*\*\***[[didn’t we already integrate w enhancers and etg links]]** In particular, we used Hi-C data to find all possible enhancer-target gene relationships if enhancers and targets’ promoters are in the same TADs. We then found TF binding motifs using ENCODE data and imputed TF-target gene relationships if TFs have enriched binding motifs on the target gene’s promoters and enhancers. In total, we included ~3.2million enhancer-gene\*\*\***[[shouldn’t this be earlier]]**, **xxx** TF-gene **[[TBU]]** and ~1.8million eQTL-gene (FDR<0.05) regulatory linkages, providing a reference wiring network for gene regulation in brain.

Finally, using these “wiring” relationships, we inferred the final gene regulatory network linkages, which include the active regulatory links relating QTLs, enhancers, and transcription factors to target gene expression (Methods). This network also has a few particular characteristics such as scale-free and hierarchical structures, which have been revealed by previous network analyses (Figure Sxx). Given a target gene, we associated coefficients with each of these wiring linkages predicting the target gene’s expression from the activities of their regulatory elements. We model them as simple linear relationships but regularize to minimize the number of connections using an elastic net model (Methods). Overall, we found this model could successfully predict expression of >xx% genes with the minimum mean square errors < **xxx**. For example, the expression of gene, **XXX** can be predicted by its TFs expression with accuracy = zzz based on our model **[[[TBD]]**]. **[[cut]]**.

**VIII. GWAS (530)**

Intrigued by the regulatory map built upon Hi-C and eQTLs, we are further interested to predict potential target genes of GWAS variants with psychiatric disorders. First, we found significant associations between eQTLs and GWAS disease traits. In particular, we calculated the enrichment in cis-eQTL SNPs of GWAS SNPs of three brain related disorders (schizophrenia, bipolar disorders and parkinson’s disease) and non-brain related disorders (CAD, asthma and type 2 diabetes). As expected, cis-eQTLs SNPs have more significant enrichment for GWAS SNPs of brain disorders than non-brain disorder GWAS SNPs. Schizophrenia GWAS SNPs have the highest enrichment on Cis-eQTLs SNPs among those three brain disorders. Collectively, these QTLs annotate a larger fraction of GWAS SNPs involving the brain (e.g., 21% in schizophrenia, 18% in bipolar) than previously observed, providing important leads on which genes are affected in disease. This suggests that GWAS SNPs have potential target genes if they overlap with eQTLs.

Thus, to further predict the target genes of GWAS variants, as second step, we exploited the combined Hi-C and eQTL data and the gene regulatory linkages.\*\*\***[[all linkages??]]** For example, to identify putative target genes of newly identified 142 schizophrenia GWS loci \cite{27869829},

**[[move to suppl]]** we categorized 5,996 putative causal (credible) SNPs reported in the original study into promoter/exonic and intergenic/intronic SNPs. Promoter/exonic SNPs were directly assigned to the target genes based on the genomic coordinates, while intergenic/intronic SNPs were annotated based on chromatin interactions, which led to the mapping of 92 loci into 377 genes. Credible SNPs colocalize with 2,029 eQTLs associated with 83 e-Genes, 43 of which overlap with those identified by the Hi-C driven approach. To confirm that this overlap is mediated by the shared causal variants in GWAS and eQTLs, we performed a colocalization test (PMID: 24830394), from which we identified 190 genes across 79 loci in which GWAS and eQTLs share common causal variants. **[[end to suppl]]**

In total, we identified 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, providing a high-confidence gene list (Figure 5xx). This is a huge increase from the previously annotated 22 genes across 19 loci based on CMC adult 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 **xxx**), consistent with the previous observations that regulatory relationships often do not follow linear genome organization.

**[[rewrite – see dict. ]]** Moreover, we found that these SCZ genes can reveal additional biological mechanisms and functions that GWAS is unable to identify. They were enriched for genes and co-expression modules dysregulated in DLPFC of schizophrenia-affected individuals \cite{27668389}, suggesting that common variation-mediated gene regulation contributes to the gene dysregulation in schizophrenia (Figure 5xx). This also hints that there likely exists shared genetic etiology between common and structural variation since the SCZ genes are often affected by recurrent CNVs in schizophrenia. Also, the SCZ genes are enriched with the loss-of-function mutation intolerant genes \cite{27869829}, translational regulators, cholinergic receptors, calcium channels, and synaptic genes (Figure **xxx**). We further leveraged our single-cell expression data to examine cell-type specific expression signatures of SCZ genes, and found that they have significantly higher expression levels at neuronal cells than non-neuronal cells.**[[end-rewrite]]**

**[[rewrite para]]** We looked at the characteristics of the genes that were associated with the SCZ loci. First, as expected, these genes shared many of the characteristics of known SCZ genes. In particular, they're enriched in copy number variants, lost of function variants associated with functions attributed to schizophrenia. They were also enriched in differentially expressed genes associated with schizophrenia **[[Ref 00:44] Capstone 1]. We further integrated this gene list with the single-cell profiles and found, interestingly, they were enriched in a variety of different neural types.**

**IX. Integrative modeling to relate genotype to molecular and high-level phenotypes in the adult brain (880)**

**The interaction between genotype and phenotype involves multiple intermediate levels; in this section therefore, we perform another level of integrative analysis by embedding our gene regulatory network from the previous section into a larger model. For this purpose, we introduce an interpretable deep-learning framework, a Deep Structured Phenotype Network (DSPN, Figure 6)** \*\*\***[[name]]**. This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from the QTLs and gene regulatory connections predicted from our elastic net regression. As shown (Figure 6a), traditional classification methods such as logistic regression predict the phenotype directly from genotype, without inferring intermediates such as the transcriptome. We build the DSPN via a series of intermediate models which add layers of structure to a logistic regression model, including a layer for intermediate molecular phenotypes such as gene expression and chromatin state, multiple layers for functional modules and other mid-level phenotypes which may be inferred as hidden nodes in the network, and a layer for high-level phenotypes such as brain traits. Finally, we use special forms of connectivity (enforcing sparsity and adding lateral intra-level connections) to integrate our knowledge of QTLs, regulatory network structure, and co-expression modules from earlier sections of the paper (Supplement). By using a generative architecture, we ensure that the model is able to impute intermediate phenotypes when needed, as well as providing a predictive model for high-level traits and 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 high-level traits than without (Figure 6b) **[[discuss]]**. **[[should I use this or not??]]** In paritcualr, we achieve, prediction of **XXX** with the extra layers vs **XXX** without. Further **[[or in paritcular?]]**, comparison with a simple logistic predictor from the genome alone shows that the transcriptome carries significant further trait relevant information, which the DSPN is able to optimally extract (Figure 6a). For instance, in the case of Schizophrenia, a logistic predictor is able to gain a 2.8 times improvement when using the transcriptome versus the genome (+13% vs. +4.6% from 50% chance), while the DSPN is able to gain a 5 times improvement (+23% vs. 4.6%); this may reflect the need to incorporate non-linear interactions between intermediate phenotypes at multiple layers as in the DSPN. **[[enddiscuss]Moreover, the model also enables practical imputation of a subset of the transcriptome and epigenome (xx enhancers), with an accuracy of ~66-72% using Top 50 genes** \*\*\***[[TBD]][[maybe a diff stat]]** (Figure 6c, Supplement). We can thus perform joint inference of the imputed intermediate phenotypes (ie the transcriptome and epigenome) and high-level traits from the genotype alone using the DSPN, which achieves between 57.9-66.7% for disease trait prediction (Figure 6c), which is better than **XXX**. **[[[add in]]** These results demonstrate the usefulness of even a limited amount of functional genomics information for unraveling gene-disease relationships, and that the structure learnt from such data can be used to make more accurate predictions of high-level traits even when absent.

**[[we need to discuss]]**We transform the results above to the liability scale in order to compare with heritability estimated on this scale using GCTA (Figure 6d). Using the PsychENCODE cohort, we estimate that common SNPs and eSNPs explain x% and x% of liability for Schizophrenia respectively, which is comparable to previous estimates. **[[[fill in]]**] The imputation-based DSPN model explains a comparable level of variance to the eSNPs (4.5%), although we note that the DSPN may be capturing epistatic interactions not modeled in SNP-based heritability.**[[[explain]]**] The full DSPN model estimates that the transcriptome-based liability for the DLPFC is ~32.8%. Although we expect that a large portion of this will overlap with the common SNP based liability (which has previously been estimated as 25.6%) and genetically determined non-linear interactions, it may also include environmental and trait-influenced contributions (see Supplemental Figure), meaning that it is an upper-bound on the genetically determined liability modeled by the DSPN. Similar estimates of the liability explained for Bipolar and ASD by the DSPN (imputation and full models) are given (Figure **xxx**).

We examined the connections learnt by the DSPN between intermediate and high-level phenotypes for potentially relevant biological interactions. We included specific known co-expression modules and submodules as 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 common prioritization scheme (Supplement). For instance, in Schizophrenia, we found that the highest prioritized module in the DSPN was associated with Dopaminergic and Glutamatergic synapse and calcium signaling pathways, with other modules associated with Oligodendrocyte markers, and the Complement cascade pathways, which confirms and extends previous smaller scale analyses **[[refs]]**. Further, we found that excitatory neuronal markers were enriched in the highest prioritized module for age, while the gene NRGN occurred in many of the top prioritized modules/submodules, in agreement with the earlier analyses. We further used eQTLs and cQTLs to link SNPs to the genes/enhancers of each module, and show that the modules prioritized by the DSPN are strongly enriched for GWAS variants (Supplement). Examples showing specific associations between modules, genes and variants for schizophrenia are shown (Figure 6e), and we provide a full summary of the functional enrichment analysis for all disease and high-level traits in supplement (Supp section xx).

**IX. Discussion (532)**

We integrated PsychENCODE datasets with other resources, and developed a comprehensive resource consisting of various functional genomic elements for the adult brain including data from 1931 individuals. This resource serves as an important step for gaining biological insights from genomic functional data in neuroscience. Overall, our study has identified a very large-scale set of eQTLs and eGenes for adult brain, several folds more than previous studies, almost achieving saturation of protein coding genes. Therefore, 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, in addition to eQTLs. **[[shorten rest of para by 50%]]**The first would be chromatin QTLs. Increasing the sample size may potentially help identify more cQTLs, which also can be further interrelated to eQTLs . Moreover, the enhancers that this study used for cQTLs are defined from the current techniques such as ATAC-seq and ChIP-seq, especially from K27AC. In the future, methods such as STARR-seq may provide more accurate definitions on enhancers, and thus can be further used to better identify chromatin associated variants. **[[end-shorten]]**

Another area of future development is single cell analysis. Current techniques suffer from the low capture efficiency, and so it remains challenging to reliably quantify low-abundant transcripts/genes and interrogate biological variation \cite{26949524, 25053837}. In this study, we found that 25 basic and known cells could explain large expression variations across tissues. However, there still exist the gene expression heterogeneities even among the same cell types (Figure Sxx\_magic), implying potential additional cell types in the adult brain. Thus, increasing single cell data and more advanced techniques in the future are expected to identify a considerable number of novel cell types, which might contribute to the unexplained variation. **[[expl next 2 sent]]**Also, given the issue of RNA decay in single cell RNA-seq, we intend to relate this resource to recent in situ transcriptomic data such as the spatial gene expression by optogenetic techniques, allowing us to find consistent expressed genes driving the brain phenotypes at the cellular and tissue levels. In addition, recent single-nucleus RNA sequencing techniques have emerged to complement the gene expression measurement at the whole cell level that can be impacted by various aspects such as intercellular interactions \cite{28729663, 29227469}. [[single doesn't work well with axons and dendrites]]

**[[cut para or do better 2 sentences]]**More accurate cQTLs and fQTLs can be input into our deep learning model, which is expected to improve the model performance. Also, the integrative model is readily expandable to include additional data types such as imaging and medical data, allowing a broader range of intermediate phenotypes to explain the connection between genotype and high-level traits. Furthermore, while providing better prediction, some model connections are deliberately set based on prior knowledge from the other analyses, such as the gene regulatory networks linkages, to make the model more interpretable and easier to use. Thus, another major goal of the model is to provide a useful compression of the resource as a whole; e.g., **XXX** KB for the model representation vs. **XXX** TB for the original functional genomic brain datasets. **[[[[jw/dc better]]**]

#

# **\*\*\*\*\*\* 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)

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

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

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

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

[Understanding the molecular mechanisms that genomic variants change associated phenotypes in brain disorders is still a key challenge. The PsychENCODE consortium has generated and assembled a robust large-scale dataset on the adult human brain to address this challenge, including genotyping, RNA-seq, ChIP-seq and single-cell transcriptomic data on many individuals with different phenotypes including normal, mental diseases. We integrate this dataset with complementary genomic information from other large consortia, particular from ENCODE, GTEx and Epigenomics Roadmap to develop a comprehensive resource for the brain functional genomics (~2000 samples in total) and compare it against various phenotypes. This resource comprises the regulatory variants such as QTLs, brain active enhancers, differentially expressed genes and transcripts, novel transcribed regions and non-coding RNAs, and putative genome-wide regulatory networks. We make the derived resource downloadable and available on the PyschENCODE website (xxxx).](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)

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

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

[We then use this resource and compare against various phenotypes to reveal the unique brain genomic activities, particularly relating to transcriptomic and regulatory binding activities. In particular, we performed the spectral analysis for comparing the similarities of gene expression and epigenetic data with other tissue samples from GTEx and Epigenomics Roadmap (Figure A). Our analysis revealed that the samples can be clustered together from their tissue types using either protein-coding gene and non-coding RNA expression or epigenetic data; e.g., brain and other tissue clusters. It shows that brain samples separated more significantly from other tissues in terms of gene expression. Additionally, we found that the brain has more non-coding transcriptional activity than most other tissues. These results suggest that the brain has specific and distinct expression and epigenetic profiles. These brains related clusters and activities that the brain has specific and distant expression and epigenetic profiles and transcription, involved by the brain elements in the resource. Furthermore, we identified regulatory variants significantly associated with brain transcriptional and epigenomic activity: >5 million expression QTL for gene expression and >5 thousand chromatin QTL for histone modification signals (Figure B). These variants cover a larger fraction of disease-associated brain GWAS SNPs than any previous analyses, suggesting potential molecular targets for these associations.](https://docs.google.com/document/d/1vJ1PIW1AVwkMSpR036AOJbrNCnlFrd5RImXSX3rRNTk/edit?usp=sharing)