Manuscript link: goo.gl/f228aK

PEC Capstone 4 Supplement

# Introduction on PsychENCODE data

PE Data tables: <https://goo.gl/bDEo8s>

Large datasets produced by the psychENCODE consortium include over 2,000 human brain samples for healthy controls and individuals afflicted by neuropsychiatric diseases. These datasets include full genotyping, RNA-seq, ChIP-seq, as well as single-cell data. It also includes processed data such as expression QTLs and chromatin QTLs trait loci, enhancers that are active in different brain regions, in addition to differentially expressed genes, transcripts, and novel non-coding RNAs. These datasets are also provided at the resolution of sub-regions of the brain, and they provide valuable resources for investigating genomic signatures and potential underlying factors for an array of psychiatric diseases.

In addition to providing new opportunities for studying the genetic basis of psychiatric disorders, however, the very richness of this data introduces considerable challenges with respect to data integration and organization. Our analyses rely on multiple methodologies, the details of which are difficult to include within the main text of this paper. As such, the purpose of this Supplement is to provide a clear and organized reference to support and explain the datasets, pipelines, and analyses associated with this study. In addition to supplementary text, the supplementary figures and tables provide additional information not included among the main figures.

The data resources may be organized into a pyramid-like structure, with large raw data files at the base, and more processed summary data organized at higher levels. The raw data files include datasets from PyschENCODE, ENCODE, CommonMind, GTEx, Epigenomics Roadmap, and others. These comprise RNA-seq expression quantification data, ChIP-seq signal track qualifications and peak identifications using ENCODE standard pipelines, in addition to private data such as imputed genotypes. Further up the pyramid, more readily human-interpretable data and descriptors populate the top. These more processed datasets include patient metadata and phenotypes (such as disease status), fully processed epigenomic signals and peaks, active enhancers, QTLs, differentially expressed genes and transcripts, and regulatory networks.

As reflected in the main text, our study is broadly organized into several main parts: a description of the the datasets that constitute the data pyramid (raw and processed), identification of brain-specific genomic and transcriptomics signatures through comparative analysis, single cell analysis and deconvolution to elucidate gene expression changes across multiple phenotypes, and integrative modeling to investigate molecular mechanisms for genotype-phenotype associations. This supplement is presented in roughly a parallel fashion to the main text. It is also connected to the main text through the major results presented in the form of main text figures – captions associated with main text figures point to relevant sub-sections within this supplement. In cases where the related supplementary section is not readily apparent, we note "see supp. section xyz” to refer to a specific section.

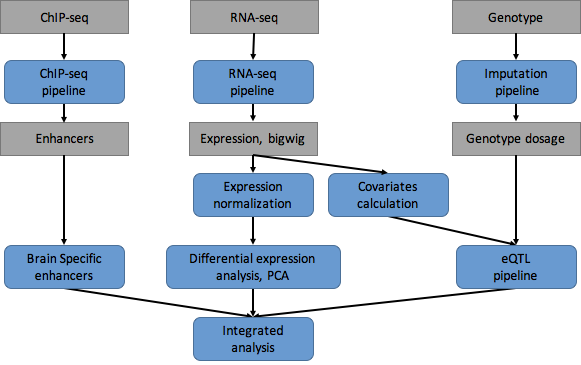
With the aim of presenting data and results (including software packages) in an organized way, we have written about this study in roughly a hierarchical fashion. The main text lies at the top of this hierarchy and synthesizes everything in a broad manner. It refers to more detailed descriptions of our methods and datasets, as provided in this supplement. Raw data files, which lie at the bottom of the hierarchy (and which are hosted as online resources) form the bedrock from which our results are built.

We note that, in preparing this supplement, we adopt the conventions prescribed in the recent opinion piece by Greenbaum et al\cite{28381262}. As such, labels that correspond to sections, sub-sections, figures, tables, and data files are labeled to indicate whether these items directly parallel (||) or do not parallel ([[dc2dc: need to insert symbol after importing into MS Word -- use tag “[[NPAR]]” for now]]) the main text, as well as whether these items are high-level or technical in nature (designated by “HL” and “TL”, respectively).

# 1 S (TL, ||) Datasets

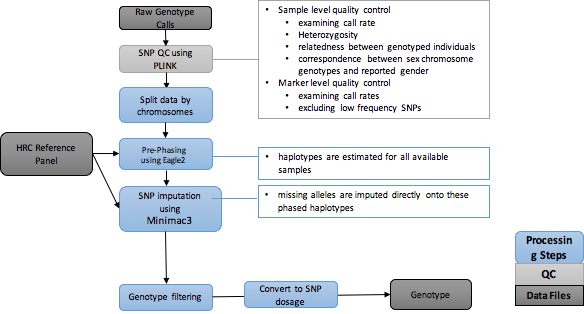
## 1.1. S (TL, ||) Phenotype Data (JP, DW)[[shadow figure & caption for fig 1 - PE]]

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



# 2 S (TL, ||) Transcriptome data processing and analysis

## 2.1 S (TL, [[NPAR]]) Genotype



### 2.1.1 S (TL, [[NPAR]]) Genotyping arrays, data generation and quality control

**[[Sl to do]]**

Genotyping was done on several different genotype chips listed in supplemental table. Initial QC was performed using PLINK (6) to remove markers with: zero alternate alleles, genotyping call rate < 0.95, Hardy-Weinberg p-value < 1 x 10-6, and individuals with genotyping call rate < 0.95.

|  |  |  |
| --- | --- | --- |
| **Datasets** | **#samples** | **DataPlatform** |
| BipSeq | 179 | Illumina\_1M and Illumina\_h650 |
| LIBD\_szControl | 493 | Illumina\_1M, Illumina\_Omni5, Illumina\_h650 |
| CMC-HBCC | 696 (896 totoal) | Illumina\_1M, Illumina\_Omni5, Illumina\_h650 |
| BrainSpan | 41 | HumanOmni2.5 |
| CommonMind | 620 | IlluminaInfiniumHuman Omni Express Exome 8 v 1.1b chip |
| GTEx | 450 (97 DFC) | Illumina OMNI 5M or 2.5M |
| BrainGVEX | 138+280 | Affymetrix6.0, PsychChips |
| UCLA-ASD | 97 | Omni-2.5 and Omni-2.5-Exome |
| iPSC | 3 | WGS |
| EpiGABA | 9 | Illumina\_HumanOmni1-Quadv1.0 |

### 2.1.2 S (TL, [[NPAR]]) Imputation of genotypes

Genotypes will be imputed using the Michigan Imputation Server, in order to streamline quality control, genotype imputation, and statistical analysis of genome-wide single nucleotide polymorphism (SNP) data. This imputation pipeline consists of four primary, independent modules: (1) pre-imputation data processing and quality control; (2) principal components analysis (PCA); (3) genotype imputation of untyped variants; and (4) post-imputation statistical analysis. Briefly, in the pre-imputation step, input genotype data (PLINK binary format) is reformatted for downstream analysis, and initial summaries of classic technical parameters (e.g. minor allele frequency, per-individual and per-site missing rates, case/control missingness, Hardy-Weinberg equilibrium) are produced. The second module consists of data filtering and relatedness testing, followed by PCA using EIGENSTRAT(36) to identify ancestry outliers and any detectable population substructure. Prior to imputation, SNP positions, identifiers, and alleles are aligned to the relevant reference genome assembly (using LiftOver), and genotype data is divided into overlapping 5 megabase (Mb) segments (~1000) for subsequent, parallel haplotype pre-phasing and imputation using SHAPEIT2/IMPUTE2 (37, 38). We used the recently released HRC Reference Panel for imputation of rare SNPs. We use the summary of R2 values from Minimac3 to evaluate the imputation accuracy .

**[[SL to fill]]**

## 2.2 (TL, ||) Bulk RNA-seq

### 2.2.1 (TL, ||) PsychENCODE and other brain consortia (DC, SL) [[polish by DC & SL]] [[we need to insert & harmonize the mike gandal text]]

We processed gene expression read count data (as quantified by FPKM and measured by RNAseq) from 9 studies: UCLA-ASD, Yale-ASD, BrainGVEX, the The Lieber Institute for Brain Development (LIBD), GTEx, the CommonMind Consortium (CMC), the CMC’s NIMH Human Brain Collection Core (CMC HBCC) and Bipseq a Bipolar cohort and from Yale. The detailed descriptions of these studies could be found on supplemental table S1 and PsychENCODE Knowledge Portal(https://www.synapse.org/#!Synapse:syn4921369/wiki/390659 ). An initial quality control step was taken in which all datasets were first pre-processed to remove outliers using a hierarchical clustering based global outlier detection. Samples from UCLA were sub-divided into three different brain regions (vermis, Brodmann area 9, and Brodmann area 41).

The gene expression data from these 9 centers were merged into one gene expression matrix, and subsequently normalized using the protocol detailed by GTEx \cite{GTEx7}.

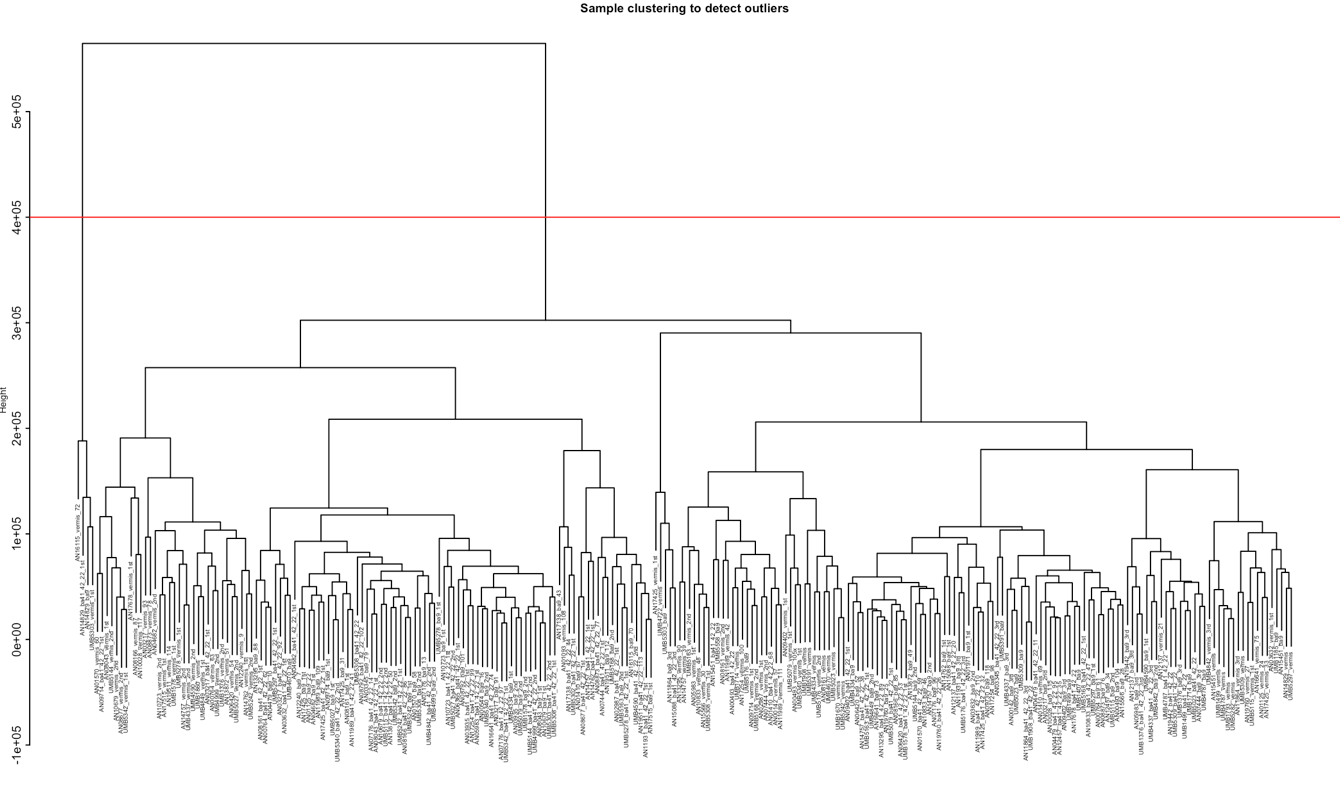


Table S1 (HL, ||) Dataset [[to be updated PE]]

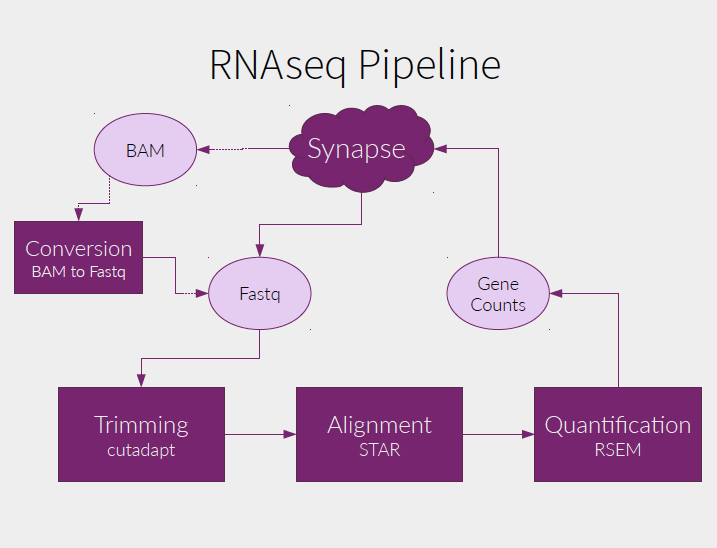
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Datasets** | **Disease status** | **# samples** | **Brain region** | **Data Type** |
| ENCODE/Roadmap | CTL |  | Multiple regions | ChIP-seq, RNA-seq |
| BrainSpan | CTL | 606 | 16 regions | ChIP-seq, RNA-seq, genotype |
| CommonMind | SCZ, BD, CTL | 613 | DFC | RNA-seq, genotype |
| GTEx | CTL | 92 | DFC | RNA-seq, genotype |
| BrainGVEX | SCZ, BD,C TL | 429 | DFC | RNA-seq, genotype |
| UCLA-ASD | ASD, CTL | 253 | DFC, CBC | ChIP-seq, RNA-seq, genotype |
| Yale-ASD | ASD, CTL | 45 | DFC, TC, V1, CBL | ChIP-seq, RNA-seq, genotype |
| CNON | SCZ, CTL | 63 | Olfactory Neuroepithelium | ChIP-seq |
| EpiGABA | CTL | 4 | OFC (BA11) | RNA-seq, genotype, ERRBS, Methylation |
| iPSC | CTL | 51 | iPSC | ChIP-seq, RNA-seq, WGS |
| TOTAL |  | **~2156** |  |  |

### 2.2.2 S (TL, ||) GTEx brain and other tissues (FN, JW) [[6p or 7?]]

### We use several types of data from the GTEx version 7 dataset \cite{GTEx7}. GTEx version 7 contains RNAseq and matching genotype data for 10 brain regions (Anterior cingulate cortex, Caudate nucleus, Cerebellar hemisphere, Cerebellum, Cortex, Frontal cortex, Hippocampus, Hypothalamus, Nucleus accumbens, and Putamen). We use the raw RNA sequencing data to quantify the proportion of the transcribed non-coding genome. For the eQTL calculations and WGCNA analysis, we used the individual TPM data, and renormalized it using the PEER factors calculated in combination with the PsychENCODE data. Further, for the eQTL calculations, we re-imputed the genotype data using from the raw genotype calls using the pipeline described below to match the processing of the PsychENCODE data.

We used further data from GTEx 7 \cite{GTEx7} to compare the transcriptome of the brain to that of other tissues. GTEx 7 contains RNAseq data from 34 other tissues. As above, we use the raw RNA sequencing data to quantify the proportion of the transcribed non-coding genome. For WGCNA analysis, we used the individual TPM data, pre-normalized by the PEER factors calculated in GTEx 7 to identify modules in individual tissues, and the median TPM data by tissue to identify modules across tissues.

### 2.2.3 S (TL, ||) RNA-seq processing (FN, SL, DC)[[FN to write]] [[SL+DC: add QC details in this sect.]]



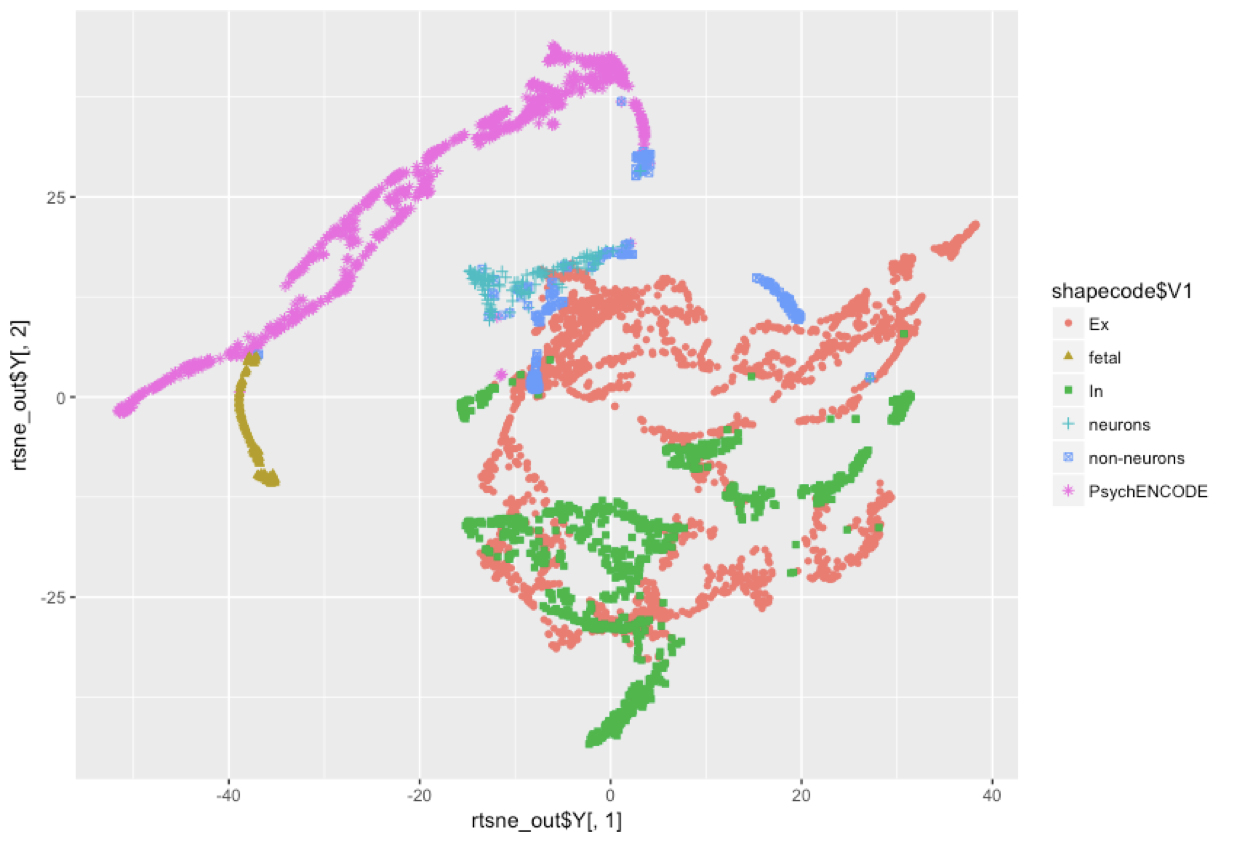
The PsychENCODE RNA-seq pipeline is mostly based on the ENCODE RNA-seq pipeline. The PsychENCODE RNA-seq pipeline is compatible with stranded and unstranded mRNAs from (poly-A(+)), rRNA-depleted total RNA, or poly-A(-) RNA libraries. The pipeline takes as inputs both RNA-seq reads (from paired-end stranded or single end unstranded libraries), a reference genome and a gene annotation file (by default GENCODE). For this release, we used GRCh37 (hg19) as reference genome and Gencode v19 as gene annotation. Both coding and non-coding transcript were used. For each sample the pipeline outputs: A bam file with reads mapped to the genome, a bam file with reads mapped to the transcriptome, bigwig files with normalized RNA-seq signal track for unique and multimapping reads (splitted between +strand and -strand if the library is stranted), gene quantifications and transcript quantifications.

== The mapping of the reads is done using STAR (2.4.2a) and the quantification of genes and transcripts is done with RSEM (1.2.29). Although there is general agreement between the mappings and the gene quantifications produced by different RNA-seq pipelines, quantifications of individual transcript isoforms, being much more complex, can differ substantially depending on the processing pipeline employed, and are of unknown accuracy. Therefore, mapping and gene quantifications can be used confidently, while transcript quantifications should be used with care.

## 2.3 S (TL, ||) Single cell RNA-seq (FN, XS, DW)[[XS]]

### 2.3.1 S (TL, ||) [[XS to merge]] Datasets of single cell transcriptomics.

We integrated and used the same pipeline including ENCODE RNA-seq analysis to uniformly process the single cell RNA-seq data for ~3000 neuronal cells with 8 excitatory and 8 inhibitory types [Lake’s 2016], and ~400 cells including two fetal types, one adult neuronal type and 5 adult non-neuronal types, astrocytes, endothelial, microglia, oligodendrocytes and OPC [Quaker 2015], and ~900 cells from PsychENCOCE with xxx novel cell types in embryonic and fetal tissues [Brainspan]. We first apply quality control on ~900 cells using R ‘scater’ package [scater paper] to filter the cells with low library size and high mitochondrial RNA concentration. Furthermore, the cells with total library size less than 0.2 million were also filtered for future analysis. In total, we built a gene expression profile of ~800 high quality cells quantified in TPM. We merged the PsychENCODE data, Lake et al data and Quake et al. by matching the gene names. As the single cell data suffers from high dropout rates, we used MAGIC to impute the missing values in the expression matrix. We compared these single cells based on the (biomarker) gene expression similarity using tSNE, and found that the same-type cells generally can be clustered together (Figure S xxx). In particular, xx% PsychENCODE cells have been found to cluster together with known cell types (xx% neuronal, xx% non-neuronal, details in supplement). In addition, xx% PsychENCODE cells form their own clusters, away from known cell types, suggesting that the potential novel cell types found by PyschENCODE for brain tissues. We also include these single cell data and cell-type biomarker genes in the resource.



Specifically, we integrated and used the same pipeline to uniformly process the single cell RNA-seq data for ~3000 neuronal cells with 8 excitatory and 8 inhibitory types [Lake’s 2016 paper], and ~400 cells including 5 non-neuronal types, astrocytes, endothelial, microglia, oligodendrocytes and Oligodendrocyte progenitor cell (OPC), and ~800 cells from PsychENCOCE for potentially additional cell types in embryonic and fetal brain tissues [ref brainspan]. In total, we included 23 single cell types (Supplement). We first compared these single cells based on the (biomarker) gene expression similarity using tSNE, and found that the same-type cells generally can be clustered together (Figure Sxxx). This suggests that our integration has removed the batch effects of single cell data from different studies. In particular, xx% PsychENCODE cells have been found to cluster together with known cell types (xx% neuronal, xx% non-neuronal, details in supplement). In addition, xx% PsychENCODE cells form their own clusters, away from known cell types, suggesting that the potential novel cell types found by PyschENCODE for brain tissues.

We also found that the gene expression changes across individual tissue samples can be largely explained by the single cell gene expression, and the changes of single cell fractions are also associated with the individual phenotypes. Therefore, we deconvolved the tissue-level gene expression data of all 1931 individuals’ tissue samples using single-cell gene expression data of 450 biomarker genes to find the fraction of different cell types corresponding, and compare cell fractions across different phenotypes.

**Quantification of gene expression: [[XS to add]]** The gene expression in both bulk and single-cell RNA-seq data were quantified in TPM and further transformed into log scale by log2(TPM+1). The transformed gene expression will be used in the decomposition and devolution analysis.

## **2.**4 **S (TL, ||) Decomposition of brain tissue gene expression data**

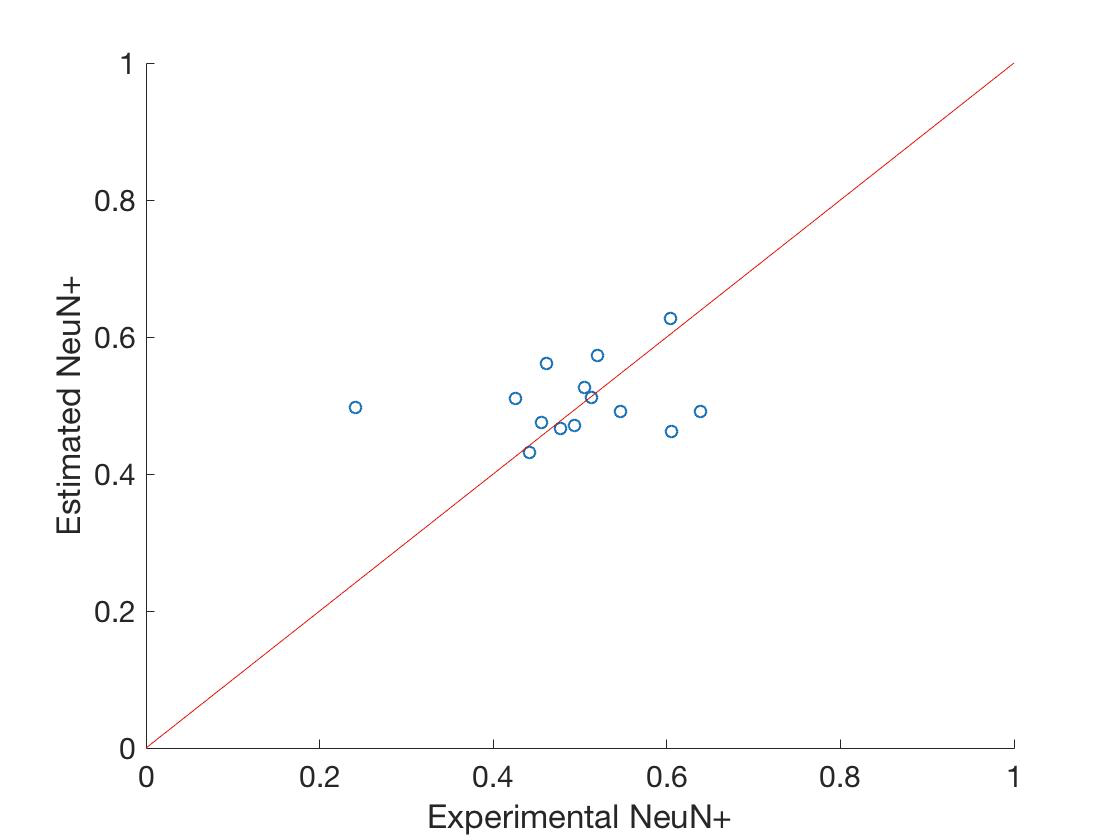
To check if the brain tissue expression is attributed to the combinations of single cell types in Section 2.4; i.e., the cell fractions, we decomposed the brain tissue gene expression data using unsupervised approach to find the principal components of tissue data, and compared them with single cell expression data. Specifically, given the brain tissue gene expression matrix X (M by N) for a phenotype/disorder where M is the number of tissue samples and N is the number of select genes (e.g., the cell biomarker genes), we used non-negative matrix factorization (NMF) to decompose X into the product of two matrices, H and V so that ||X-H\*V||^2 is minimized with constraints that all elements of H. H is an M by K matrix with the (i,j) element describing the contribution coefficient of jth NMF principal component (NMF-PC) to ith tissue sample, K is the number of select NMF-PCs (e.g., equal to the number of select cell types as above), and V is a K by N matrix with the (i,j) element being the expression level of jth select gene on ith NMF-PC. We then correlated NMF-PCs with the select gene expression data of different single cell types, and obtain the correlation map between NMF-PCs and single cells (Figure XX). For example, No. 22 and 23 NMF-PCs of the non-neuronal group highly correlate with astrocytes, No. 2 NMF-PC correlate with fetal cells, and No. 1, 5, 10, 24 and 25 NMF-PCs of the neuronal group correlate with excitatory neuronal cell types. This suggests that the large portion of tissue’s gene expression changes is a linear combination of these cell types’ gene expression. Thus, we want to further identify the cell fractions showing how individual single cells contribute the tissue’s gene expression, using the deconvolution.

## **2.**5 **S (TL, ||) Deconvoluting brain tissue gene expression data using single cell data to estimate cell fractions**

Because we used unsupervised approach (NMF) to decompose the tissue expression and found NMF-PCs recovered the expression patterns (spaces) of both neuronal and non-neuronal cells. This suggests that the brain tissue expression is highly likely contributed by a linear combination of single cells. Therefore, to more accurately identify the single cell fractions that determine the tissue expression, especially for various phenotypes/disorders, we further applied the supervised approach that uses the single cell expression data to deconvolve the brain tissue expression data to find the fractions of different cell types of individual tissues. In particular, the brain tissue gene expression matrix X (M by N) for a phenotype/disorder, where M is the number of tissue samples and N is the number of select genes (e.g., the cell biomarker genes), and the single cell gene expression matrix Y(K by N), where K is the number of select cell types, we used the non-negative least square method to find a non-negative M by K matrix, W to minimize ||X-W\*Y||^2. The (i,j) element of W represents the linear combination coefficient of jth single cell type to ith tissue expression, which is proportional to the jth single cell fraction. In the deconvolution analysis, the gene expression quantified in TPM was transformed into log scale by log2(TPM+1). We further evaluate the goodness-of-fit for the deconvolution model by calculating the coefficient of determination (also known as R^2), which accounts for the percentage of variance in the individual gene expression of tissue samples that has been explained by varying the cell proportions of cell types. Specifically, the variance in the gene expression of tissue samples is ||X|| and the variance that has not been explained by the model is ||X-W\*Y||. The R^2 can be calculated as 1-||X-W\*Y||/||X||, which was further normalized to adjusted R^2 by incorporating the degree of freedom. In addition, we deconvolved the tissue expression data and compared the cell fraction changes for various phenotypes and psychiatric disorders.

**N+/N-1 validation figure**

We have validated our estimated cell fractions on a subset of samples from EPIMAP study with experimental measured NeuN+ fractions. Figure xxx shows the NeuN+ fractions measured in experiment and estimated in our deconvolution analysis on 14 samples with RIN > 7.3. It can be seen that our estimation is very close to the experimental NeuN+ fractions.



## 2.6 S (HL, ||) Differentially expressed genes for brain phenotypes (JP, DW)

We used the R package, limma for linear modeling to find genes that are differentially expressed for neuropsychiatric disorders, sex, and brain regions. [cap1]

## 2.7 S (TL, ||) Gene co-expression network analysis (JW, DW) [[JW to add submodules…]]

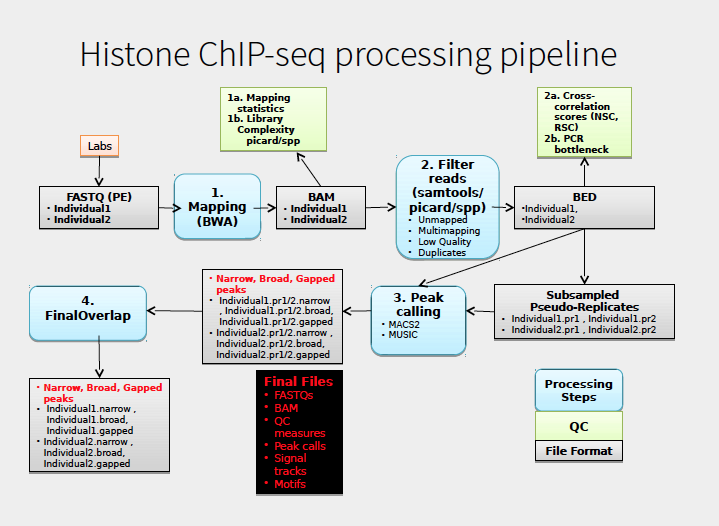
We use Weighted Gene Co-expression Network Analysis (WGCNA) to identify modules of co-expressed genes, both within tissues and between tissues \cite{Zhang05}. Briefly, each gene is associated with a vector of normalized expression values across either individuals or tissues (using median expression). A weighted network is constructed where the weight between any two genes is a similarly score, calculated by normalizing the Pearson correlation of their expression vectors to lie between 0 and 1, and raising this to the power . We follow Zhang et. al. in setting such that connectivity of the network is as close to scale-free as possible (using the statistic described in Zhang et. al.). The genes are then hierarchically clustered using a *topological overlap score*, which compares how similar the patterns of connection are from each node to all other nodes, and disjoint modules are extracted using the Dynamic Tree Cut algorithm \cite{Langfelder08}. To find brain specific modules using clusters calculated on median expression variation across tissues, we further calculate the *module eigengenes* (as described in Zhang et. al.), and calculate the correlation of each eigengene with a binary vector which is 1 for brain regions and 0 otherwise. We call a module ‘brain specific’ if this correlation is significant at the 0.05 level (under a permutation test of the tissue labels).

Our co-expression analysis indeed found several modules whose eigengenes show very different expression levels between brain and non-brain samples (Figure Sxxx, Supplement), which suggests that there exist brain specific regulatory mechanisms drive these brain co-expression modules.

# 3 S (TL, ||) Epigenome data processing and analysis

## 3.1 S (HL, ||) PsychENCODE ChIP-seq (SL)

We use the modified parallel version of ENCODE ChIP-seq pipeline. The pipleline is shown in supplemental figure x.



## 3.2 S (HL, ||) Epigenomics Roadmap, ENCODE ChIP-seq (PE, MTG)

We incorporated ChIP-seq datasets from the Roadmap Epigenomics Consortium and the ENCODE project in our analysis. To integrate them consistently with the PsychENCODE dataset, all these ChIP-seq experiments are uniformly processed using ENCODE standard pipeline (See below section 2.3 S), including alignment, quality control and peak-calling. Each of the released experiment consists of the raw sequencing data (in FASTQ format), and processed output, including alignment files, signal files and peak files.

Based on the set of uniformly processed ChIP-seq experiments, a comprehensive statistical model was used to generate a registry of candidate regulatory elements (cREs) for major cell lines and tissues (Moore et al, under review). The cREs are based on a combined set of high quality DHSs. For a particular cell or tissue, z-scores for DNase, H3K4me3, H3K27ac and CTCF is calculated for these high-quality DHSs. Based on the maximum z-score across all cell types and the distance to the nearest TSS, they are classified into promoter-like elements, enhancer-like elements and regions bound by CTCF only. As described in later sections (2.6 S and 2.10 S), we based on these cREs to calculate cQTLs and annotate sets of enhancers in different brain regions.

## 3.3 S (HL, ||) Activated brain enhancers (MTG)

We took the uniformly processed H3K27ac ChIP-seq data from healthy individuals and integrated the peaks to generate a set of candidate enhancers. The H3K27ac peaks are first merged across samples. The average size of the merged peaks is around 1.5Kb. We then filtered out peaks that occur only in 3 individuals or less since they are likely sample specific. This filters out 53% of the merged peaks. Upon intersecting the remaining regions with candidate regulatory elements (cREs) from ENCODE annotation (Moore et al, in review), we found that more than 91% of these filtered regions overlap with at least one cRE. The cRE that is closest to the center of each filtered H3K27ac peak is chosen to represent the regulatory region. In addition, we identified 394 merged peaks that, though not overlapping with any cREs, are present in more than half of the control samples. We included these regions as well in the brain regulatory enhancer set. Finally, we excluded any regions in the set that are within 2Kb upstream of any TSS. In the final set we annotated around 88K regions that are predicted to be active enhancers in human prefrontal cortex, 50K active enhancers in the cerebellum and 90K in the temporal cortex. We checked out annotation against an independent set of ATAC-seq peaks from . Compared to the Roadmap prefrontal cortex enhancer annotation, our have significantly more overlap with

## 3.4 S (HL, [[NPAR]]) Comparison of brain enhancers using REMC datasets (from Suhn)

To compare brain enhancers with enhancers from more than a hundred different cell types, we obtained 2.3M enhancer elements reported from REMC that were identified by overlapping DNase peaks with ChromHMM enhancer chromatin states (EnhG, Enh, Enhbivalent) using real and imputed histone ChIP-seq data in 111 different cell/tissue samples (Kundaje et al REMC, Nature 2015) (<http://egg2.wustl.edu/roadmap/data/byDataType/dnase/BED_files_enh/>). Of these 2.3M enhancers, 539,220 were identified in at least one of the tested brain tissue samples (Angular Gyrus, Anterior Caudate, Cingulate Gyrus, Germinal Matrix, Hippocampus Middle, Inferior Temporal Lobe, Mid Frontal Lobe/Dorsolateral Prefrontal Cortex, or Substantia Nigra). Brain is known to produce more divergent transcript variants than other tissue samples (Mele et al GTEx, Science 2015). Therefore, we determined the number of REMC enhancers that were present in brain tissue samples and the number of REMC enhancers that were brain-specific in each brain tissue sample, and then compared these to the numbers of enhancers for the rest of tissues analyzed by REMC. On average, there were no statistically significant differences in the number of enhancers present in brain as well as tissue-specific enhancers for brain than for other tissues (t-test, adj. p-value > 0.05). We also compared brain enhancers identified by REMC with the 88,000 brain enhancers which were identified as part of the PsychENCODE Consortium, identifying more than 37,000 additional enhancers that are active in the brain than were found in the REMC collection.

# 4 S (TL, ||) Comparative analysis

## 4.1 S ([[?? -- TL or HL? Cannot tell from content]], ||) Spectral analysis like PCA, RCA (FN, DW)

Gene expression, chromatin

DW to add PCA, tSNE

## 4.2 S (HL, ||) Non-coding RNAs and TARs (FN, 6p or 7?)

We used uniformly processed RNA-seq signal data from healthy individuals deriving from GTEx and from PsychENCODE to quantify the expression activity of annotated and non-annotated regions of the human genome. In order to create signal files we used alignment files (bam files) as input to RSEM to create both uniquely aligned and multiple aligned signal tracks. Singal values were normalized within samples using the total number of reads mapped to the genome and generating RPM values. We took bins of 100 base pairs of the genome and calculated the average expression (RPM) in windows. We finally selected regions in the genome with RPM higher than 0.1 to filter transcriptionally active regions. The union of all bins in the human genome above the threshold was used to build the resource of active regions of the human brain. Furthermore we calculated the number of novel transcribed bin as individuals were analysed. We finally fit the curve of novel transcribed bins to a cumulative exponential curve to calculate the upper bound of windows in different tissues.

# 5 S (HL, ||) QTL analysis

## 5.1 S (HL, ||) eQTL (SL, DC)

We used QTLtools package for eQTL analysis. The gene expression matrix will be normalized according to gender, Age, and genotype PCs for eQTL analysis. Probabilistic Estimation of Expression Residuals (PEER) factors, genotype PCs and gender will be used as covariates input for Matrix eQTL/fastQTL. Based on our sample size, we will calculate both cis-eQTL and trans-eQTL. Finally we will correct for the multiple hypothesis tests of SNPs in LD for a given gene for eQTL analysis.

## 5.2 S (HL, ||) cQTL (SL,MTG)

To calculate cQTL, we used the uniformly processed ChIP-seq data from PsychENCODE (3 different brain regions) and Roadmap ChIP-seq data for different tissues. The cQTL signal were calculated based on candidate regulatory regions (cREs). We first extend (in rare cases truncate) each cRE to 1kb, which is the typical size of enhancers. We then calculate the average signal on each of the extended regions across PsychENCODE and roadmap samples.

## 5.3 S (HL, ||) splicing QTL (MX, SL)

RSEM iso\_pct value “ 'IsoPct' stands for isoform percentage. It is the percentage of this

transcript's abandunce over its parent gene's abandunce. If its parent

gene has only one isoform or the gene information is not provided,

this field will be set to 100.”

## 5.4 S (HL, ||) Cell fraction & residual QTL (XS, DC, SL) -- also detail why we selected this specific set of 9 cell types

We used the QTLtools package\cite{28516912} to calculate the cell fraction and residual QTLs based on the cell fractions and residuals estimated in Section 3.4. QTLtools was run in nominal pass mode to identify fQTLs in 9 different cell types (Ex1, Ex3, Ex4, Ex5, In6, Astrocytes, Endothelial, Microglia, Oligodendrocytes). All nominal p-values were adjusted by Bonferroni correction, yielding 508 significant fQTLs (at a significance threshold of 0.05 for each Bonferroni-corrected p-value). A separate supplementary data file listing all fQTLs (along with associated data) is made available online.

# 6 S (TL, ||) Gene regulatory network

## 6.1 S (TL, [[NPAR]]) Hi-C and TADs (Heijin, Dan Geschwind)

**Hi-C.** Hi-C libraries were generated as previously described3. Briefly, adult dorsolateral prefrontal cortices (DLPFC) from three individuals (sample information provided below) were acquired through a Reference Brain Project as a component of the psychENCODE project. Briefly, frozen pulverized tissue (100mg) was homogenized in 2mL of ice-cold lysis buffer (10mM Tris-HCl pH8.0, 10mM NaCl, 0.2% NP40, protease inhibitor). Ten million nuclei were collected and chromatin was crosslinked in 1% formaldehyde (diluted in 1X PBS) for 10 min. Crosslinked chromatin was first digested by HindIII (NEB, R0104), and digested sites were labelled by biotin-14-dCTP (ThermoFisher, 19518-018). Proximity-based ligation was performed within nuclei to prevent random collision-based ligation4. Biotin-marked DNA was then purified and sequenced by Illumina 50 bp paired-end sequencing.

**Hi-C data processing.** Hi-C reads were mapped and filtered as previously described using hiclib (<https://bitbucket.org/mirnylab/hiclib)>3,5. Only cis reads (which refer to intra-chromosomal interactions) were used to construct contact matrices at 40kb and 10kb resolution for compartment and loop analyses, respectively. To obtain maximum resolution for loop detection (10kb), we pooled datasets from three individuals. To compare interaction profiles in adult and fetal brain, we combined previously generated Hi-C datasets from two fetal cortical laminae2 to obtain comparable read depths (see below for read depths for pooled samples). Compartments were analyzed by calculating the leading principal component (PC1) values from Pearson’s correlation matrix generated from contact matrices in 40kb resolution2. Regions with PC1s positively and negatively correlated with the gene density were defined as compartment A and B, respectively.

|  |  |  |  |
| --- | --- | --- | --- |
| **Samples** | **Sample information** | **cis filtered reads** | **total filtered reads** |
| HSB189 | Male 36yr  (Ancestry unknown) | 197,394,146 | 251,515,059 |
| HBS106 | Male 64yr  (Ancestry unknown) | 170,057,582 | 209,571,512 |
| HBS181 | Male 44yr  (Caucasian) | 243,396,052 | 299,801,452 |
| Pooled samples – Adult brain |  | 610,847,780 | 760,888,023 |
| Pooled samples – Fetal brain | Won et al.3 | 855,987,816 | 1,834,759,860 |

**Detection of promoter-enhancer interactions.** Promoter-enhancer interactions were identified as previously described3. Briefly, we constructed background interaction profiles from randomly selected length- and GC content-matched regions to promoters (defined as 2kb upstream of transcription start sites based on Gencode v19). Using these background interaction profiles, we fit interaction frequencies into Weibull distribution at each distance for each chromosome using the *fitdistrplus* package in R. Significance of interaction from each promoter was calculated as the probability of observing higher interaction frequencies under the fitted Weibull distribution, and interactions with FDR<0.01 (which corresponds to P-values~1x10–4) were selected as significant promoter-enhancer interactions.

## 6.2 S (TL, ||) Imputed gene regulatory networks (TFs) (PE, DW)

We integrated and imputated all possible regulatory relationships in brain including the enhancers, transcription factors (TFs), miRNAs and target genes in this resource (Methods). For example, we found the TF binding motifs using ENCODE data and inferred the TF-target gene relationships if TFs have enriched binding motifs on the target gene’s regulatory regions such as promoters and enhancers. The enhancer-target gene pairs were identified by JEME. In total, we included xxx enhancer-gene, xxx TF-gene, and xxx miRNA-gene regulatory linkages, providing a reference wiring network on gene regulation in brain, which consists of the regulatory factors and elements (e.g., TFs, miRNAs, enhancers) and target genes. In order to identify the activated regulatory wires for particular phenotype or disorder, we further used the computational method to determine such activated regulation. In particular, given a gene and a phenotype/disorder, we applied the Elastic net regression, linearly combining the L1 and L2 regularizations to predict its gene expression data from the expression data of the TFs that have the binding sites on the gene’s JEME enhancers and promoter and overlap the QTLs; i.e., the QTLs break the binding sites. We then identified the activated TF-target regulatory relationships if TFs have large regression coefficients. In detail, suppose *Y* is an *N*-dimensional vector with elements being the gene’s expression levels across samples, where *N* is the sample number for the phenotype/disorder. *X* is an *N* by *M* matrix whose columns are the TFs’ expression levels, where *M* is the number of potential TFs. The Elastic net regression estimates the coefficients of *M* TFs, denoted by an M-dimensional vector, B= argmin\_B ||Y-XB||^2+alpha\*||B||^2+beta\*||B||\_1, where alpha and beta are parameters to adjust the contributions from L2 and L1 regularizations of B.

Furthermore, we used Hi-C data to highlight the activated cis-regulatory relationships if they are in the Hi-C TAD regions.

# 7 S (TL, ||) Integrative modeling

## 7.1 S (HL, ||) Genotype - gene expression/enhancers - modules - phenotype (JW, DW) [[JW to update]]

We integrate data of the kinds described above into a single model connecting genotype, phenotype and functional genomics data from PsychENCODE in the Prefrontal Cortex. As for the eQTL and cQTL calculations above, we normalize the RNAseq and epigenomic data by calculating PEER factors to remove the effects of known and unknown confounders. Further, we calculate PFC-specific modules using WGCNA. We use the phenotype information in PsychENCODE to construct a number of different high-level traits, including discrete traits for disease phenotypes (Affective disorders, Bipolar and Schizophrenia phenotypes) and Sex (Male/Female), and an ordinal representation of age by binning the samples into quintiles.

## 7.2 S (TL or HL? - unclear from content, || or [[NPAR]] - unclear from content]]) Enrichment analysis for modules (JP, DW) [[JW to run enrichments for final modules]]

xxx

## 7.3 S (TL, ||) Deep learning model for predicting brain genotype-phenotype (JW, DW)

We construct a *Structured Deep Boltzmann Machine* (sDBM) model to learn the conditional distribution of the transcriptome / epigenome / trait data given the genotype data. This is based on the Deep Boltzmann Machine (DBM) framework \cite{Salakhutdinov09}, which we adapt so as to include layers with sparse intra-level connectivity, which we use to represent gene regulatory interaction dependencies within the transcriptome, and enhancer-gene dependencies between transcriptome and epigenome. Further, we conditionalize the sDBM model globally so that it is dependent on genotype, hence forming a Conditional Random Field (CRF) (see \cite{Koller09}). The model can be expressed as:

where is a layer of visible units representing the transcriptome (gene expression levels) and epigenome (H3k27ac levels), are two levels of hidden units representing levels modules / pathways at multiple scales, represents a phenotypic trait (a binary or multinomial variable), and represents the genotype. The potentials have the form of restricted RBMs between layers (with conditional dependence on the genome in the first potential), a multinomial activation potential between the final hidden layer and the phenotype, and a sparse Gene Regulatory Network (GRN) based potential between visible units, as described above.

To learn the network described, we perform pre-training using Contrastive Divergence as in Salakhutdinov et. al. We then perform global training using mean-field and persistent MCMC as in Salakhutdinov et. al, which we adapt to include the sparse GRN potential and the conditional dependence on genotype. To incorporate conditional dependence, we adapt mean-field and persistent MCMC steps to estimate conditional expectations as in CRF training (see \cite{Koller09}), requiring us to sample from a set of persistent Markov Chains for each genotype.

In particular, we enforce sparse connectivity between the genotype and gene expression layers based on the estimated eQTLs, and between genotype and enhancer layers based on estimated cQTLs. Further, we introduce sparse intra-layer (lateral) connections within the gene-expression layer using the TF-promoter links estimated in our Gene Regulatory Networks, and sparse connections between the gene-expression and enhancer sub-layers using the estimated TF-enhancer enhancer-target links. fQTLs and modQTLs are represented by sparse connections jumping directly from the genotype to the first intermediate layer, while the connectivity between module/cell-fraction and expression/enhancer nodes is fixed based on prior gene-sets. Other than these relationships, full connectivity is used between nodes at all adjacent layers.

In particular, our inference is performed using a mean-field approximation, and training is performed using a Persistent Markov Chain Monte Carlo algorithm which is able to ensemble multi-dimensional datasets (Supplement).

## 7.4 S (TL, ||) Imputation using integrative modeling (JW, DW)

We perform imputation of transcriptome and epigenomic data by performing mean-field inference in the model expressed in Eq. 1, treating the variables to be imputed as hidden variables. Briefly, for imputation of transcriptome/epigenome data given genotype data along with prediction of the phenotype, we estimate argmax, hence jointly predicting the transcriptome/epigenome and phenotype variables, while marginalizing out the hidden layers. To impute epigenome data and predict the phenotype given transcriptome and genotype data, we estimate argmax, where are the units representing the transcriptome and epigenome respectively. Imputation when the trait variable is known is performed similarly, but we additionally condition on the phenotype.

## 7.5 S (TL or HL? - unclear from content, || or [[NPAR]] - unclear from content]]) Inheritable variation

# 8 S (HL, ||) Resource website (PE, SL) S

[https://docs.google.com/document/d/1Vq6vdVPIXdkZ5LrmXzsqWuxt2RZDiScidkq2aNRmkCY/edit?usp=sharing](https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdocs.google.com%2Fdocument%2Fd%2F1Vq6vdVPIXdkZ5LrmXzsqWuxt2RZDiScidkq2aNRmkCY%2Fedit%3Fusp%3Dsharing&data=02%7C01%7Cdaifeng.wang%40stonybrookmedicine.edu%7C85ad7a29fb38455df1f708d52c7e9819%7Ceafa1b31b194425db36656c215b7760c%7C0%7C0%7C636463844605497590&sdata=Vn8iIwfjjtYEhDHbEQFJ8ry6j50DxY%2BxSQSFSmUQxGM%3D&reserved=0)

[http://adult.psychencode.org/](http://psychencode.gersteinlab.org/)

# 9 S (HL [[NPAR]]) References

1 Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome research* **27**, 1939-1949, doi:10.1101/gr.220640.117 (2017).

2 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

3 Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527, doi:10.1038/nature19847 (2016).

4 Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

5 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).

B. Zhang and S. Horvath. A General Framework for Weighted Gene Co-expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 2005.

P. Langfelder, B. Zhang and S. Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 2008.

R. Salakhutdinov and G. Hinton. Deep Boltzmann Machines. *AISTATS*, 2009.

D. Koller and N. Friedmann. *Probabilistic Graphical Models*, 2009.

GTEx consortium. Genetic effects on gene expression across human tissues. Nature, 2017