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

Abstract

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

Introduction

Disorders of the brain affect nearly 20% of the world's population (ref). Unlike cardiac disease, where lifestyle and pharmacological modification of environmental risk factors has had a profound effect on disease morbidity and mortality (ref), or cancer, which is now understood to be a disorder of the genomic functions (ref), until recently, little progress has been made in our fundamental understanding of the molecular cause of the brain disorders. This recent progress has come is the form of genetic association signals from large GWAS studies of the psychiatric and neurological disorders and currently hundreds of genomic locations that alter the disease risk are known (ref of review, or list disorders in text below, depending on space). Unfortunately, for most of these locations, we have little to no understanding of which base pairs alterations constitute the functional genomic alteration, which transcripts and networks are altered, and what are the molecular mechanisms that cause those alterations. It is presumed that these changes in transcription modify the proteome, which leads to changes in brain

Formatted: Not Highlight

structure and function, and these changes interact with environmental factors to change the probability of developing a brain disorder.

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

In fact, understanding the molecular mechanisms on how these genomic elements affect various brain functions and phenotypes is still a key challenge in neuroscience. To address it, the PsychENCODE Consortium integrates a group of projects to produce a public resource of multi-dimensional genomic data from thousands of high quality healthy and diseased human post-mortem brains (PEC ref) (6). Particularly, it has generated and assembled a robust large-scale dataset on the adult human brain to address this challenge, including genotyping, RNA-seq, ChIP-seq and single-cell transcriptomic data on ~2000 (or 1945) adult brain tissue samples with different phenotypes and these data are housed in a central, publically available depository (xxxx). In addition, for these analyses, we have supplemented the PEC data with the primary data from other related genomic resources, such as: ENCODE, CommonMind, GTEx, Epigenomics Roadmap, etc, and uniformly processed all the data together and performed integrated analyses with up to X,XXX samples. We have also supplemented the PEC single-cell data with the primary data from recent publications (refs), reprocessed and analyzed all the

Deleted: detect

Deleted: mega-

data jointly to find gene expression signatures and calculate the fractions of neuronal and non-neuronal cell types in normal and disease states. We provide all the PEC data and mega-analyses in an online resource which contains all possible functional genomic elements for adult brain including the brain-active enhancers, transcripts, expression models, imputed regulatory networks, eQTLs and cQTLs for various phenotypes, and an integrated deep-learning model, Deep Structured Phenotype Networks (DSPN) for predicting and imputing brain phenotypes. We then use this resource to discover the properties of brain gene expression, non-coding transcription and enhancers, and to build this DSPN model, to describe how interactions between genomic variants, gene expression, enhancers might work together molecularly to alter disease risk.

Deleted: , the [[PEAR]]

Deleted: . We then use the [[PEAR]]

Deleted: an integrative deep neural network model, named eTWAS

Comprehensive resource for adult brain functional genomics

The PsychENCODE consortium has generated and assembled a large-scale dataset of genotypes, RNA-seq, ChIP-seq, ATAC-seq, Hi-C and single-cell transcriptomic data from adult brains of 1945 individuals, with and without several mental illnesses (Figure 1, Assay summary in Methods). To harmonize and integrate the datasets across multiple consortia, we processed these datasets using standard bioinformatic pipelines in common use (Methods). For instance, we adopted the ENCODE processing pipelines for the bulk and single cell RNA-seq and ChIP-seq data. Likewise, we used the GTEx eQTL pipeline and associated parameters, to allow comparison to previously published eQTL maps. All these uniformly processed datasets are available in our XXXX resource (URL here). Finally, we also compared the resource data against various phenotypes, and identified the brain specific data (derived data type). For example, this resource includes the regulatory variants such as QTLs, brain active enhancers, differentially expressed genes and transcripts, novel transcribed regions and non-coding RNAs, and putative genome-wide regulatory networks. It is also publicly accessible and available on the PyschENCODE website (xxxx), such as using the interactive web app.

Deleted: ~2000 (or

Formatted: Not Highlight

Deleted:)

Overall, this resource is structured in a pyramid shape (Figure 1), with the largest scale and raw data at the bottom level and the lightest and most interpretive data at the top level.

Next generation sequencing data for brain functional genomics

At the bottom, we have the large scale raw data and the phenotype information for 1945 individuals, much of which is private and under controlled access. Based on this, we have then uniformly processed raw datasets from PyschENCODE and other consortia (ENCODE, CommonMind, GTEx, Epigenomics Roadmap, etc), including RNA-seq expression quantifications, ChIP-seq signal track qualifications and peak identifications using ENCODE standard pipelines, and private imputed genotypes. The processed functional genomic data is much easier to interpret but still rather large scale. In details, they include the following major types:

Deleted: ~2000

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

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

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

Chromatin structures – we generated and processed the Hi-C data for adult brain, and identified the xxx regions on which the enhancers and promoters interact. These regions (e.g., TADs) enable the systems identification of potential regulatory enhancers if they interact the target genes' promoters. [more from HJ&DH]

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

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

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

Topologically associating domains – we used a full Hi-C data for adult brain and identified xxx in xxx Topologically Associating Domains (TADs) of adult brain. These TADs provide the regions at which the enhancers interact with target gene promoters in adult brain. [more from HJ&DH]

Differentially expressed genes, transcripts and brain splicing patterns - we compared expression changes in uniformly processed RNA-seq data from brain samples across PsychENCODE-related studies, ENCODE, and GTEx, and found xxx expressed genes and ~79,000 transcripts in pre-frontal cortex, ~11k eGenes associated with eQTLs (Methods), and xxx non-coding RNAs and novel transcribed regions. We also derived phenotype-specific genes and transcripts.

Deleted: We

Deleted: summarize

Deleted: ~88k

Formatted: Font color: R,G,B (34,34,34)

Deleted: cortex and cerebellar

Formatted: Font color: R,G,B (34,34,34)

In addition, we calculated the alternative splicing patterns at the transcript level; i.e., the percentage of the transcript abundance over its gene abundance, and found the brain-specific spliced transcripts. Our resource contains differentially expressed and spliced genes and transcripts across a number of biological variables, including neuropsychiatric disorders and developmental stages.

We should emphasize that our comparative analysis is consistent for finding various brain elements including brain enhancers, genes and transcripts. More specifically, we compared them against a same set of brain and non-brain tissues; e.g., the RNA-seq gene expression data from GTEx and the ChIP-seq H3K27AC binding signal data from Epigenomics Roadmap for brain pre-frontal cortex vs. other non-brain tissues including liver, lung, blood, etc.

Deleted: to find

Deleted: is consistent; e.g., via comparing with

Deleted: regions

Deleted: organs

System identification of the QTLs and gene regulatory networks associated with adult brain transcriptomics and epigenomics

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

Deleted:) and

Formatted: Font color: R,G,B (34,34,34)

Given a great number of QTLs we identified, we are further interested to see how they relate to the known variants for brain. In particular, we compared them with existing QTLs databases and subdivided our QTLs into different functional categories, mainly including the disease GWAS SNPs, the SNPs breaking the TF binding sites, etc (Table/Figure xxx). Collectively, these QTLs annotate a larger fraction of GWAS SNPs involving the brain (e.g., 6% in schizophrenia, 10% in bipolar) than previously observed, providing leads on which genes are affected in disease. We also evaluated the overlap of eQTLs with cQTLs and found that XX% of cQTLs are overlapped with eQTLs. The SNPs in cis-eQTL list(Cis-eSNPs) were enriched within XXXX, and depleted XXXXXX (Fig. X). We examined the enrichment of most significant eQTLs per gene in

Deleted: For example, we found that

Formatted: Font color: R,G,B (34,34,34), Highlight

Deleted: variants cover

Formatted: Font color: R,G,B (34,34,34), Highlight

Deleted: disease-associated brain

Formatted: Font color: R,G,B (34,34,34), Highlight

Formatted: Font color: R,G,B (34,34,34), Highlight

Deleted: any previous analyses, suggesting potential molecular targets for these associations (xx% for SCZ, xx% for BP, ASD) and approaching the saturation of human mutations (Figure xxx).

Roadmap Epigenomics Consortium and ENCODE enhancers across XX human tissues and cell lines. Cis-eQTL were enriched for enhancer sequences present in brain tissues and the strongest enrichment is observed in DLPFC enhancers. We also calculate the enrichment of cis-QTLs on GWAS SNPs of brain related disorders (schizophrenia, bipolar disorders and parkinson's disease) and non-brain related disorders (CAD, asthma and type 2 diabetes). Cis-QTLs have more significant enrichment for GWAS SNPs of brain related disorders than the ones of non-brain related disorders. In addition, we link the QTLs that overlap the enhancers and promoters in the resource to reveal the potential regulatory activities. We thus classified the QTLs into subgroups in terms of their gene regulatory characteristics including the regulatory QTLs (rQTLs) that break TF binding sites on promoters and/or enhancers, and the modular QTLs (mQTLs) that highly associate with a set of co-expressed genes. Finally, we found that the eQTLs/eGenes number can be predicted from the sample size using a fitted curve (Figure xxx).

Gene regulatory networks - we also integrated and imputated the regulatory relationships in brain such as the enhancers, transcription factors (TFs), miRNAs and target genes [refs] in this resource (Methods). For example, we found the TF binding motifs using ENCODE data and inferred the TF-target gene relationships if TFs have enriched binding motifs on the target gene's regulatory regions such as promoters and enhancers. We also used Hi-C data to filter the enhancers that are not in the TAD regions for given target genes. In total, we included xxx enhancer-gene, xxx TF-gene, and xxx miRNA-gene regulatory linkages, providing a reference wiring network on gene regulation in brain. It should be noted that activations of these regulatory wires are highly attributed to the genotypes of QTLs, leading to various phenotypes. Thus, using these "wiring" regulatory relationships, we inferred the gene regulatory networks that identify the regulatory relationships on how QTLs, enhancers, and transcription factors relate to target gene expression (Methods). In particular, given a target gene, we found its related regulatory elements from the resource including the eQTLs, the enhancers that control its gene expression [JEME] plus their cQTLs, and predicted the transcription factors (TFs) that have enriched binding sites on these enhancers and its promoter. We then used RNA-seq and ChIP-seq data based on the Elastic Net model with regularization that combines the L1 and L2 penalties of the lasso and ridge regressions to predict the regression coefficients of genotypes of various QTLs, the chromatin stages of enhancers, splicing patterns and TFs gene expression to the target gene expression, and identified the highly predictive relationships (i.e., large coefficients). We repeated this for all genes and found how various subgroups of QTLs affect gene expression; e.g., a significantly number of predictive QTLs break the TFBSs on the enhancers or promoters (xx%, Figure xxx). We thus constructed a gene regulatory networks consisting of the QTLs, enhancers, TFs and target genes with high predictive relationships (coeff. > xxx, Methods), revealing the biological mechanisms on how QTLs regulate the target gene expression in the adult brain.

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

activities, and the biological mechanisms explaining how the brain specific elements affect the phenotypes and diseases in the adult brain.

Comparative analysis reveals the brain related transcriptomic and epigenomic activity

We leveraged this resource to compare the human brain with other tissues. To reveal potential brain specific genomic activities, particularly relating to transcriptomic and epigenomic activities, we performed a consistent spectral analysis and compared the similarities of gene expression and H3K27AC binding signals on enhancers and found that the brain has more distinct expression patterns compared to most other tissues, including a greater amount of non-coding transcription. However, the differences in epigenetics are relatively smaller.

For gene expression, we compared the adult brain samples from our resource with the other tissue samples from GTEx, using uniformly reprocessed RNA-seq data. It shows that the brain samples, though from different studies are clustered together in a major cluster, significantly separated from the other major cluster consisting of non-brain samples from their leading reduced dimension (left vs. right clusters in Figure xxx). This suggests that the brain has unique and distinctive gene expression programs, which are involved by the brain elements including brain expressed genes, transcripts and non-coding RNAs in our resource. In addition, the samples of PsychENCODE that include psychiatric disorders have larger variations than the reference brain samples and other tissue clusters (Figure xxx). The cluster radiuses were estimated by fitting the two main principal components into a multivariate normal model and finding a 0.95 confidence interval (Methods). This suggests that the psychiatric diseases still have larger variations of gene expression, and different gene regulatory programs from the normal, though even more distant from other organs. Additionally, to understand where the human brain sits in regards of its the transcription diversity compared to other tissues, we estimated the proportion of genome that is transcriptionally active across hundreds of samples. We first found that transcript diversity is mostly saturated at the scale of hundreds of individuals (Figure xxx). The saturation is observed for both the coding and non-coding portions of the genome. The human brain does not stand as a highly diverse in protein coding regions. For example, the tissues such as the testis is highly diverse [Ref]; however, we found that the brain has more transcriptional activity at the non-coding and novel transcribed regions than most other tissues (Figure xxx). Which implies that the non-coding transcription is highly likely another factor to make the brain tissues unique.

As shown above, the brain samples have different chromatin and gene expression activities from other organs, implying that the brain also has specific gene regulatory activities. Therefore, we are further interested to compare the enhancers between brain and other tissues to see any brain epigenomic activities. In particular, we integrated the H3K27Ac ChIP-seq signal data of enhancers in the resource and performed the consistent spectral analysis for gene expression as above to compare the similarities of epigenetic profiles of PsychENCODE samples with Epigenomics Roadmap data. It is also interesting to find dissimilar patterns with the gene expression comparison; e.g., while the brain samples separates from other tissues when using

Deleted: that

Deleted: has specific

Deleted: This comprehensive resource allows us to discover the specific functional genomic elements that relate the brain functions and phenotypes as above. Thus, we

Deleted: against various phenotypes and compared

Deleted: tissue types to

Deleted: the unique

Deleted: . In particular

Deleted: same

Deleted: for comparing

Deleted: between brain versus

Deleted: (Figure xxx),

Deleted: .

Deleted: there exist

Deleted: distinct

Deleted: differentially

Deleted: that make brain very different from other tissues.

Deleted: this major brain cluster has a particular geometric pattern showing that

Deleted: normal

Deleted: (e.g., GTEx) form an inner core, and the disease samples form a subcluster having

Deleted: radius

Deleted: subcluster consisting of the normal

Deleted: (e.g., GTEx

Deleted: radii

Deleted: Gaussian mixture

Deleted: Also, the major cluster can be further subdivided into several subclusters, each of which mainly comprises the samples from same brain region; e.g., the cortex and cerebellum clusters in Figure xxx. However, the distances among these sub brain clusters are significantly less than the ones among other organs, suggesting that the brain regions, though functionally different, still need to more closely coordinate with each other than other organs.

Deleted: regulatory regions

Deleted: specific regulatory

Deleted: We

Deleted: somewhat similar

Deleted: can also cluster together in terms of active enhancer similarity

genes expression data, the active enhancers are not able to separate brain from other tissues (Figure xxx). This result suggests that the brain has less specific and distinct epigenomic activities, involving the brain active enhancers from our resource. Thus, there may exist more complex regulatory mechanisms among the brain enhancers with low signal variability than other tissues to drive the brain distinct gene expression. One important mechanism is that the brain active enhancers or gene expression patterns are intermediate phenotypes, potentially driven by particular large set of brain regulatory variants such as our QTLs as previously described.

Our comparative analysis reveals that the brain is different from other organs in gene expression. Thus, we are then interested to identify the functional genomic elements in brain that give rise to the uniqueness of brain. To systematically find the specific expressed functional elements in brain, we identified the differentially expressed genes and non-coding RNAs for various phenotypes including mental disease, gender, regions (Methods and Figure XX) for the resource. For example, xxx genes have been found to differentially express between SCZ and normal samples; i.e., SCZ DEX genes, and they are also enriched with the pathways and functions relating to SCZ (Figure Sxxx). Moreover, we identified a group of genes that differentially express across different ages (Figure xxx). For example, the gene involved in early growth response is down-regulated at elder samples whereas the gene with ceruloplasmin is down-regulated around the middle ages. Finally, we report the DEX genes for all phenotypes in our resource along with their enriched functions and pathways in supplement. Also, the brain specific gene expression is likely driven by a group of genes, rather than individual genes, so we constructed the gene co-expression network using all PsychENCODE and GTEx samples, and clustered it into gene co-expression modules using WGCNA [Methods]. The genes clustered in a same module are highly likely co-regulated by similar mechanisms. Our co-expression analysis indeed found several modules whose eigengenes show very different expression levels between brain and non-brain samples (Figure Sxxx, Supplement), which suggests that there exist brain specific regulatory mechanisms drive these brain co-expression modules.

Single cell analysis and deconvolution explain gene expression changes across adult phenotypes

The brain tissues have been found to comprise a variety of cell types including neuronal and non-neuronal cells such as astrocytes [refs]. One issue with the changes of gene expression in our brain tissue samples is whether the changes are driven by gene expression in a particular cell type or different cell-type populations. To address this tissue, we integrated the single cell gene expression data to discover how the gene expression from various cell types including both neuronal and non-neuronal contribute to the gene expression at the tissue level. In particular, we used the biomarker genes with strong expression signals in single cell to deconvolve the gene expression data of individual tissues over both novel and known cell types to find the cell fractions for individuals, and relate to the individual phenotypes. We found that the gene expression changes across individual tissue samples can be largely explained by the

Deleted: as well

Deleted: More importantly,

Deleted: either

Deleted: Table XXX

Deleted: XXX

Deleted: male

Deleted: female

Deleted: . We

Deleted: checked the

Deleted: among the SCZ genes, and indeed found that many are

Deleted: male.

Deleted: also found that these brain dex genes are significantly less/greater than DEX genes for other tissues

Deleted: GTEX (p<xxx), which suggesting that

Deleted: brain expression uniqueness is highly driven by a small/large set of genes. As previously described

Deleted: xxx

Deleted: changes of brain tissue genes across

Deleted: . We also

Deleted: adult brain phenotypes at the

Deleted: level

Deleted: more easily

single cell gene expression, and the changes of single cell fractions are also associated with the individual phenotypes.

Specifically, we integrated and used the same pipeline to uniformly process the single cell RNA-seq data for ~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 include these single cell data and cell-type biomarker genes in the resource. Moreover, for those differentially expressed genes at the tissue level from our resource, we further checked their expression changes across various single cells, and found that a group of psychiatric disorder related genes indeed show the expression dynamic changes among cells. For example, the dopamine receptor genes (DRD) that associate with SCZ, are significantly more highly expressed in neuronal cells than others (Figure Sxxx), and their expression levels across cells vary significantly larger than tissue samples, suggesting that the cell fraction changes potentially equalize the tissue expression variability. Therefore, we are then interested to see if the brain gene expression at the tissue level in our resource is contributed by the above cell types and affected by the cell fractions.

To this end, we decomposed the gene expression data across individuals at the tissue level from our resource using non-negative matrix factorization (NMF, see Methods). Indeed, we found that three groups of top principal components of NMF (NMF-PCs) capturing the most covariance of brain gene expression across individual tissues, highly correlate with the biomarker gene expression signatures of neuronal, non-neuronal and fetal cell types as above, respectively (three blocks in Figure xxx). 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.

Therefore, we deconvolved the tissue-level gene expression data of all 1945 samples using single-cell gene expression data of 450 biomarker genes to find the fraction of different cell types corresponding, and compare cell fractions across different phenotypes (Supplement). The single cells used in deconvolution cover all 16 neuronal types, five non-neuronal types and xxx additional fetal types from PsychENCODE single cell data [ref: brainspan]. It is very interesting that the linear combinations of single cell expression of 23 cell types, where combinational

Deleted: 800 cells from PsychENCOCE, ~

Deleted: ,

Deleted: xxx novel

Deleted: .

Deleted: then

Deleted: xxx).

Deleted: . ... [1]

Deleted: . Thus

Deleted: two

Deleted: and

Deleted: neuronal and non-neuronal cells'

Deleted: For those differentially expressed genes at the tissue level from our resource, we further checked their expression changes across various single cells, and found that a group of differentially expressed genes indeed show the expression dynamic changes among cells. For example, the SCZ gene, XXX is (or ww% of SCZ genes) significantly more highly expressed in YYY and ZZZ neuronal cells than others (Figure xxx), suggesting that the cell fractions of YYY and ZZZ drive the SCZ gene expression changes across tissues[ref].

Deleted: 2000

Deleted: xxx

Deleted: proportions

Deleted: Y=WX, Methods

Deleted: types. For example, it

Deleted: we

coefficients, can explain >80% of the gene expression variations across 1945 individual tissues (Figure xx). The coefficients of cell types for linear combination are estimated from our deconvolution analysis (Methods in supplement), and proportional to the cell fractions of individuals. In addition, we found that the cell fractions of individuals (i.e., deconvolution coefficients) vary, and a number of cell population changes highly associate with different phenotypes and disorders (Figure xxx). For example, the fraction(s) of neuronal type(s) (Inhibitory X) is significantly anti-correlated with Age (r = xxx). The excitatory neuronal cell populations (e.g., EX1) increase significantly in ASD samples (p<xxx) while the non-neuronal cells decreasing (e.g., oligodendrocytes). Finally, we report the individual cell populations along with significantly associated relationships between particular cell type fractions and phenotypes (Supplement).

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

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

The interaction between genotype and phenotype is a very complex process, involving multiple intermediate stages including gene expression, signaling, modulation and so on. Thus, to understand the entire process of how genotype and phenotype relate to each other, we introduce an interpretable deep-learning framework, Deep Structured Phenotype Networks (DSPN), which provides insight into how the brain genomic variants affect gene expression and regulation, and eventually predict phenotypes (Figure xxx). This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from the QTLs and regulatory networks estimated in our resource. It integrates all high dimensional functional data types in this resource including genomics, transcriptomics, epigenetics and regulatomics, and genotype-phenotype relationships, and also allows us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only. The model is trained as a deep generative model to represent the conditional distribution of all variables given the genotype. Unlike a feed-forward network architecture, the undirected form of the Boltzmann machine allows information to flow in top-down, bottom-up and lateral directions during inference, so that intermediate and high-level phenotypes may be jointly inferred while respecting their mutual dependencies. This allows us for instance to impute transcriptome and epigenome data when it is missing. Inference is performed using a mean-field approximation, and training is performed using a Persistent Markov Chain Monte Carlo algorithm (see supplement).

Deleted: much (R2=~
Formatted: Font color: Black, Pattern: Clear
Deleted: %)
Deleted: individual variation in
Deleted: of both male
Formatted: Font color: Black, Pattern: Clear
Formatted: Font color: Black, Pattern: Clear
Formatted: Font color: Black, Pattern: Clear
Formatted: Font color: Black, Pattern: Clear
Deleted: female samples in terms of changing proportions of basic cell types, rather than changes in individual genes (Figure xxx covariance).
Deleted: vary across different phenotypes (Figure xxx),
Deleted: brain
Deleted: .
Deleted: non-
Deleted: SCZ (or Male)
Deleted: .
Deleted:).
Deleted: SNPs

Deleted: interactions
Deleted: genotypes
Deleted: phenotypes
Deleted: experiencing
Deleted: processes
Deleted: genotypes
Deleted: phenotypes affect
Deleted: built
Deleted: integrative model, eTWAS to understand
Deleted: the
Deleted: is built based on
Deleted: ,
Deleted: integrated
Deleted: allowed
Deleted: It uses the undirected edges rather than the directed edges of deep neural network modeling because the phenotypes potentially impact back to the intermediate stages like gene expression. As shown in Figure xxx, the eTWAS consists of four layers: 1) genotypes such as QTLs; 2) gene expression and enhancers; 3) intermediate modules and 4) phenotypes such as brain traits, and provides the additively predictive relationships between layer nodes. In particular, the model is constructed based on the Deep Boltzmann Machine (RBM) but has a hybrid structure. On one hand, it incorporates the contemporary deep learning ideas to model these large scale datasets with a multi-layer architecture with interconnections ... [2]

As shown in Figure xxx, the DSPN consists of four layers: 1) genotypes such as QTLs; 2) molecules and genomic elements, including genes and enhancers; 3) functional modules and other mid-level phenotypes at a series of intermediate layers; i.e., the hidden nodes of deep learning modeling; 4) high-level phenotypes such as brain traits. In addition, we enforce the DSPN to have sparse connectivity (Supplement). Specifically, we built each layer of our model as follows. We first used the imputed gene regulatory networks that identify the regulatory connectivities on how QTLs, enhancers, and transcription factors relate to target gene expression (Supplement). We then connected the nodes on Layer 2 of our model to follow the inferred gene regulatory network structures; i.e., embedding the gene regulatory network. In particular, many intermediate-layer modules (i.e., strongly predictive features on Layer 3) that correspond to known gene sets associated with well-characterized pathways and functions in the brain; e.g., the module xxx is connected to genes enriched in the dopaminergic and glutamatergic synapse (GSEA enrichment score > xxx, Figure xx). Also, some modules are used to capture the information on single cell populations; e.g., the module yyy is connecting to Age, and represents the neuronal cell fractions (Figure xxx). Furthermore, we used this model to recapitulate the pathways comprising the cross-layer nodes and predictive edges for particular phenotypes. For example, as highlighted in Figure xxx, the schizophrenia (SCZ) trait is activated by two modules on the layer of hidden nodes corresponding to glutamatergic signaling and excitatory synapse, respectively. The modules are connected by a set of genes including GRIN1, which are regulated by corresponding QTLs (e.g., rs1146020) and enhancers (e.g., GH09H137166) as shown in the blowup gene regulatory mechanism. In addition, we discovered novel molecular mechanisms for SCZ such as module(s) corresponding to dopamine-related pathways and complement pathways (Figure xxx). These modules are connected to the C4 family genes, regulated by eQTLs and enhancers (p<1e-4).

Moreover, the model also enables practical imputation of a subset of the transcriptome and epigenome, with an accuracy of ~70% (Figure xxx). We use the model to improve prediction of biological variables and psychiatric diseases by the addition of transcriptomic data to genotype, as compared to genotype alone. In particular, we can predict bipolar disease and schizophrenia with much higher accuracy from the transcriptome than from genotype alone; i.e., three times improvements (+18% vs. +6%) from the random prediction 50% for schizophrenia, Figure XXX). The imputed transcriptome also clearly adds predictive value, as we can predict schizophrenia with an accuracy of 61% using our model and an imputed transcriptome compared to 56% with genotype alone. This result demonstrates the usefulness of even a limited amount of functional genomics information for unraveling gene-disease relationships. On the resource website, we provide a list of DSPN pathways for each endophenotype and disease. We also make the model available as distributive software and as a set of simplified files summarizing represented genotype-phenotype pathways.

Discussion

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

Deleted:

Deleted: this

Deleted: from the resource

Deleted: Methods

Deleted: connecting

Deleted: the

Deleted: with ZZZ pathways (p<xxxx

Deleted: populations. We show that this integrated model has significantly improved the prediction accuracy over individual genomic data types. For example, its AUC/MSE for classifying SCZ and health samples is xxx beating other classification methods using gene expression only (Table XXX).

Deleted: The trait of

Deleted: , x,

Deleted: y corresponding to dopamine-related pathways and complement pathways

Deleted: Each module is

Deleted: C4 genes

Deleted: eTWAS

genomic elements for the adult brain. Developing this resource and integrated model to a population-level scale serves as an important step in gaining meaningful biological insights from functional genomics studies in neuroscience. In particular, we compared it with other tissues such as GTEx data and identified the genotypes and QTLs, the specific expressed genes, transcripts and noncoding RNAs, active chromatin regions, the regulatory networks that significantly relate with different brain phenotypes at both cellular and tissue levels. For example, the QTLs allow one to potentially interpret most of the known brain-associated GWAS SNPs in terms of perturbations to specific genes. Thus, the neuroscientist can use this resource as a reference to compare with their data, generate hypotheses and help design experimental validations. In addition, this resource is publicly available online and can be extendable and scalable to integrate additional data types and phenotypes. For example, it can add the individual's fMRI image features measuring functional neuro-connectivity, and use our model to identify the genotypes that associated with image features such as image-QTLs (iQTLs) [xx]. Also, our resource can incorporate with the neurodegenerative diseases like Alzheimer or developmental stages.

Moreover, we built an integrative epigenome- and transcriptome-wide association model (eTWAS), built on the Deep Boltzmann Machine (RBM) and integrates the high dimensional functional genomic and phenotypic data at multiple layers, using the hierarchical structures in deep learning. The model reveals the relationships among various data types from a number of directions for genotype to phenotype. In particular, this model also incorporates the derived data types into its hierarchical structure such as imputed gene regulatory networks and QTLs, and provides the additional statistical powers to better predict the genotype to phenotype. This model allows us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only. More importantly, it integrates high-dimensional functional genomics data with genotype-phenotype associations to highlight key brain genes and modules and relate how variants in these regulate gene expression. This integrative model is also available online as a general purpose platform. The users can apply it to impute missing data , predict the genotype-phenotype relationships, and reveal potentially novel gene regulatory mechanisms and modules for additional phenotypes. Also, the model can be used to make in-silico predictions for the perturbation outcomes. For example, we can identify the module X that have the extremely highest connection weights to Austin, and thus knocking down the genes connecting to the module highly likely will deactivate Autism. Furthermore, while the model does provide better predictive performance, some of these correlations are deliberately set to be interpreted simplifications, such as the known enhancers, or gene regulatory network structure, to make the model more interpretable and easier to use. Thus, another major goal of the model is to provide a compression of larger amount of functional genomic datasets for brain; e.g., XXX KB of model files vs. XXX TB of total resource data, beyond a purely predictive network from genotype to phenotype.

Though single cell remains challenging to reliably quantify the low-abundant transcripts/genes and interrogate the biological variations using single-cell sequencing technology, it is still worthwhile using the biomarker genes with strong expression signals in single cell to deconvolve the gene expression data of individual tissues over both novel and known cell types

to find the cell populations for individuals, and relate to the individual phenotypes. With increasing amount of single cell data in near future, we could deconvolve the resource data at tissue level to find potential new cell types and obtain more complete cell populations. The current single-cell sequencing technology suffers from the low capture efficiency [PMCID: PMC4758375, PMCID: PMC4132710]. Due to this reason, the single-cell sequencing will only measure a small fraction of cellular transcriptome as the final sequencing library only contains a subset of input materials. Furthermore, the limited amount of RNA molecules in single cell makes it even harder to capture the weak signals, which makes the data sensitive to technical noise. Thus, given that the RNA decaying issues in single cell RNA-seq, we could also relate this resource to the in situ transcriptomic data such as optogenetic techniques measuring the spatial gene expression, and find the consistent expressed gene for the brain phenotypes at the tissue level.

Formatted Table

References

1. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, Shah HR *et al*: **Gene expression elucidates functional impact of polygenic risk for schizophrenia**. *Nat Neurosci* 2016, **19**(11):1442-1453.
2. Consortium GT: **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans**. *Science* 2015, **348**(6235):648-660.
3. Psych EC, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S *et al*: **The PsychENCODE project**. *Nat Neurosci* 2015, **18**(12):1707-1712.
4. Neale BM, Sklar P: **Genetic analysis of schizophrenia and bipolar disorder reveals polygenicity but also suggests new directions for molecular interrogation**. *Curr Opin Neurobiol* 2015, **30**:131-138.
5. Schizophrenia Working Group of the Psychiatric Genomics C: **Biological insights from 108 schizophrenia-associated genetic loci**. *Nature* 2014, **511**(7510):421-427.
6. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida *et al*: **Genetic effects on gene expression across human tissues**. *Nature* 2017, **550**(7675):204-213.
7. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A *et al*: **Population Variation and Genetic Control of Modular Chromatin Architecture in Humans**. *Cell* 2015, **162**(5):1039-1050.
8. Roshyara NR, Horn K, Kirsten H, Ahnert P, Scholz M: **Comparing performance of modern genotype imputation methods in different ethnicities**. *Sci Rep* 2016, **6**:34386.
9. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K *et al*: **A reference panel of 64,976 haplotypes for genotype imputation**. *Nat Genet* 2016, **48**(10):1279-1283.

10. Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D *et al*: **Chromosome conformation elucidates regulatory relationships in developing human brain**. *Nature* 2016, **538**(7626):523-527.
11. Geschwind DH, Flint J: **Genetics and genomics of psychiatric disease**. *Science* 2015, **349**(6255):1489-1494.
12. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O: **Fast and efficient QTL mapper for thousands of molecular phenotypes**. *Bioinformatics* 2016, **32**(10):1479-1485.
13. What constitutes the prefrontal cortex? Science 2017, DOI: 10.1126/science.aan8868

Supplement

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

Page 9: [1] Deleted	Daifeng Wang	1/15/18 6:22:00 PM

We

Page 10: [2] Deleted	Daifeng Wang	1/15/18 6:22:00 PM

It uses the undirected edges rather than the directed edges of deep neural network modeling because the phenotypes potentially impact back to the intermediate stages like gene expression. As shown in Figure xxx, the eTWAS consists of four layers: 1) genotypes such as QTLs; 2) gene expression and enhancers; 3) intermediate modules and 4) phenotypes such as brain traits, and provides the additively predictive relationships between layer nodes. In particular, the model is constructed based on the Deep Boltzmann Machine (RBM) but has a hybrid structure. On one hand, it incorporates the contemporary deep learning ideas to model these large scale datasets with a multi-layer architecture with interconnections between layers, and also explicitly allow integrating additional genomic elements into the model such as incorporating imputed eQLTs and cQTLs. The RBM architecture, especially undirected edges can reveal the relationships among functional genomic elements across layers from a number of directions, rather than one direction in classical deep neural networks. Moreover, using these relationships, the model can be used to better predict phenotypes from genotypes, through adding predictive powers from gene expression and chromatin data; e.g., gene regulatory networks. On the other hand, given known associated genotypes and phenotypes, this model can trace their all possible connectivities and better pinpoint them to a predictive trajectory including specific gene expression, activate enhancer(s) and dysregulated gene modules across different layers. For example, this latter use, of course, enables us to better localize the specific activities at the molecular level happening from genotypes to associated phenotypes such as psychiatric disorders.