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

The brain is the most complex organ in adult human, playing a commander role for the human body. [[need more and Jim's help to lead off]].

A variety of genomic elements have been found to associate with psychiatric behaviors such as ones in mental diseases. For example, the Psychiatric Genomics Consortium (PGC) identified a wide variety of genomic variants including SNPs and CNVs associated with psychiatric disorders; e.g., 108 GWAS loci associated with schizophrenia (SCZ) and >90k SNPs imputed from GWAS 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. For example, the differentially expressed genes were found for SCZ (n=xxx), BP (n=xxx), ASD (n=xxx). In another context, to systematically discover the gene expression and regulation patterns in brain, recent large consortia like GTEx, ENCODE and Epigenomics

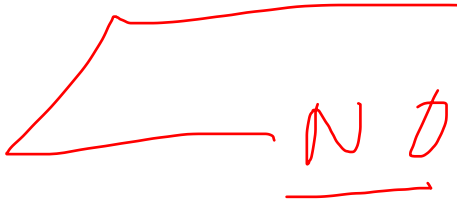
Deleted: brain phenotypes including

Deleted: GWAS has identified XXX SNPs significantly associated with SCZ. Also

Deleted: for

Deleted: Therefore,

Deleted: studies



Roadmap have generated large-scale RNA-seq and ChIP-seq data for brain tissues and cell lines, trying to systematically detect the brain specific genes, transcripts and regulatory elements. However, these studies only focused on the normal healthy brains, so their data is unable to find the specific genomic elements for additional phenotypes especially for mental health. The CommonMind Consortium and others thus have provided the gene expression and genotyping data for both healthy and schizophrenia samples (N=279 vs. 258) and identified ~693 differentially expressed genes in schizophrenia, but their results still suggested needing thousands of samples to achieve the statistical power of 0.8 for detecting differentially expression of eQTL-associated genes [refs]. Therefore, given that the complexity of brain samples of mental diseases, we need a variety of additional samples to gain more statistical powers for discovering a complete set of genomic elements for mental diseases or other phenotypes. Moreover, individual molecules are not independent to affect brain. Instead, they interact with each other forming a network. Thus, effort is needed to model and analyze the molecular interactions that drive the phenotypes of adult brain including psychiatric 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 (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 brain tissue samples with different phenotypes. The rich data generated by the PsychENCODE Consortium are an integrative resource for studying regulatory mechanisms in the human brain [1], such as for major psychiatric disorders, age, gender, etc. PsychENCODE datasets have been assembled by many investigators over several years, and they are housed in a central depository (xxxx) and shared with the public. Integration of these multi-dimensional and large-scale datasets potentially benefits understanding the molecular mechanisms for adult brain, which however still remains challenge.

To address this challenge, we integrated the ~~PsychENCODE~~ and relate datasets over all ~2000 samples, compared them against various phenotypes and diseases in adult brain, and developed a comprehensive and online available resource for the adult brain. The core datasets in PsychENCODE are a large amount of functional genomic and genotype data with phenotype information relating to the brain. However, to maximize the data size, we also uniformly reprocessed and interconnected data with other related genomic resources to develop such a comprehensive resource for the brain functional molecules across genomic, transcriptomic, epigenomic and regulatomic levels. In particular, these other resources include ENCODE, CommonMind, GTEx, Epigenomics Roadmap, etc. Also, we uniformly processed and integrated the recent single cell RNA-seq data for both neuronal and non-neuronal cell types to find gene expression signatures and calculate the corresponding cell fractions of individual tissues. Thus, this resource comprises 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. We then analyzed this resource and found the specific

CHROM.
SS.

JIM

Deleted:

Deleted: disease brains such as SCZ for xxx

Deleted: [refs]

Deleted: .

Deleted: G

Deleted: the adult brain phenotypes are highly likely driven by interactions among various molecules, rather than individual molecules.

Deleted: brain

Deleted: Also

Deleted: preeminent

Deleted: brain phenotypes

Deleted: brain

Deleted: . to

Deleted: a

genomic and transcriptomic activities on genome wide in brain including gene expression, non-coding transcription and enhancers. Finally, we developed and built an integrative model to reveal how the interactions among genomic variants, gene expression, enhancers and phenotypes, trying to explain the molecular mechanisms from genotypes to brain phenotypes.

Comprehensive resource for adult brain functional genomics

The PsychENCODE consortium has generated and assembled a robust large-scale dataset on the adult human brain, including genotyping, RNA-seq, ChIP-seq and single-cell transcriptomic data on ~2000 individual brain tissues with different phenotypes including mental diseases (Assay summary in Methods). To harmonize and integrate the datasets across multiple consortia, we uniformly processed these datasets using common standard pipelines (Methods). In particular, we adopted the ENCODE standard such as RNA-seq and ChIP-seq data processing pipelines for PsychENCODE and reprocessed all other major datasets from other resources using this standard. This step uniformly processed cross-resource functional genomic data and enables the comparisons across phenotypes such as brain data from PsychENCODE vs. other tissues from GTEx and Epigenomics Roadmap. We also used the GTEx eQTL pipeline to uniformly process the PsychENCODE genotype data and find the QTLs for various brain genomic activities in the resource. All these uniformly processed datasets are also available in our resource. 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.

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.

At the bottom, we have the large scale raw data and the phenotype information for ~2000 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:

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).

ATAC & HIC

SS UNI.

Formatted: Heading 2, Line spacing: multiple 1.38 li

Deleted: . ... [1]

Deleted: and can be used as

Deleted: tool

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.

System identification of the QTLs associated with adult brain transcriptomics

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 activities. 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) and splicing QTLs (sQTLs). We used the GTEx 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: >2 million expression QTLs (eQTLs) with ~11k eGenes, >5 thousand chromatin QTLs (cQTLs) for histone modification signals, and xxx splicing QTLs for alternative splicing patterns.

GIVEN THE LARGE #

We also 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). For example, we found that these variants cover a larger fraction of disease-associated brain GWAS SNPs than 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). We also evaluated the overlap of eQTLs with cQTLs and found that XX% of cQTLs are overlapped with eQTLs. The SNPs in cis-eQTL list(Cis-eSNPs) were enriched within XXXX, and depleted XXXXXX (Fig. X). We examined the enrichment of most significant eQTLs per gene in Roadmap Epigenomics Consortium and ENCODE enhancers across XX human tissues and cell lines. Cis-eQTL were enriched for enhancer sequences present in brain tissues and the strongest enrichment is observed in DLPFC enhancers. We also calculate the enrichment of cis-QTLs on GWAS SNPs of brain related disorders (schizophrenia, bipolar disorders and parkinson's disease) and non-brain related disorders (CAD, asthma and type 2 diabetes). Cis-QTLs have more significant

GTEx
STD
?

Moved (insertion) [1]

Deleted: In addition, we link the QTLs that overlap the enhancers and promoters in the resource to reveal the potential regulatory activities such as QTLs break TF binding sites.

Moved (insertion) [3]

Deleted: Our resource includes a great number of regulatory variants significantly associated with brain transcriptional and epigenomic activity: >5 million expression QTL for gene expression, >5 thousand chromatin QTL for histone modification signals, and xxx splicing QTLs for alternative splicing patterns.

Deleted: ,).

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).

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

In addition to identifying the genomic variants such as QTLs for adult brain, at the middle of resource, we also 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 summarize a list of xxx PsychENCODE brain enhancers, mainly active on DLPFC and CBC (Supplement).

Differentially expressed genes, transcripts and brain splicing patterns - we compared the expression changes from uniformly processed RNA-seq data across PsychENCODE-relate studies, ENCODE and GTEx, and found the xxx genes and transcripts that express in brain samples,and xxx eGenes associated with eQTLs (Methods). We also discovered xxx non-coding RNAs and novel transcribed regions in brain. Also, we compared them against all possible phenotypes and derived the phenotype-specific genes and transcripts. In addition, we calculated the alternative splicing patterns at the transcript level; i.e., the percentage of the transcript abundance over its gene abundance, and found the brain-specific spliced transcripts. We finally created a map linking all these brain differentially expressed and spliced genes and transcripts with corresponding phenotypes such as disorders and developmental stages.

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. 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. The activations of various wires may change across

Deleted:

Formatted: Font:16 pt

Formatted: Font:16 pt

Deleted:

Formatted: Font:16 pt

Deleted: A

Deleted: used the

Deleted: the

Deleted: of QTLs,

Deleted:

Deleted: *QTLs* - 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 expression QTLs, chromatin QTLs and splicing QTLs. In particular, we used the GTEx 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: >xxx million expression QTLs (eQTLs), >5 thousand chromatin QTLs (cQTLs) for histone modification signals, and xxx splicing QTLs for alternative splicing patterns. We compared our QTLs with known disease associated SNPs such as GWAS and found that the number of eQTLs in this resource is significantly greater than previous studies, approaching the saturation of human mutations (Figure xxx). We also showed that the eQTLs number can be predicted from the sample size using a fitted curve (Figure xxx).

Moved (insertion) [2]

phenotypes. Thus, using these “wiring” regulatory relationships, we inferred the gene regulatory networks that identify the regulatory connectivities 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 that brain has specific genomic and transcriptomic activity

This comprehensive resource allows us to discover the specific functional genomic elements that relate the brain phenotypes as above. Thus, we leveraged this resource against various phenotypes and compared with other tissue types to reveal the unique brain genomic activities, particularly relating to transcriptomic and regulatory binding activities such as RNA abundances and open chromatins. In particular, we first performed the spectral analysis for comparing the similarities of gene expression other tissue samples from GTEx (Figure xxx), 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. This suggests that there exist the brain has unique and distinct gene expression programs, involved by the brain elements including differentially expressed genes and non-coding RNAs in our resource that make brain very different from other tissues. In addition, this major brain cluster has a particular geometric pattern showing that the normal samples (e.g., GTEx) form an inner core, and the disease ones expand this core to the entire cluster. This suggests that the psychiatric diseases still have different expression programs from the normal, though even more distant from other organs. 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

Moved up [2]: The activations of various wires may change across phenotypes.

Moved up [1]: In addition, we link the QTLs that overlap the enhancers and promoters in the resource to reveal the potential regulatory activities such as QTLs break TF binding sites.

Deleted: that drive the brain phenotypes

Deleted: to

Deleted: eventually

Deleted: elements

Deleted: these brain

Deleted: affect

Deleted: System identification of

Deleted: s

Deleted: via comparative analysis

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. 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.

Deleted: .

Deleted: It is shadowed, for

Deleted: by highly diverse

Our comparative analysis reveals that the brain is different from other organs in either 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 Table XXX) for the resource. For example, XXX genes have been found to differentially express between male and female samples. We also checked the enriched pathways and functions among the SCZ genes, and indeed found that many are relating to male. Moreover, we also found that these brain dex genes are significantly less/greater than DEX genes for other tissues in GTEX (p<xxx), which suggesting that the brain expression uniqueness is highly driven by a small/large set of genes. As previously described, 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 xxx, Supplement), which suggests that there exist brain specific regulatory mechanisms drive these brain co-expression modules.

Deleted: SCZ

Deleted: healthy

Deleted: SCZ

Deleted: [[SCZ is replaceable by other features]]

Deleted: W

As shown above, the brain samples have different chromatin and gene co-expression activities from other organs, implying that the brain also has specific gene regulatory activities. Therefore, we are further interested to compare the regulatory regions between brain and other tissues to see any brain specific regulatory activities. We integrated the chromatin data 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 somewhat similar patterns with the gene expression comparison; e.g., the brain samples can also cluster together in terms of active enhancer similarity (Figure xxx). This result suggests that the brain has specific and distinct epigenomic activities as well, involving the brain active enhancers from our resource. More importantly, the brain active enhancers or gene expression patterns are intermediate phenotypes, potentially driven by particular

regulatory variants, so we are further interested to find such variants highly associated with gene expression and enhancer signal changes across brain samples.

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

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 the expression changes of brain tissue genes across various cell types including both neuronal and non-neuronal. We also 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 adult brain phenotypes at the tissue level can more easily be explained by the changes of cell fractions.

Specifically, we integrated and used the same pipeline to uniformly process the single cell RNA-seq data for ~800 cells from PsychENCOCE, ~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 OPC, and xxx novel cell types in embryonic and fetal tissues. We then 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 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.

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]. Therefore, we deconvolved the tissue-level gene expression data of all 2000 samples using single-cell gene expression data of xxx biomarker genes to find the proportions of different cell types corresponding, and compare cell fractions across different phenotypes (Y=WX, Methods). The single cells used in deconvolution cover all 16 neuronal types, five non-neuronal types and xxx additional PsychENCODE types. For example, it is very interesting that we can explain much (R2=~80%) of the individual variation in gene expression of both male and female samples in terms of changing proportions of basic cell types, rather than changes in individual genes (Figure xxx covariance). In addition, we found that the cell fractions of individuals vary across

Moved up [3]: Our resource includes a great number of regulatory variants significantly associated with brain transcriptional and epigenomic activity: >5 million expression QTL for gene expression, >5 thousand chromatin QTL for histone modification signals, and xxx splicing QTLs for alternative splicing patterns. We also 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). For example, we found that these variants cover a larger fraction of disease-associated brain GWAS SNPs than any previous analyses, suggesting potential molecular targets for these associations (xx% for SCZ, xx% for BP, ASD,). We also evaluated the overlap of eQTLs with cQTLs and found that XX% of cQTLs are overlapped with eQTLs. The SNPs in cis-eQTL list(Cis-eSNPs) were enriched within XXXX, and depleted XXXXXX (Fig. X). We examined the enrichment of most significant eQTLs per gene in Roadmap Epigenomics Consortium and ENCODE enhancers across XX human tissues and cell lines. Cis-eQTL were enriched for enhancer sequences present in brain tissues and the strongest enrichment is observed in DLPFC enhancers. We also calculate the enrichment of cis-QTLs on GWAS SNPs of brain related disorders (schizophrenia, bipolar disorders and parkinson's disease) and non-brain related disorders (CAD, asthma and type 2 diabetes). Cis-QTLs have more significant enrichment for GWAS SNPs of brain related disorders than the ones of non-brain related disorders.

Deleted: **[[should we move up??]]**

Deleted: 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

Deleted: populations

Deleted: populations

Deleted: First

Deleted: from

Deleted: W

Deleted: these

Deleted: s for the brain genes in the resource

Deleted: brain

Deleted: at the tissue level

Deleted: In contrast, we also found that a number of brain genes don't show expression changes across cell types, which implies that their expression changes at the tissue level are potentially explained by the cell proportions, rather than individual genes.

Deleted: 4

Deleted: we used the heatmaps (Figure xxx) to display

Deleted: populations

different phenotypes (Figure xxx), and a number of cell population changes highly associate with brain phenotypes. For example, the fraction(s) of neuronal type(s) (Inhibitory X) is significantly anti-correlated with Age (r = xxx). The non-neuronal cell populations increase significantly in SCZ (or Male) samples (p<xxx) while the neuronal cells decreasing. 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). This suggests that these SNPs 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 interactions between genotypes and phenotypes is a very complex process experiencing multiple intermediate stages including gene expression, signaling, modulation and so on. Thus, to understand the entire processes how genotypes and phenotypes affect to each other, we built an integrative model to understand how the brain genomic variants affect gene expression and regulation, and eventually predict the phenotypes (Figure xxx). This model integrated all high dimensional functional data types in this resource including genomics, transcriptomics, epigenetics and regulatomics, and genotype-phenotype relationships, and also allowed us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only. Specifically, we named this model, a Deep Boltzmann Machine-based eTWAS model that directly embeds regulatory network information to predict genotype-phenotype associations. 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

Deleted: . We found that there exist

Deleted: that

Deleted: brain

Deleted: called

Deleted: feed-forward

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.

Specifically, we built this model as follows. We first inferred the gene regulatory networks that identify the regulatory connectivities 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 that combines lasso and ridge regressions to predict the target gene expression from genotypes of various QTLs, the chromatin stages of enhancers, splicing patterns and TFs gene expression using the resource samples, 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.

EARLIER

We then connected the nodes on Layer 2 of our model to follow the inferred gene regulatory network structures. 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 connecting to the genes enriched with ZZZ pathways (p<xxxx). 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 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). 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 trait of schizophrenia (SCZ) is activated by two modules, x, and y corresponding to dopamine-related pathways and complement pathways, respectively. Each module is connected by a set of genes including C4 genes, which are regulated by corresponding QTLs and enhancers as shown in blowup gene regulatory mechanism. For each phenotype, we also provide a list of such eTWAS pathways on resource websites. In addition, this model also allows us to quantitatively impute the missing transcriptional and epigenetic information by inputting given genotype data only. Moreover, we also make the model available as a set of distributive software from the resource. Finally, the model is made available as a set of simplified files, where one can explicitly see the correlations being used at various stages.

Deleted: Autism

Deleted: neuronal cell fractions

Discussion

We integrated the genomic, transcriptomic and regulatomic PsychENCODE datasets from ~2000 samples and developed this comprehensive resource consisting of various functional genomic elements for the adult brain. Developing this resource and integrated model to a population-level scale serves as an important step in gaining meaningful biological insights from functional genomics studies in neuroscience. In particular, we compared it with other tissues such as GTEx data and identified the genotypes and QTLs, the specific expressed genes, transcripts and noncoding RNAs, active chromatin regions, the regulatory networks that significantly relate with different brain phenotypes at both cellular and tissue levels. For example, the QTLs allow one to potentially interpret most of the known brain-associated GWAS SNPs in terms of perturbations to specific genes. Thus, the neuroscientist can use this resource as a reference to compare with their data, generate hypotheses and help design experimental validations. In addition, this resource is publicly available online and can be extendable and scalable to integrate additional data types and phenotypes. For example, it can add the individual's fMRI image features measuring functional neuro-connectivity, and use our model to identify the genotypes that associated with image features such as image-QTLs (iQTLs) [xx]. Also, our resource can incorporate with the neurodegenerative diseases like Alzheimer or developmental stages.

Deleted: integrate

Deleted: clinical data like

Deleted: s

Deleted: connectivties

Deleted: functional genomic elements

Deleted: for

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.

Deleted: .

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

To systematically understand the molecular functions and mechanisms how genomic variants affect associated phenotypes in the brain, we need to find the related molecules that have specific activities for the brain phenotypes. Therefore, t