Comprehensive resource and integrative model for functional genomics of the adult brain (~5300)

Abstract (192)

Understanding how genomic variation influences adult brain phenotypes and disorders remains a key challenge. To this end, the PsychENCODE consortium has generated large-scale datasets on the adult human brain, including genotyping, RNA-seq, ChIP-seq, ATAC-seq, HiC and single-cell data on healthy and diseased brain tissues of thousands of adult individuals with different phenotypes. Using this data, we developed a comprehensive resource on functional genomics of adult brain including a variety of QTLs for expression and chromatin, and active enhancers. Leveraging the single cell data, we deconvolved tissue-level gene expression to find the cell fraction changes along with associated QTLs for various phenotypes. Comparing this resource with others using spectral analysis, we show that the brain has unique expression and greater non-coding transcription than most other tissues. Moreover, we integrated the Hi-C and regulatory data to predict the gene regulatory network linking all possible functional genomic elements including QTLs, regulatory factors and target genes. Based on this, we developed a deep-learning model, significantly outperforming previous methods to predict genotype-phenotype associations and highlight intermediate genes and functional modules, revealing potential mechanisms, and enable quantitatively imputation of missing transcriptional and epigenetic information from genotype data only.

I. Introduction (495)

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

To this end, a number of genomic studies have recently focused on discovering genomic functions relating to the phenotypes in adult brain. A variety of genomic elements and variants have been found to be associated with brain and psychiatric disorders (for instance, the Psychiatric Genomics Consortia (PGC) identified 142 GWAS loci associated with schizophrenia). Large consortia have also identified the reference sets of genomic elements across the entire body; e.g., eQTLs and eGenes in GTEx, and enhancers from ENCODE and Epigenomics Roadmap that are associated with various human cells and tissues. Though some of these elements relate to the brain, none of the consortia have specifically tailored their efforts toward comprehensively identifying the functional elements in the brain.

To address this gap, recent technologies have started to detect the specific molecular activities [[dg: define gene regulatory mechanisms?]] within the brain. Recent HiC and ATAC-seq studies have been used to identify specific chromatin structural and regulatory elements, such as brain-active enhancers. Single-cell sequencing techniques offer great promise for studying the

Style Definition: Normal: Font: (Default) Times New Roman, (Asian) Times New Roman

Formatted: Font color: Text 1

Deleted: In

Deleted:]], little progress has yet been made in our fundamental understanding of the molecular causes of psychiatric disorders.

Deleted: .

Deleted: governing

Deleted: as they directly relate to the

Deleted: 108

Deleted: these

Deleted: within the brain. Particularly, single-cell sequencing techniques offer great promise for studying the transcriptome. Also, recent

transcriptome. However, each of the studies that leverage such technologies have generally focused on individual aspects of brain functional genomics (CommonMind) or have been a small component of broad surveys (ENCODE, Epigenomics Roadmap). These data have not yet been fully integrated at scale, so as to comprehensively understand the basis of psychiatric disease.

[[[discuss trans]]] Larger sample sizes and more comprehensive data are warranted to obtain a fuller view of brain-relevant functional genomics [[refs]]. To this end, the PsychENCODE Consortium has generated and assembled a large-scale dataset on the adult human brain, including data derived through genotyping, RNA-seq, ChIP-seq, ATAC-seq, HiC and single-cell analysis on high-quality brain tissue from both healthy and diseased samples of thousands of adult individuals with different phenotypes. We have built a central, publically available comprehensive resource (http://adult.psychencode.org/) for adult brain functional genomics, including all the raw and uniformly processed data at both tissue and single cell levels from PsychENCODE and other related projects, including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, in addition to single-cell data [refs] with up to X,XXX samples. By leveraging this resource, our analyses identified various functional genomic elements and quantitative trait loci (QTLs) specific to the adult brain. We also combined these elements and built an integrated deep-learning model to impute missing data. The results obtained from this model are then studied in relation to specific brain phenotypes and psychiatric disorders.

II. Comprehensive resource for adult brain functional genomics (224)

We designed this comprehensive resource to provide a coherent data structure. Broadly, it organizes a large amount of data for brain functional genomics pyramidally, with a large base of raw data files (much of these exist as restricted-access data, such as individual genotyping and raw next-generation sequencing data of transcriptomics and epigenomics), a middle layer of uniformly processed and shareable results (such as open chromatin peaks and gene expression quantifications), and a compact cap at the top, consisting of an integrative model (based on imputed regulatory networks and QTLs). As shown in Fig 1, to build the base layer, we included all the datasets from PsychENCODE related to the adult brain and merged these datasets with other relevant data from additional project,s including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, and recent brain single cell studies. In total, this resource constitutes XXXX data samples derived from 1931 individual adult brains from multiple cohorts, which covers a large representation of brain phenotypes and psychiatric disorders. The major data types include genotyping, RNA-seq, ChIP-seq, ATAC-seq, HiC and single-cell data (this required large-scale imputation for all the PsychENCODE datasets, and we make full genotype sets available). Furthermore, the PsychENCODE project developed a specific "reference brain" project utilizing many assays on the same set of brain tissues, which we used to develop an anchoring annotation for the entire resource (Supplement).

III. Bulk and single cell transcriptome analysis and deconvolution explain gene expression and cell fraction changes (813)

To identify the genomic elements that exhibit transcriptional activities in the specific to the adult brain, we used the ENCODE standard pipeline to uniformly process the RNA-seq data of all available samples from PsychENCODE and GTEx. Using these data, we identified more interpretable functional elements, such as sets of differentially expressed and co-expressed

Deleted: . They

Deleted: Previous efforts, such as CommonMind, have included integrative analyses to identify brain-specific genomic elements. Though promising, larger

Deleted: still needed

Deleted: specific

Deleted: thus

Deleted: We are interested in identifying

Deleted: . In particular

Deleted: RNA-seq

genes characterizing various brain regions, phenotypes and disorders [[ref:capstone1]]; these are provided as part of our resource. Moreover, we constructed a gene co-expression network using the samples across brain and other tissues, and we clustered this into a number of gene co-expression modules and submodules (representing clusters at multiple resolutions, see Supplement), many of which reveal the expression patterns specific to brain samples.

Brain tissues has been found to comprise a variety of cell types, including neuronal and non-neuronal cells such as astrocytes. One issue with measuring gene expression changes over a population in our brain tissue samples is reliably determining whether the changes are driven by gene expression in a particular cell type or whether they are driven by changes in relative proportions of various cell-types. To address this, we integrated the single cell transcriptome data to discover how the gene expression from various cell types contribute to bulk gene expression using two strategies.

First, we used the standard pipeline to uniformly process single cell RNA-seq data in PsychENCODE, in conjunction with a number of other single-cell studies on the brain, in order to assemble a list of cell types in the brain (i.e., 16 neuronal types, 5 non-neuronal types and 4 additional fetal-related types from PsychENCODE; see Supplement). This list constitutes a matrix (C) of the gene expression signatures of 25 basic or expression-clustered cell types, which are mostly concordant with what has been published, with some minor modifications (Figure Sxxx). Across these cell types, we found that the number of genes whose expression levels vary much more substantially than they do amongst individual tissues; e.g., the dopamine receptor genes (DRD) that associate with SCZ (Figure xxx). This implies that the gene expression variation of cell types can give rise to substantial changes in bulk gene expression at the tissue level, [[dq]]which has been demonstrated by various forms in healthy tissue [[Oldham et al. nat neuro 2008; PMC5325728, etc…]], as well as in brain disease [[Voineagu et al. 2011, PMC3706780; DOI: 10.1126/science.aad6469]].

To explore this further, we performed an unsupervised analysis for the bulk tissue expression data to identify the primary components as they relate to different single cell types. We decomposed the bulk gene expression matrix (B) from our resource using non-negative matrix factorization (NMF, see Methods), and we then determined whether the top components (TCs) of the NMF that capture a variety of data covariance (a.k.a., NMF-TCs) and the 25 reference gene expression signatures of single cells are consistent. As shown in Figure XX, we found a number of NMF-PCs that are highly correlated with the gene expression signatures of neuronal, non-neuronal and fetal cell types. This demonstrates that an unsupervised analysis derived from the main components of the bulk tissue data roughly matches the single cell data, partially corroborating the 25 cell types.

As shown in Figure xx, we then de-convolved the bulk tissue expression matrix B using the single cell data matrix C to estimate the cell fractions W, by solving the equation “B=WC” (See methods). We found that multiplying estimated cell fractions and single cell expression data can explain much of the expression variation at the population level (i.e., across tissue samples). Specifically, $1-||B-WC||^2/||B||^2>0.85$, where $||.||$ is the Frobenius norm of matrix (see Methods). This shows that over 80% of the bulk gene expression variation across samples can be explained by variation in the proportions of basic cell types. Moreover, we found that our estimated fractions of NEU+/- cells match the experimental measurements for reference brain samples (r=xxx, Figure xxx).

Furthermore, we found considerable variation in the cell fractions in different individuals (i.e.,

Deleted: network

Deleted: one

Deleted: heterogeneous

Deleted: over

Deleted: reference

Deleted: type signatures. These signatures

Deleted: previously

Deleted: in terms of cell clusters based on their gene expression similarities

Deleted: .

Deleted: Second

Deleted: in order

Deleted: In particular, we

Deleted: PCs

Deleted: and

Deleted: validates these

Deleted: single

Deleted: of individual tissues

deconvolution coefficients from W), and cell fraction changes were found to be highly associated with different phenotypes and psychiatric disorders (Figure xxx, Supplement for complete individual cell population estimates). For example, the excitatory and inhibitory neurons (Ex3 and In6) exhibit significantly different fractions between healthy male and female samples. The fraction of Ex3 cell types are also significantly reduced in ASD samples (p<xxx), while non-neuronal cells (e.g., oligodendrocytes) are represented in much greater abundance. Another interesting association we found was that cell fractions change with age. In particular, the fractions of neuronal type(s) (Ex3 and Ex4) are significantly positively correlated with age (r = xxx), but non-neuronal types (oligodendrocytes) are found to be negatively correlated with age. Furthermore, these age-related cell fraction changes are also potentially associated with differentially expressed genes across age groups (Figure xxx). For example, the gene involved in early growth response is down-regulated in older age groups, whereas the gene ceruloplasmin is down-regulated among middle-aged groups.

IV. Active enhancers in adult brain (215)

In addition to the transcriptome data, the uniformly processed chromatin data in the resource gave rise to uniform quantifications, peak calling lists and single tracks for adult brain epigenomics. Thus, we developed a consistent set of brain active enhancers. In particular, we used the H3K27ac and the H3K4me3 ChIP-seq data, together with the ATAC-seq data from the reference brain sample. [[move to Suppl?]]The ATAC-seq peaks from the reference brain helps to identify the open chromatin regions. For a region to be considered as an active enhancer, it needs to contain both ATAC-seq signal and H3K27ac signal (Z-score > 1.64). We also exclude the regions within 2kb of any TSS that has H3K4me3 signals as they are most likely to be promoters. Moreover, we find that most of the reference brain enhancers overlap the Roadmap enhancers from Roadmap. Totally, we identified ~80K active enhancers in the DLPFC. Similarly, we have also developed reference sets in additional brain regions including CBC and ACC.

We then looked at the variations of the epigenetic signals across individuals at these enhancers. We found that ~10% of the identified enhancers appear to be active in all control samples, and around 68% are active in more than half of the population. This suggests that enhancer activity varies across individuals, yet the majority of brain enhancers are active in most of the population. We also compared the distribution of the saturation curve on the normal samples (n=50) and the ASD samples (n=43). The ASD samples seem to have slightly more variation in terms of the enhancer distribution across individuals, yet a two-sided K-S test does not suggest a significant distribution among these two groups.

V. Consistently comparative analysis reveals the brain related transcriptomic and epigenomic activity (475)

One key aspect of our analysis is that we uniformly processed the transcriptomic and epigenomic data across PsychENCODE, ENCODE, GTEx and Roadmap. This allows us to compare the brain to other organs in a consistent fashion in order to delineate gene expression and chromatin activities that are unique to the brain. Moreover, we attempted several methods for an appropriate comparison.

Principal component analysis (PCA) and t-SNEs are two popular techniques for performing comparisons. The former tends to capture only global structures, ignoring most of the local structure, but it can easily be influenced by outliers. On the other hand, t-SNE analysis

Deleted: a number of

Deleted: population

Deleted: The

Formatted: Space After: 0 pt

Deleted: We used these data and derived further simplified epigenomic representations for the adult brain. In particular, we first

Deleted: We processed

Deleted: and

Deleted: of

Deleted: using the standard ENCODE ChIP-seq processing pipeline. We then identified an overall set of brain enhancers based on these experimental data of the same reference

Deleted: using the ENCODE 3 candidates regulatory element (cRE) pipeline.

Formatted: Highlight

Deleted: indicate the

Formatted: Highlight

Deleted: in the brain, and the histone marks, together with their distances to ... [1]

Deleted: (Moore et al, in review). ... [2]

Formatted: Highlight

Formatted: Highlight

Deleted: peaks to consistently find ... [3]

Formatted: Highlight

Deleted: developed

Deleted: sets in additional brain ... [4]

Formatted ... [5]

Deleted: uniform processing

Deleted: This comparison could not be ... [6]

preserves local structure but “shatters” global structure. For example, t-SNE tends to separate samples from the same tissue so that the cluster distances on t-SNE space are not proportional to real gene expression dissimilarities. It thus does not give a sense of overall effects. We thus found another technique that is capable of capturing local structure while maintaining meaningful distances in global structure space. Reference Component Analysis (RCA) projects the gene expression in individual sample against a reference panel, and then essentially reduces dimensionality of individual projections. [[check]] In fact, we did RCA consistently for comparing brain and other tissues in terms of their similarities of both the transcriptome (RNA-seq gene expression) and the epigenome (ChIP-seq signals on our consistent set of enhancers).

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

Our RCA analysis focuses on inter-tissue differences in well-annotated regions (i.e. genes, promoters and enhancers). In addition to the expression differences in protein-coding genes, a tremendous amount of transcriptional diversity is present across tissues in intergenic and noncoding regions. Thus, we looked at the overall level of transcriptional diversity across tissues. For protein-coding regions, it has previously been demonstrated that testes and lung tend to have the largest transcriptional diversity in terms of the percentage of transcribed regions (Figure SYYY sat’d for genes). However, when we shift to non-coding and unannotated regions, we find that brain tissues (such as cortex and cerebellum) do, to some degree, stand out by exhibiting greater transcription than most other tissues. This transcriptional diversity tends to increase with the number of samples (Figure xxx sat’d).

VI. QTL analysis (568)

To understand how the genotype affects the transcriptome and epigenome in the adult brain, we used the PsychENCODE resource data to identify quantitative trait loci (QTLs) affecting gene expression and chromatin activity. In particular, we calculated the association of SNPs with normalized gene expression and chromatin state (Methods) to find quantitative trait loci associating with gene expression and epigenomic activity in adult brain, including several major categories: expression QTLs (eQTLs), chromatin QTLs (cQTLs), splicing QTLs (sQTLs) and cell fraction QTLs. For the eQTLs, we adopted a standard approach, adhering closely to the established GTEX eQTL pipeline. We identified ~2M eQTLs and ~17000 e-genes in the DLPFC. This conservative estimate is a substantially larger number of eQTLs than previous brain eQTL studies and reflects the very large sample size and statistical power we have. We believe this is close to saturation, in terms of associating almost every variant with some expression modulating characteristic. We also applied the same QTL calculation pipeline to calculate sQTLs and identified ~590k sQTLs.

For the cQTLs, the situation is more complicated. There are no established standard methods for calculating these on a large scale. To properly identify them, we focused on a reference set of enhancers to define the region associated with the activity of the chromatin and then looked

Deleted: region

Deleted: cQTLs

at how this activity varies in these enhancers across individuals, correlating this with nearby variants. (See methods). Overall, we were able to identify ~2000 cQTLs in addition to [[Crawford]].

Furthermore, we were interested to see if any SNVs were associated with the single cell fractions. In particular, we used our QTL pipeline to identify 443 distinct SNVs whose genotypes are significantly associated with differential cell fractions across individuals; i.e., cell fraction QTLs (fQTLs). In total, the 443 distinct SNVs constitute 508 different fQTLs between different cell types. Significant fQTLs are those with associated Bonferroni-corrected p-values of no more than 0.05. Different cell types exhibit a great deal of heterogeneity in terms of their abundance within the set of high-confidence fQTLs. For instance, we identified 45, 15, and 33 significant fQTLs associated with the endothelial cells, astrocytes, and microglia, respectively, but there were no significant fQTLs that were found to be associated with oligodendrocytes. Moreover, we also identified XXX SNPs significantly associated with the gene expression changes across individual tissues unexplained by our single cell deconvolution; i.e., B-W*C(Methods).

Given the QTLs we identified, we overlapped and annotate them with a variety of different genomic annotations and look at the degree to which they overlapped. The distributions of detailed QTL annotations across genomic regions are shown in Figure xxx. For example, we observed a significantly number of predictive QTLs break the TFBSs on the enhancers or promoters (xx%, Figure xxx), and also found xxx e-promoters on which eQTLs lie associating with distal genes. As expected, there is a very large amount of overlap between the cQTLs, sQTLs, and eQTLs, and with ~50% of the cQTLs also being eQTLs. We calculated the enrichment in cis-QTLs of 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 disorders than non-brain disorder GWAS SNPs. Collectively, these QTLs annotate a larger fraction of GWAS SNPs involving the brain (e.g., 21% in schizophrenia, 18% in bipolar) than previously observed, providing important leads on which genes are affected in disease.

VII. Gene regulatory networks in adult brain (905)

In this section, we provided an integrative analysis at the gene regulation level for the data and genomic elements in the resource and predicted a gene regulatory network revealing how the genotype and regulators control target gene expression in adult brain.

To this end, we first process a full Hi-C dataset for adult brain, which provides direct physical evidence for potential interactions between enhancers and promoters (Figure 5A). Specifically, we generated and processed the Hi-C data for the same reference adult brain that was used to identify the brain active enhancers, as previously described (PMID: 27760116, Supplement). In total, we identified 2,735 topologically associated domains (TADs) which set the physical boundaries of enhancer-promoter interactions and then 149,097 putative enhancer-promoter interactions in the adult DLPFC. As expected, we found that ~75% of enhancer-promoter interactions occur in the same TADs (Figure 5xx), suggesting that TADs provide physical boundaries for cis-regulatory relationships between enhancers and target genes. Promoters tend to interact with other regulatory elements such as other promoters and enhancers (Figure xxx), the type and number of which affect the target gene expression levels (Figure 5xx, Figure xxx).

We next integrated the Hi-C dataset with eQTLs to assess how much of the common variation-associated gene regulation is mediated by chromatin interactions. Interestingly, 30.7% of e-

Deleted: .

Deleted: ¶

Deleted: our

Deleted: step is to

Deleted: in the format of topologically associated domains (TADs)

Deleted: using the protocol of XXX (

Deleted: xxx

Deleted: in adult brain. Overall, this data showed a number of established properties, such as that gene expression tends to increase with increasing number of interactive enhancers (Figure 5xx). More importantly, we found that >xx%

Deleted: happen

Deleted: in the adult brain

Deleted: potentially

Deleted: a large number of

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

Intrigued by the regulatory map built upon Hi-C and eQTLs, we exploited these two key datasets to identify putative target genes of newly identified 142 schizophrenia GWS loci [[ref:clozuk]]. We categorized 5,996 putative causal (credible) SNPs reported in the original study into promoter/exonic and intergenic/intronic SNPs. Promoter/exonic SNPs were directly assigned to the target genes based on the genomic coordinates, while intergenic/intronic SNPs were annotated based on chromatin interactions, which led to the mapping of 92 loci into 377 genes. Credible SNPs colocalize with 2,029 eQTLs associated with 83 e-genes, 43 of which overlap with those identified by the Hi-C driven approach. To confirm that this overlap is mediated by the shared causal variants in GWAS and eQTLs, we performed a colocalization test (PMID: 24830394), from which we identified 190 genes across 79 loci in which GWAS and eQTLs share common causal variants. In total, we identified 488 putative schizophrenia-associated genes, hereby referred as SCZ genes, and 99 genes that show evidence both at the level of Hi-C and eQTLs, providing a high-confidence gene list (Figure 5xx). This is a huge increase from the previously annotated 22 genes across 19 loci based on CMC adult brain eQTLs [[ref:clozuk, PMID: 27668389]], mainly due to the dramatic increase in power of eQTLs. The majority of SCZ genes (288 genes, ~59%) were not in linkage disequilibrium (LD, r^2>0.6) with index SNPs (Figure xxx), consistent with the previous observations that regulatory relationships often do not follow linear genome organization.

Intriguingly, SCZ genes were enriched for genes and co-expression modules dysregulated in DLPFC of schizophrenia-affected individuals (PMID: 27668389), suggesting that common variation-mediated gene regulation contributes to the gene dysregulation in schizophrenia (Figure 5xx). SCZ genes are often affected by recurrent CNVs in schizophrenia, hinting the shared genetic etiology between common and structural variation. We also recapitulated the key findings that SCZ genes were enriched with loss-of-function mutation intolerant genes [[ref:clozuk]]. Functional analyses showed that SCZ genes were enriched for translational regulators, cholinergic receptors, calcium channels, and synaptic genes (Figure xxx). We also leveraged a single-cell expression atlas to examine cell-type specific expression signatures of SCZ genes. RESULT. Collectively, these results demonstrate the strength of an integrative resource for providing rich biological insights into psychiatric disorders.

As a second step to build the gene regulatory network, we integrated the TADs with other regulatory elements and relationships such as the enhancers, transcription factors (TFs), miRNAs, eQTLs to target genes in the resource (Methods). In particular, we used Hi-C data to find all possible enhancer-target gene relationships if enhancers and targets' promoters are in the same TADs. We then found TF binding motifs using ENCODE data and imputed TF-target gene relationships if TFs have enriched binding motifs on the target gene's promoters and enhancers. In total, we included xxx enhancer-gene, xxx TF-gene, xxx eQTL-gene and xxx miRNA-gene regulatory linkages, providing a reference wiring network for gene regulation in brain.

Finally, using these "wiring" relationships, we inferred the final gene regulatory network linkages, which include the active regulatory links relating QTLs, enhancers, and transcription factors to target gene expression (Methods). In particular, given a target gene, we associated coefficients with each of these wiring linkages predicting the target gene's expression from the activities of

Deleted: inferred

Deleted: Overall

Deleted: regulatory

Deleted: to try to predict our

their regulatory elements. We model them as simple linear relationships but regularize to minimize the number of connections using an elastic net model (Methods). Overall, we found this model could successfully predict expression of >xx% genes with the minimum mean square errors < xxx. We repeated this for all genes and constructed a gene regulatory network consisting of the QTLs, enhancers, TFs and target genes with high predictive connections (Methods), revealing biological mechanisms underlying how QTLs regulate target gene expression in the adult brain. This network also has a few particular characteristics such as scale-free and hierarchical structures, which have been revealed by previous network analyses (Figure Sxx).

VIII. Integrative modeling to relate genotype to molecular and high-level phenotypes in the adult brain (870)

The interaction between genotype and phenotype involves multiple intermediate stages; in this section therefore, we perform another level of integrative analysis by embedding our gene regulatory network from the previous section into a larger model. For this purpose, we introduce an interpretable deep-learning framework, a Deep Structured Phenotype Network (DSPN, Figure 6) [[name]]. This model combines a Deep Boltzmann Machine architecture with conditional and lateral connections derived from the QTLs and gene regulatory connections predicted from our elastic net regression. As shown (Figure 6a), traditional classification methods such as logistic regression predict the phenotype directly from genotype, without inferring intermediate variables such as the transcriptome. We build the DSPN via a series of intermediate models which add layers of structure to a logistic regression model, including a layer for intermediate molecular phenotypes such as gene expression and chromatin state, multiple layers for functional modules and other mid-level phenotypes which may be inferred as hidden nodes in the network, and a layer for high-level phenotypes such as brain traits. Finally, we use special forms of connectivity (enforcing sparsity and adding lateral intra-level connections) to integrate our knowledge of QTLs, regulatory network structure, and co-expression modules from earlier sections of the paper (Supplement). By using a generative architecture, we ensure that the model is able to impute intermediate phenotypes when needed, as well as providing a predictive model for high-level traits and phenotypes.

Using the full model with genome and transcriptome data provided, we show that adding the extra layers of structure in the DSPN allows us to achieve substantially better prediction of disease and other high-level traits than without (Figure 6b) [[discuss]]. Further, comparison with a simple logistic predictor from the genome alone shows that the transcriptome carries significant further trait relevant information, which the DSPN is able to optimally extract (Figure 6a). For instance, in the case of Schizophrenia, a logistic predictor is able to gain a 2.8 times improvement when using the transcriptome versus the genome (+13% vs. +4.6% from 50% chance), while the DSPN is able to gain a 5 times improvement (+23% vs. 4.6%); this may reflect the need to incorporate non-linear interactions between intermediate phenotypes at multiple layers as in the DSPN. Moreover, the model also enables practical imputation of a subset of the transcriptome (50 genes) and epigenome (xx enhancers), with an accuracy of ~66-72% (Figure 6c). We can thus perform joint inference of the imputed intermediate phenotypes and high-level traits from the genotype alone using the DSPN, which achieves between 57.9-66.7% for disease trait prediction (Figure 6c). These results demonstrate the usefulness of even a limited amount of functional genomics information for unraveling gene-disease relationships, and that the structure learnt from such data can be used to make more accurate predictions of high-level traits even when absent.

Deleted: them

Deleted: the

Deleted: observed 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

Deleted: regulatory

Deleted: Overall, our

Deleted:).

Deleted:).

Deleted: .

Deleted:), substantially higher than the logistic predictors from the genome (Figure 6a). These results demonstrates

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

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

IX. Discussion (469)

We integrated PsychENCODE datasets with other resources, and developed a comprehensive resource consisting of various functional genomic elements for the adult brain including data from 1931 individuals. This resource serves as an important step for gaining biological insights from genomic functional data in neuroscience. Overall, our study has identified a very large-scale set of eQTLs and eGenes for adult brain, several fold more than previous studies (Figure xx), almost achieving saturation of protein coding genes. Therefore, we suspect that larger population studies will not significantly expand on these. However, there exist other aspects of brain QTLs that can be extended in the future, in addition to eQTLs. The first would be chromatin QTLs. Increasing the sample size may potentially help identify more cQTLs, which also can be further interrelated to eQTLs and other regulatory variants using our deep learning model. Moreover, the enhancers that this study used for cQTLs are defined from the current techniques such as ATAC-seq and ChIP-seq, especially from K27AC. In the future, methods such as STARR-seq may provide more accurate definitions on enhancers, and thus can be further used to better identify chromatin associated variants.

Deleted: Autism

Deleted:), and applied gene set

Deleted: underlying mechanisms.

Deleted: .

Deleted: for the adult brain, which are currently much fewer in number than eQTLs in the resource.

Deleted: signals

Deleted: state of the art

Another area of future development is single cell analysis. Current techniques suffer from the low capture efficiency, and so it remains challenging to reliably quantify low-abundant transcripts/genes and interrogate biological variation [[ref]]. In this study, we found that these basic and known cells could explain large expression variations across tissues. However, increasing single cell data and more advanced techniques in the future are expected to identify a considerable number of novel cell types, which might contribute to the unexplained variation. Using additional single cell data, our deconvolution analysis is expected to characterize cell populations more completely and estimate more accurate fQTLs in brain tissues. Also, given the issue of RNA decay in single cell RNA-seq, we intend to relate this resource to recent in situ transcriptomic data such as the spatial gene expression by optogenetic techniques, allowing us to find consistent expressed genes driving the brain phenotypes at the cellular and tissue levels.

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

Deleted: single cell

Deleted: thus integrated recent single cell data including thousands of neuronal and non-neuronal cells along with almost 1000 PsychENCODE single cells, mainly consisting of fetal cells, and

Page 4: [1] Deleted **Daifeng Wang** **2/18/18 7:23:00 PM**

in the brain, and the histone marks, together with their distances to gene transcription start sites (TSS), identify the enhancer

Page 4: [2] Deleted **Daifeng Wang** **2/18/18 7:23:00 PM**

(Moore et al, in review). Finally, we intersect these brain enhancers with

Page 4: [3] Deleted **Daifeng Wang** **2/18/18 7:23:00 PM**

peaks to consistently find brain active enhancers across all the PsychENCODE and Roadmap datasets, including ~88,800 active enhancers in the dorsolateral prefrontal cortex (see Supplement). We have

Page 4: [4] Deleted **Daifeng Wang** **2/18/18 7:23:00 PM**

sets in additional brain regions, including CBC and ACC. We also developed reference enhancer sets for the other tissues.

Page 4: [5] Formatted **Daifeng Wang** **2/18/18 7:23:00 PM**

Font: Arial, 11 pt, Font color: Black

Page 4: [6] Deleted **Daifeng Wang** **2/18/18 7:23:00 PM**

This comparison could not be achieved without such a large-scale uniform data processing.