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

I. Introduction (616)

Disorders of the brain affect nearly a fifth 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 genome (ref), until recently, little progress has been made in our fundamental understanding of the molecular cause of the brain disorders. 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). 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 changes in transcription modify the proteome, which leads to changes in brain structure and function, and these changes, in turn, interact with environmental factors to change the probability of developing a brain disorder.

To this end, a number of genomic studies have been created to focus on discovering the genomic functions for adult brain. On one hand, a variety of genomic elements and variants have been found to associate with brain and psychiatric disorders; e.g., the Psychiatric Genomics Consortia (PGC) that identified 108 GWAS loci associated with schizophrenia. On the other hand, large consortia have identified the reference sets of genomic elements across the entire body; e.g., the eQTLs and eGenes from GTEx, and the 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 has specifically tailored its work towards comprehensively identifying the functional genomics for adult brain.

To address this, recent technologies have started to detect the specific molecular activities for brain. Particularly, single-cell sequencing techniques show great promise to study the transcriptome of different neuronal and non-neuronal cells. Also, recent HiC and ATAC-seq studies found the specific chromatin structure and activity of the regulatory elements such as brain active enhancers. However, all of these studies have focused on individual aspects, and not fully been integrated to comprehensively understand the brain functional genomics such as how single cells contribute to the bulk tissue gene expression. Therefore, more efforts, especially on computational modeling and analysis for these datasets are of crucial importance to systematically understand the molecular mechanisms for brain and psychiatric disorders.

Some work such as CommonMind attempted an integrative analysis to identify the brain genomic elements including 693 differentially expressed genes associated with schizophrenia. Though promising, larger sample size, more comprehensive data, deeper modeling and analysis is further essential to obtain comprehensive view of brain functional genomics [refs]. To this endeavor, the PsychENCODE Consortium (PEC) has generated and assembled a robust large-scale dataset on the adult human brain, including genotyping, RNA-seq, ChIP-seq, ATAC-seq, HiC and single-cell data on the high quality healthy and diseased brain tissue samples of thousands of adult individuals with different phenotypes. We have thus built a central, publically available comprehensive resource (http://adult.psychencode.org/pec/) for adult brain functional genomics, including all the raw and uniformly processed data at both tissue and single cell

Deleted: **543**
Deleted: 20%… fifth of the world's ... [1]
Formatted ... [2]
Deleted:
Formatted ... [3]
Formatted ... [4]
Deleted: , including genomic variants ... [5]
Formatted ... [6]
Deleted: have generated large-scale ... [7]
Formatted ... [8]
Deleted: ChIP-seq data for dozens
Formatted ... [9]
Deleted: tissues and cell lines to ... [10]
Formatted ... [11]
Deleted: genes, transcripts
Formatted ... [12]
Formatted ... [13]
Deleted: regulatory elements (N=xxx) ... [14]
Formatted ... [15]
Deleted: show
Formatted ... [16]
Deleted: [ref], and single cell ... [17]
Formatted ... [18]
Deleted: these results still suggested ... [19]
Formatted ... [20]
Deleted: molecules do
Formatted ... [21]
Deleted: independently affect
Formatted ... [22]
Deleted: , and instead interact with ... [23]
Deleted: that drive the
Formatted ... [25]
Deleted: phenotypes
Formatted ... [24]
Formatted ... [26]
Deleted: ¶ ... [27]
Formatted ... [28]

levels from PEC and other related projects, including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, recent brain single cells [refs] with up to X,XXX samples. Using the resource, our analyses identified the functional genomic elements and activities for the adult brain on the genome scale. We also combined these elements and built an integrated deep-learning model to impute missing data and reveal the mechanisms about how they interact to drive the brain phenotypes and psychiatric disorders.

II. Comprehensive resource for adult brain functional genomics (250)

We built this comprehensive resource to have a coherent data structure. Broadly, it organizes a large amount of data for brain functional genomics in a pyramid shape (Figure 1). The bottom includes the largest scale data with often controlled access such as individual genotyping and raw next generation sequencing data of transcriptomics and epigenomics. It is followed by the uniformly processed and summarized data from the bottom such as open chromatin peaks and gene expression quantifications. Derived from these data, the middle part then includes the brain related genomic elements and interactions such as QTLs, enhancers and gene regulatory networks. Finally, at the top, the resource contains an intuitive and interpretive model revealing how the genomic elements interact to affect brain functions and phenotypes.

In terms of the data corpus for building this large-scale comprehensive resource, we included all the datasets from PsychENCODE related to the adult brain and merged them with the data from other relevant projects including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, and recent brain single cell studies. In total, this resource has XXXX data samples of 1931 individual adult brains from multiple cohorts, covering high variability among brain phenotypes and psychiatric disorders. The major data types include genotyping, RNA-seq, ChIP-seq, ATAC-seq, HiC and single-cell data. In particular, we used the annotations from reference brain project to define the brain related genomic elements, and the ENCODE standard pipelines to uniformly process all raw next generation sequencing data for both bulk and single cell data to find their activities.

III. Bulk and single cell transcriptome analysis and deconvolution explain gene expression and cell fraction changes (810s)

Given the large-scale bulk transcriptomic data in resource, we are interested to identify the genomic elements that have specific transcriptional activities in adult brain at the tissue level. In particular, we used the ENCODE standard RNA-seq pipeline to uniformly process the RNA-seq data of available samples from PEC and GTEx to quantify the expression levels for the protein coding genes, transcripts, noncoding RNA and novel transcribed regions of brain and other tissues. Using these data, we found more interpreted functional elements such as sets of differentially expressed and co-expressed genes characterizing various brain regions, phenotypes and disorders [cap1], and reported them in our resource. Moreover, we constructed a gene co-expression network using the samples across brain and other tissues and clustered it into a number of gene co-expression modules, many of which reveal the expression patterns specific for brain samples.

Deleted: ¶

Deleted: 334

Deleted: First, we

Deleted: (or hierarchical?)

Deleted: Second, in

Deleted: Moreover, we further analyzed these data and systematically identified the functional genomic elements for adult brain; i.e., the derived data including the brain-active enhancers, differentially expressed genes and transcripts, and QTLs associated with various phenotypes. In addition, our analyses and model revealed the interactions among these brain genomic elements such as imputed gene regulatory networks in which enhancer to gene linkages are identified by HiC data, and the single cell fractions of individual tissues for both neuronal and non-neuronal cells.

Deleted: 714

Moved down [1]: For example, we identified a group of genes that differentially express across ages (Figure xxx). In particular, 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.

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

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, we integrated the single cell transcriptome data to discover how the gene expression from various cell types contribute to the bulk gene expression using two strategies.

First, we used the standard pipeline to uniformly process single cell RNA-seq data in PEC in conjunction with the number of other single-cell studies on the brain to create a list of basic and primary cell types in the brain; i.e., 16 neuronal types, five non-neuronal types and xxx additional fetal types from PsychENCODE (Supplement). These are mostly concordant with what has been previously published with some minor modifications in terms of cell clusters based on their gene expression similarities (Figure Sxxx). Across these cell types, we found a number of genes varies much more substantially than they do amongst individual tissues and so forth; 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.

Second, we used an unsupervised analysis for the bulk tissue expression data and try to find its main components that potentially relate to the single cell types. In particular, we decomposed the decomposed the bulk gene expression matrix from our resource using non-negative matrix factorization (NMF, see Methods), and compared if top principal components of NMF (NMF-PCs capturing most data covariance) and the gene expression of single cells are consistent. As shown in Figure XX, we can see a number of NMF-PCs highly correlate with the biomarker gene expression signatures of neuronal, non-neuronal and fetal cell types as above; e.g., the NMF-PCs shown in Figure xxx. This shows that our unsupervised analysis derived the main components from the bulk tissue data, roughly matching the single cell data, and suggests that these cell types do make sense and contribute bulk tissue gene expression.

After this analysis, as shown in Figure xx, we further devolved 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 the multiplication of estimated cell fractions and single cell expression data can explain large variation of expression at the population level (i.e., across tissue samples). That is, $||W*C||^2/||B||^2 > 85\%$, where ||.|| is the Frobenius norm of matrix (Methods), which shows that over 80% bulk gene expression variation across samples can be accounted for a variation in single cell types. Moreover, we found that our estimated fractions of NEU+/- cells match the experimental measurements for reference brain samples (r=xxx, Figure xxx). we found that the cell fractions of individual tissues (i.e., deconvolution coefficients from W) vary, and a number of cell population changes highly associate with different phenotypes and psychiatric disorders (Figure xxx). For example, the excitatory and inhibitory types (EX3 and In6) have significantly different fractions between healthy Male and Female. The EX3 cell fractions also decrease significantly in ASD samples (p<xxx) while the non-neuronal cells increasing (e.g., oligodendrocytes). Another interesting association we found was the cell fraction changes with Age. In particular, the fractions of neuronal type(s) (EX 3 and 4) are significantly correlated with Age (r = xxx), but non-neuronal type, Oligodendrocytes anti-correlate. The cell fraction changes also potentially drive the differentially expressed genes in Age at the tissue level (Figure xxx). For example, we identified a group of genes that differentially express across ages (Figure xxx). In particular, 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 individual cell populations along with

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

Formatted: Font color: Black

Deleted:

Deleted:

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

Deleted: Y

Deleted: X

Deleted: Y=WX

Deleted: a variety

Deleted: variation

Deleted: ; i.e., ||WX||_F/||Y||_F>80% showing

Formatted: Font color: Black

Deleted: The fraction(s) of neuronal type(s) (Inhibitory X) is significantly anti-correlated with Age (r = xxx), and Inhibitory X cells have functions of XXX involving the differentially expressed genes in Age from our resource (Figure xxx). The excitatory neuronal cell populations (e.g., EX3)

Moved (insertion) [1]

significantly associated relationships between particular cell type fractions and phenotypes (Supplement).

IV. Active enhancers in adult brain (195)

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. Then, we used these data and derived further simplified epigenomic data sets in adult brain.

First, we developed a consistent set of brain active enhancers. In particular, we processed the H3K27ac and H3K4me3 ChIP-seq and ATAC-seq data of the reference brain 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 sample using the ENCODE 3 candidates regulatory element (cRE) pipeline, where the ATAC-seq peaks indicates the open chromatin in the brain, and the histone marks together with the distance to genes transcription start site (TSS) identifies the enhancer regions (Moore et al, in review). Finally, we intersect these brain enhancers with H3K27ac peaks to find brain active enhancers consistently across all the PEC and Roadmap data, including ~88,800 active enhancers in dorsal lateral prefrontal cortex (Supplement). We have also developed reference sets in additional brain regions including CBC and ACC. We also developed reference enhancer sets for the other tissues. [[Numbers about the Enhancer variation across Roughly 80k enhancers but across the 50 samples on avg 1K additional in each]]

V. Consistently comparative analysis reveals the brain related transcriptomic and epigenomic activity (499)

One key aspect of our analysis is that we, as consistently as possible, processed the transcriptomic and epigenomic data across PEC, GTEx and Roadmap together. This allows us to compare the brain to other organs in a consistent fashion to see if brain has unique gene expression and chromatin activities. This comparison couldn't be achieved without such a large-scale uniform data processing. Moreover, we attempted several methods for an appropriate comparison. Principal component analysis (PCA) and t-SNE are two popular techniques. We found that the former captured only global structures, ignoring most of the local structure. On the other hand, the latter keeps local structures intact but ignores global structures. 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 and does not give one a sense of the overall effect. Therefore, we found another technique that is capable of capturing local structures while maintaining meaningful distances in the global structure. Reference Component Analysis (RCA) projects the gene expression in individual sample against a reference panel, and then essentially reduces dimensionality of individual projections. In fact, we did RCA consistently for comparing brain and other tissues in terms of their similarities of both transcriptome (RNA-seq gene expression) and epigenome (ChIP-seq signals on our consistent set of enhancers).

Our comparative analysis for gene expression shows that brain tends to separate from the other tissues in the first component of RCA, showing it has a more distinct expression pattern and that all the brain tissue samples from the different projects grouped together (due to our uniformly processing). This difference is even accentuated when one looks not at all the individual but simply looks at the tissue cluster centers and the distribution about them. The difference between brain and other tissues is much larger than the one within any of the given tissues.

Deleted: VI
Deleted: 215
Formatted: Font color: Black
Formatted: None, Space Before: 0 pt
Formatted ... [29]
Deleted: ¶
Deleted: activte
Formatted: Font color: Black
Formatted: Font color: Black
Deleted:
Formatted: Font color: Black
Deleted: The signal variability of these ... [30]
Formatted ... [31]
Formatted ... [32]
Deleted: 463
Formatted ... [33]
Formatted ... [34]
Deleted: , but we
Formatted ... [35]
Deleted: tended to be overly ... [36]
Formatted: Font color: Black
Deleted: just uniformly
Formatted: Font color: Black
Deleted: all the clusters and
Formatted: Font color: Black
Deleted: very useful
Formatted: Font color: Black
Deleted: to be
Formatted: Font color: Black
Deleted:), which
Formatted: Font color: Black
Formatted: Font color: Black

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), implying the sophisticated interactions among brain enhancers, rather than individual enhancers per se that potentially drive brain unique gene expression.

In addition to the expression differences confined to well annotated regions, such as 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. On protein-coding regions, it has been previously known 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 stand out to some degree in having more transcriptions than most other tissues. This transcriptional diversity tends to increase with the number of samples (Figure xxx sat'd).

VI. QTL analysis (625)

To understand how the genotype affects the transcriptomic and epigenetic activities in adult brain, we used the resource data to identify more interpreted association relationship data such as the quantitative trait loci (QTLs) affecting gene expression and chromatin activity. In particular, we calculated the association of 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 several major categories: expression QTLs (eQTLs), chromatin QTLs (cQTLs), splicing QTLs (sQTLs) and even cell fraction QTLs. For the eQTLs, we adopted a standard approach and emphasized the scale of the database. We adhered closely to the established GTEX eQTL pipelines. We identified ~2M of eQTLs and ~17000 number of e-genes in DLPFC region. This is a conservative larger number of eQTLs than previous brain eQTL studies and reflects the very large sample size and great power we have. We believe it's moving close to saturating 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 ~10M sQTLs. For the cQTLs, the situation is more complicated. There are no established standard methods of calculating cQTLs on a large scale. To properly identify cQTLs, we focused on a reference set of enhancers to define the region associated with the activity of the chromatin and then look at how this activity varies. We used ENCODE CREs to define the consensus regions of H3K27ac marker used to calculate average signal value. And then we correlated this with nearby variants. (See methods). Overall, we were able to identify ~2000 cQTLs.

Furthermore, we are interested to see if any genotype is also associated with the single cell fractions. In particular, we used our QTL pipeline and identified 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. This suggests that these fQTLs potentially can be used to predict the cell fractions in adult brain. Moreover, we also identified xxx SNPs significantly associated with the gene expression

Formatted ... [37]
Formatted ... [38]
Deleted: indistinguishable
Formatted ... [39]
Deleted:). Thus,
Formatted ... [40]
Deleted: gene expression difference ... [41]
Formatted ... [42]
Deleted: cannot be simply attributed to ... [43]
Formatted ... [44]
Deleted: be driven by more complex ... [45]
Formatted ... [46]
Deleted: ,
Formatted ... [47]
Deleted: that we're looking at are ... [48]
Formatted ... [49]
Deleted: known canonical
Formatted ... [50]
Deleted: however people have ... [51]
Formatted ... [52]
Deleted: tried to get a sense of this ... [53]
Formatted ... [54]
Deleted: transcript
Formatted ... [55]
Deleted: the entire genome. In terms ... [56]
Formatted ... [57]
Deleted: genes
Formatted ... [58]
Deleted: tends
Formatted ... [59]
Deleted: most
Formatted ... [60]
Deleted: does
Formatted ... [61]
Deleted: transcriptional diversity
Formatted ... [62]
Deleted: **598**
Formatted ... [63]
Deleted: In order to
Formatted ... [64]
Deleted: have to both
Formatted ... [65]
Deleted: did joint K27 peak calling ... [66]
Formatted ... [67]
Deleted: , we calculated an
Formatted ... [68]

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.

Given the QTLs we identified, we overlap and annotate them with a variety of different genomic annotations and look at the degree that they overlapped. The distributions of detailed QTL annotations on genomic regions are shown in Figure xxx. As expected, there's a very large amount of overlap between the cQTLs and eQTLs, and with ~50% of the cQTLs essentially being a subset of the eQTLs.We examined the enrichment of most significant eQTLs per gene in Roadmap Epigenomics Consortium and ENCODE enhancers across XX human tissues and cell lines. 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 leads on which genes are affected in disease. 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 disorders than the ones of non-brain disorders.

VII. Gene regulatory networks in adult brain (523)

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 to control target gene expression in adult brain. Gene regulation is a key mechanism that genotype affects phenotype. This comprehensive resource thus enables us to identify the gene regulatory relationships among the brain genomic elements.

To this end, the first step is to process a full Hi-C data for adult brain, which provides direct physical evidence for potential interactions between enhancers and promoters in the format of topologically associated domains (TADs) (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, using the protocol of XXX (Supplement). In total, we identified xxx TADs in adult brain. Overall, these TADs show a number of established properties, such as the gene expression tends to increase with increasing numbers of interactive enhancers (Figure 5xx). More importantly, we found that >xx% enhancer-promoter interactions happen in the same TADs in adult brain (Figure 5xx), suggesting that TADs potentially provide at large cis-regulatory relationships between enhancers and target genes.

Therefore, the second step to build the gene regulatory network is that we integrated the TADs with other regulatory 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 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 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 on gene regulation in brain.

Finally, using these "wiring" regulatory relationships, we inferred the gene regulatory network that include the active regulatory relationships on how QTLs, enhancers, and transcription factors relate to target gene expression (Methods). In particular, given a target gene, we then associate coefficients with each of these wiring linkages to try to predict our target gene expression from the activities of their regulatory elements. We model them as simple linear

Deleted:

Formatted: Font: Times New Roman, Font color: Auto

Formatted: Font: Times New Roman, Font color: Auto

Formatted: Font color: Black

Deleted: most

Formatted: Font color: Black

Deleted: eQTLs. We found that XX% of cQTLs are overlapped with

Formatted: Font color: Black

Deleted:

Formatted: Font color: Black

Deleted: 6

Formatted: Font color: Black

Deleted: 10

Formatted: Font color: Black

Deleted: ,

Formatted: Font color: Black

Deleted: related

Formatted: Font color: Black

Deleted: related

Formatted: Font color: Black

Deleted: In addition, we link the QTLs that overlap the enhancers and promoters in the resource to reveal the potential regulatory activities (Figure Sxx).

Formatted: Font: Times New Roman, Font color: Auto

Deleted: also …nables us to identify [76]

Deleted: we…he first used the HiC [77]

Deleted: gene co-expression network [78]

Deleted: networks…etwork that ... [79]

relationships but regularize them to minimize the number of connections using the elastic net model that combines the L1 and L2 penalties of the lasso and ridge regressions (Methods). Overall, we found this model could predict expression successfully as shown from Figure xxx. 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 network consisting of the QTLs, enhancers, TFs and target genes with high predictive relationships (Methods), revealing the biological mechanisms on how QTLs regulate the 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 our previous regulatory network analyses (Figure Sxx).

VIII. Integrative modeling to explain the molecular mechanisms for genotype-phenotype relationships in adult brain (821)

The interaction between genotype and phenotype is a complex process, involving multiple intermediate stages including gene regulatory network. Thus, in this section, we perform another level of integrative analysis for the resource by embedding our gene regulatory network into a larger model; i.e., we introduce an interpretable deep-learning framework, a Deep Structured Phenotype Network (DSPN), which provides insight into how the genomic variants link to the regulatory network, then to functional modules, and eventually predict phenotypes such as schizophrenia (Figure xxx). 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. In particular, it integrates all high dimensional functional data types in this resource including genomics, transcriptomics, epigenetics and regulation, and genotype-phenotype relationships, and also allows us to quantitatively impute missing transcriptional and epigenetic information for samples with genotypes only. This allowed us to not only feed information from the bottom of the network to the top i.e. variants all the way up to phenotypes, but also propagate information throughout the network, predicting for instance the transcriptome from genomic variants or directly predicting phenotype from the transcriptome. We also make the model downloadable as a set of simplified files encompassing the elastic net model described earlier plus additional DSPN connections between layers such as a groups of gene modules to phenotypes.

As shown in Figure xxx, traditional classification methods such as logistic regression predict the phenotype directly from genotype, missing the intermediate information such as transcriptome (Figure xx). We build the DSPN via a series of intermediate models which add layers of structure to a logistic 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)

We examined the connections learnt by the DSPN between intermediate and high-level phenotypes for potential mechanisms to see if they are biologically meaningful. For example, the module xxx is connected to genes enriched in the dopaminergic and glutamatergic synapse (GSEA enrichment score > xxx, Figure xx), and the module yyy is connecting to Age, and represents the neuronal cell fractions (Figure xxx). Furthermore, we used this model to

Deleted: eQTLs,

Deleted: enhancers that control its gene expression 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

Deleted: to predict the regression coefficients of genotypes of various QTLs, the chromatin stages of enhancers, splicing patterns and TFs gene

Deleted: to the target gene expression, and identified the highly predictive relationships (i.e., large coefficients).

Deleted: networks

Formatted: Font: Arial, 11 pt, Font color: Black

Deleted: **613**

Deleted: We thus

Deleted: Networks

Deleted: brain

Deleted: affect gene expression and regulation

Deleted: regulatory networks estimated in our resource. On the resource website, we provide a list of DSPN pathways for each phenotype and disease. We also make the model downloadable as a set of simplified files summarizing represented ... [80]

Deleted: The model is trained as a deep generative model to represent ... [81]

Deleted: .

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 also found some potentially additional 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). The complete functional annotations of modules are available in supplement.

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. Further, we transform the results above to the liability scale in order to compare with heritability estimated on this scale using GCTA (Figure xxx). Using the PEC 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%), suggesting this model is near optimal (although there may be further epistatic interactions the model can capture). The full DSPN model estimates that the transcriptome-based liability for the PFC is ~19.2%. Although we expect that a large portion of this will overlap with the common SNP based liability, it may also include environmental and epistatic contributions (see Supplemental Figure), precluding direct comparison. Similar estimates of the liability explained for Bipolar and Autism by the DSPN (imputation and full models) are given (Figure xxx).

IX. Discussion (558)

We integrated the high-dimensional brain genomic datasets of PsychENCODE and other projects from 1931 individuals, and developed a comprehensive resource consisting of various functional genomic elements for the adult brain. This resource serves as an important step in gaining biological insights from genomic functions and mechanisms in neuroscience. Neuroscientists 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 in brain such as the neurodegenerative diseases like Alzheimer and Parkinson.

Overall, our study has identified a very large-scale set of eQTLs and eGenes for adult brain (Figure xx), almost achieving the saturated numbers. Therefore, we suspect that the future larger population studies would be very helpful to this context. 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 for adult brain, which is currently much less than eQTLs in the resource. Increasing sample size such as large population potentially helps identify more cQTLs, which also can be further interrelated to eQTLs and other regulatory variants from our deep learning

Deleted: 506

Formatted: None, Space Before: 0 pt, Add space between paragraphs of the same style

Deleted: this

Deleted: In particular, our comparative analyses found that these genomic elements significantly relate with the psychiatric disorders and other brain phenotypes including developmental stages [cap2]. The neuroscientists

Formatted: Font color: Text 1

Deleted: individual's fMRI image features measuring functional neuro-connectivity to identify the associated genotypes such as image-QTLs (iQTLs) [xx]. Also, it can incorporate with

Formatted: Font color: Text 1

Formatted: Font: Arial, 11 pt, Not Bold, Font color: Text 1

Deleted: ¶
Moreover, by combining the resource data, we built an integrative deep learning model, DSPN to reveal the interactions and mechanisms among various high-dimensional functional genomic elements from a number of directions between genotype and 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 phenotype. It is also available online as a general-purpose tool and enables quantitatively imputing missing transcriptional and epigenetic information for samples with ... [82]

model. Moreover, the enhancers that this study used for cQTLs are defined from the current techniques such as ATAC-seq and ChIP-seq signals, especially from K27AC. The future and new state of the art methods available such as STARR-seq provide more accurate definitions on enhancers, and thus should be further used to better identify such as chromatin associated variants.

Another aspect that might move forward is single cell analysis. The current single cell techniques suffer from the low capture efficiency, so remain challenging to reliably quantify the low-abundant transcripts/genes and interrogate the biological variations [refs]. However, it is still worthwhile using the biomarker genes with strong expression signals in single cell to deconvolve the tissue gene expression data to find the cell fractions for individual tissues and relate to the individual phenotypes. In this study, we thus integrated recent single cell data including thousands of neuronal and non-neuronal cells along with almost 1000 PEC single cells mainly consisting of fetal cells and 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 will identify considerably large number of novel cell types, which might contribute to unexplained variations. Using these additional single cell data, our deconvolution analysis expects to estimate more complete cell populations and accurate fQTLs to brain tissues.

Moreover, more accurate cQTLs and fQTLs can be input into our deelp learning model, which might also improve the model performance. Our model represents a state of the art method at the moment to reveal genotype-phenotype at the population level but might improve with the development of machine learning and additional data types such as image and medicine. Furthermore, while providing better prediction, some model connections are deliberately set to be interpreted simplifications, such as gene regulatory networks, to make the model more interpretable and easier to use. Thus, another major goal of the model is to provide a compression of large 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.

Moved down [2]: Furthermore, while providing better prediction, some model connections are deliberately set to be interpreted simplifications, such as gene regulatory networks, to make the model more interpretable and easier to use.

Deleted: Thus,

Moved down [3]: another major goal of the model is to provide a compression of large 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.¶

Formatted: Font color: Text 1

Formatted: Space After: 4 pt

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

Deleted: ¶

Deleted: With increasing amount of single cell data in near future, we could deconvolve the tissue data in the resource to find potential new cell types and obtain more complete cell populations. 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[83]

Formatted: Font color: Text 1

Formatted: Font: Arial, 11 pt, Font color: Text 1

Moved (insertion) [2]

Formatted: Font color: Text 1

Moved (insertion) [3]

Formatted: Font: Times New Roman, 12 pt, Font color: Auto

Page 1: [1] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

20%

Page 1: [1] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

20%

Page 1: [1] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

20%

Page 1: [1] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

20%

Page 1: [1] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

20%

Page 1: [2] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font: Arial, 11 pt, Font color: Black

Page 1: [3] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font: Arial, 11 pt, Font color: Black

Page 1: [4] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [4] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [4] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [5] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

, including genomic variants and genes found by many studies. For example, 108 GWAS loci and 693 differentially expressed genes associated with schizophrenia identified by

Page 1: [5] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

, including genomic variants and genes found by many studies. For example, 108 GWAS loci and 693 differentially expressed genes associated with schizophrenia identified by

Page 1: [6] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [7] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

have generated large-scale RNA-seq

Page 1: [8] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [9] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [9] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [10] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

tissues and cell lines to systematically identify

Page 1: [11] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [12] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [13] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [14] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

regulatory elements (N=xxx). Moreover, recent

Page 1: [15] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [16] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [17] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

[ref], and single cell techniques can detect gene expression and epigenetic patterns for neuronal and non-neuronal cells from brain tissues [ref].

Page 1: [18] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [19] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

these results still suggested that thousands of samples would be required to achieve statistical power of 0.8 for detecting a complete set of brain-related genomic elements [refs]. Also,

Page 1: [20] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [21] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [22] Formatted **Daifeng Wang** **2/9/18 3:10:00 AM**

Font color: Auto

Page 1: [23] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

, and instead interact with each other in a network. Thus, effort is also needed to model and analyze the molecular interactions and

Page 1: [24] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [25] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [26] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [26] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 1: [27] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

In fact, understanding the mechanisms on how these genomic elements affect various brain functions and phenotypes is still a key challenge in neuroscience. To address it

Page 1: [28] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Auto

Page 4: [29] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font: Times New Roman, 12 pt, Not Bold, Font color: Auto

Page 4: [30] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

The signal variability of these enhancers across individuals can be further used to identify chromatin QTLs in the subsequent section.

Page 4: [31] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font: Times New Roman, Font color: Auto

Page 4: [32] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

None, Space Before: 0 pt, After: 0 pt

Page 4: [33] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font: +Body (Calibri), 12 pt, Not Bold, Font color: Text 1

Page 4: [34] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 4: [34] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 4: [35] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 4: [35] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 4: [36] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

tended to be overly influenced by data outliers, and

Page 5: [37] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [38] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [39] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [40] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [41] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

gene expression difference, in a sense, in

Page 5: [42] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [43] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

cannot be simply attributed to that chromatin, but

Page 5: [44] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [45] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

be driven by more complex gene regulatory mechanisms involving enhancers.

Page 5: [46] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [47] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [48] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

that we're looking at are those

Page 5: [49] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [50] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [50] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [51] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

however people have previously remarked about the

Page 5: [52] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [52] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [53] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

tried to get a sense of this looking

Page 5: [54] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [55] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [56] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

the entire genome. In terms of

Page 5: [57] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [58] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [59] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [60] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [60] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [61] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [62] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [62] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [63] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [64] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [65] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [66] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

did joint K27 peak calling over the hundreds of brains in PsychENCODE with H3K27ac marker. From

Page 5: [67] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [68] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [69] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [70] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [71] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [72] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

both neuronal and non-neuronal

Page 5: [73] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 5: [74] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

For example, fQTL xxx is for Ex3 fraction

Page 5: [75] Formatted	Daifeng Wang	2/9/18 3:10:00 AM

Font color: Black

Page 6: [76] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

also

Page 6: [76] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

also

Page 6: [77] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

we

Page 6: [77] Deleted	Daifeng Wang	2/9/18 3:10:00 AM

we

Page 6: [77] Deleted Daifeng Wang 2/9/18 3:10:00 AM

we

Page 6: [77] Deleted Daifeng Wang 2/9/18 3:10:00 AM

we

Page 6: [77] Deleted Daifeng Wang 2/9/18 3:10:00 AM

we

Page 6: [77] Deleted Daifeng Wang 2/9/18 3:10:00 AM

we

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [78] Deleted Daifeng Wang 2/9/18 3:10:00 AM

gene co-expression network using all PsychENCODE and GTEx samples and clustered it into gene co-expression modules using WGCNA [Methods].

To construct

Page 6: [79] Deleted Daifeng Wang 2/9/18 3:10:00 AM

networks

Page 6: [79] Deleted Daifeng Wang 2/9/18 3:10:00 AM

networks

Page 6: [79] Deleted Daifeng Wang 2/9/18 3:10:00 AM

networks

Page 6: [79] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

networks

Page 7: [80] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

regulatory networks estimated in our resource. On the resource website, we provide a list of DSPN pathways for each phenotype and disease. We also make the model downloadable as a set of simplified files summarizing represented genotype-phenotype pathways. In particular, this model

Page 7: [81] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

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, this architecture allows information to flow in top-down, bottom-up and lateral directions during inference

Page 8: [82] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

Moreover, by combining the resource data, we built an integrative deep learning model, DSPN to reveal the interactions and mechanisms among various high-dimensional functional genomic elements from a number of directions between genotype and 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 phenotype. It is also available online as a general-purpose tool and enables quantitatively imputing missing transcriptional and epigenetic information for samples with genotypes only. Also, the model can be used to prediction the outcomes of in-silico perturbations; e.g., knocking down GRIN1 potentially breaks the excitatory and glutamatergic signaling pathways to likely affect schizophrenia.

Page 9: [83] Deleted **Daifeng Wang** **2/9/18 3:10:00 AM**

With increasing amount of single cell data in near future, we could deconvolve the tissue data in the resource to find potential new cell types and obtain more complete cell populations. 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 recent in situ transcriptomic data such as the spatial gene expression by optogenetic techniques, and find the consistent expressed genes driving the brain phenotypes at the cellular and tissue levels