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

**Introduction (~562 words, aiming 550)**

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.

# prelude - dic1pre

\* some gwas

\* some ref. annotation but not much brain

\* some focused brain but could do more

To this end, a variety of genomic elements and variants in them have been found to associate with brain and psychiatric disorders. For example, 108 GWAS loci and 693 differentially expressed genes associated with schizophrenia identified by Psychiatric Genomics and CommonMind consortia. In somewhat different but related efforts, other consortia such as GTEx, ENCODE and Epigenomics Roadmap have generated large-scale RNA-seq and ChIP-seq data for dozens of brain tissues and cell lines to systematically identify brain-associated genes, transcripts and regulatory elements (N=xxx). Moreover, recent studies show the specific chromatin structure and activity of the regulatory elements such as brain active enhancers [ref], and single cell techniques can detect gene expression and epigenetic patterns for neuronal and non-neuronal cells from brain tissues [ref]. However, 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, individual molecules do not independently affect brain, and instead interact with each other in a network. Thus, effort is also needed to model and analyze the molecular interactions and mechanisms that drive the brain phenotypes and psychiatric disorders.

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, the PsychENCODE Consortium (PEC) has generated and assembled a robust large-scale dataset on the adult human brain, including genotyping, RNA-seq, ChIP-seq, HiC and single-cell data on the high quality healthy and diseased brain tissue samples of 1931 adult individuals with different phenotypes and produced these data in a central, publically available resource (PEC ref). We have also supplemented the resource with the primary data at both tissue and single cell levels from other related projects, including ENCODE, CommonMind, GTEx, Epigenomics Roadmap, recent brain single cells [refs], and uniformly processed all the data together and performed integrated analyses with up to X,XXX samples. Our analyses provide a comprehensive list of functional genomic elements for the adult brain including the brain-active enhancers, transcripts, expression models, imputed regulatory networks, single cell fractions, and QTLs for various phenotypes such as eQTLs. 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.

**Discussion (~ 506 words, aiming 500)**

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. In particular, our comparative analyses found that these genomic elements significantly relate with the psychiatric disorders and other brain phenotypes including developmental stages [cap2]. 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 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 the neurodegenerative diseases like Alzheimer and Parkinson.

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 power to better predict phenotype. It is 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. 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.

# disc - dic2disc

\* single-cell

- fundamental limitations

- more cells

\* chromatin - pop.

\* sat. RNA-seq - ref. enhancers

- new measurements

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. 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.