**Comprehensive resource, integrative single cell analysis and deep neural network model for functional genomics of the adult brain**

**Contact PI: Mark Gerstein**

**Extended abstract.** Robust phenotype-genotype associations have been established for a number of neuropsychiatric disorders such as schizophrenia and bipolar disorder; however, the molecular mechanisms underlying these associations are unknown. Addressing this deficiency is a key aspect of the PsychENCODE Consortium. To this end, the consortium has generated genotypic, transcriptomic, epigenetic, Hi-C and single-cell sequencing data from thousands of individuals and uniformly processed them to ensure consistency. In addition, we have uniformly re-processed other large-scale genomic resources (e.g., Epigenomics Roadmap and GTEx) and merged them with the PsychENCODE corpus to develop the largest data and analytic resource of the human adult brain (comprising 1945 individuals). This resource allows us to make a number of advances:

 \*We compare transcriptomic and epigenomic data between brain and other tissues in a consistent fashion, and develop the largest reference set of adult brain functional genomic elements including active brain enhancers, transcripts and gene regulatory networks. For instance, we identified the specific enhancers that active on major brain regions, e.g. ~88.8K active enhancers and ~79k brain active transcripts . Moreover, using a variety of spectral analyses, we find that the brain has more distinct expression patterns compared to most other tissues (Figure A), including relatively large amount of non-coding transcription. However, the differences in epigenetics are smaller.

 \*Using single-cell data of both neuronal and non-neuronal cells, we find that >80% of the inter-individual variation of gene expression in bulk brain tissue can be accounted for by alterations in the proportions of basic cell types, rather than by changes in individual genes. Moreover, the cell fractions vary across brain phenotypes.

\*We develop the largest set of brain quantitative trait loci (QTLs) that associate with brain phenotypes and disorders, including expression (eQTLs), chromatin (cQTLs), alternative splicing (spQTLs) and even cell fractions (fQTLs). For example, we have found >1M variants and >11K eGenes including non-coding genes, (more than previous studies, Figure B). We also observed >5K associated chromatin QTLs. Collectively, these QTLs help annotate a larger fraction of brain-GWAS SNPs (e.g., 6% in schizophrenia, 10% in bipolar) than previously observed, providing leads on which transcripts are altered in disease.

\*We merge the analytic results into an interpretable deep generative network model, where we enforce constraints on the connectivity to mirror genotype-phenotype connections and regulatory architecture at multiple levels. This allows us to relate genotypes, gene-expression levels and epigenetics to the regulatory network and QTLs. The model enables practical imputation of a subset of the transcriptome and epigenome with an accuracy of ~70%.

 \*We use the integrated model to improve prediction of biological variables and psychiatric diseases by the addition of transcriptomic data to genotype. In particular, we show that we can predict bipolar disease and schizophrenia with much higher accuracy from the transcriptome than from genotype alone (i.e., three times accuracy improvement over a random baseline of 50%, +18% vs +6% for schizophrenia, Figure C). We also demonstrate the clear predictive value of the imputed transcriptome, showing the usefulness of even a limited amount of functional genomics information for unraveling gene-disease relationships. For instance, we can predict schizophrenia with an accuracy of 61% using our model and an imputed transcriptome, versus 56% with genotype alone.

**Why Science? [DW, PE]**

**\* main psychencode that describes the bulk pec data - the largest resource on the adult human**

**\* itnegraties single cell & gene expr variation -**

**\* in addition many eqtls - we have chromatin and even cell fraction qtls**

**\* we provide resource for the & interpretable model**

**\* multiple layers that strongly predict psychiatric - shows how the data can be leveraged to predict**

This work is one of main PsychENCODE papers describing its functional genomic data and presenting the largest comprehensive resource to date on the adult brain. We have successfully integrated single cell data to deconvolve the tissue gene expression to elucidate phenotypic differences in cellular fractions. Additionally, we have compiled a list of QTLs for various functional genomic activities such as gene expression, chromatin-state and cell fractions. Finally, we construct an interpretable deep neural network model embedding gene regulatory networks to leverage QTLs, genes, enhancers and functional modules to impute and predict the brain variables and psychiatric disorders.