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

**Contact PI(s): M Gerstein & Jim Knowles**

Robust phenotype-genotype associations have been established for a number of neuropsychiatric diseases 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 adult human 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. ~88,800 active enhancers and ~79,000 brain active transcripts in pre-frontal cortex. 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**, 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 those for expression (eQTLs), chromatin (cQTLs), alternative splicing (spQTLs) and even cell fractions (fQTLs from associated with tje single-cell analysis). For example, we have found >1M variants involved in eQTLs, involving >11K genes including non-coding one (considerably more than previous studies, Figure B). We also observed >5K associated chromatin QTLs. Collectively, these QTLs help annotate a larger fraction of GWAS SNPs involving the brain (e.g., 6% in schizophrenia, 10% in bipolar) than previously observed, providing leads on which genes are affected in disease.

\* We merge the analytic results into a **generative, deep-learning model**, where we enforce interpretable 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 functional genomics 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% (over 50% random) using our model and an imputed transcriptome, versus 56% with genotype alone.

**Why Science?**This is one of the main capstone papers describing the consortium’s data and presenting the largest resource to date on adult brain. It represents a “next-generation resource,” merging human population variation, epigenetics and single-cell data with interpretable, deep-learning models for predicting psychiatric disorders.



