**Gerstein lab experience with processing and deconvolving bulk RNA-seq data to estimate cell populations in human brain tissues**

To better understand the molecular causes of psychiatric disorders, we integrated multi-omics data from PsychENCODE and other large consortia to construct a comprehensive functional genomic resource for the human brain (resource.PsychENCODE.org), which contains 3,810 genotype, transcriptome, chromatin and Hi-C datasets from PsychENCODE and 1,662 datasets using similar “bulk” assays merged from outside the consortium10. Overall, the datasets from the prefrontal cortex (PFC) comprise samples from 1,866 individuals. The resource also includes single-cell RNA-seq data for 18,025 cells from PsychENCODE and 14,012 cells from outside sources. These data span a range of psychiatric disorders, including schizophrenia (SCZ), Bipolar Disorder (BPD), and Autism Spectrum Disorder (ASD). Using this resource, we identified brain functional genomic elements, including a number of disease genes (e.g., 321 high-confidence SCZ genes) and ~79,000 brain-active enhancers. These findings enabled us to further link genes and transcription factors (TFs) in an extended regulatory network.

We uniformly processed multiple independent bulk and single cell datasets (including GTEx). We also clustered and merged single cell RNA-seq datasets with recently published datasets (Lake et al, 2016; Lake et al, 2018) to develop a consistent set of single cell profiles, such as those for inhibitory and excitatory neurons, non-neuronal cell types (e.g., microglia, astrocytes) and other cell types in development.

Gene expression changes observed at the tissue level may result from changes in the proportions of basic cell types. We previously investigated how changes in cell proportions contribute to variation in tissue-level gene expression across individuals (Eichler et al., 2010). Adding publicly available data, we first used non-negative matrix factorization (NMF) to decompose bulk tissue data. We found that the top principal components correlated with cell expression signatures, suggesting single cell contributions to the bulk tissue gene expression. We then deconvolved the bulk tissue expression across 1866 individuals in PsychENCODE and GTEx using single-cell data via non-negative least squares. Interestingly, differences in the proportions of cell types explained >85% of the cross-population variation observed. In addition, we found that the cell fraction changes were associated with aging, disorders (Wang et al, 2018).

**Gerstein lab experience with integrative data analysis of regulatory networks and their changes**

We have extensive experience building regulatory networks from bulk ‘omics data in tissues and cell lines. We used a compendium of ENCODE ChIP-seq data to investigate the hierarchical network architecture of regulation by TFs in human cell lines (Gerstein et al, 2012). We also built regulatory networks from RNA-seq data across species (Gerstein et al, 2014). Most recently, we used a variety of data types, including ChIP-seq, RNA-seq and Hi-C conducted in brain tissues in PsychENCODE to build regulatory networks that helped to predict psychiatric disease (Wang et al, 2018). In particular, this work used an elastic net model to linearly combine the *L1* and *L2*-norm regularizations to determine TF-target relationships. As regulatory networks reflect cell type compositions, we identified cell type-associated gene regulatory networks (GRNs). These revealed regulatory networks regulating key cell-type marker genes, which may determine cell fates.

We also included additional links within the bulk gene regulatory network by identifying quantitative trait loci (QTLs), as associations between genomic variants and intermediary phenotypes – such as expression (eQTLs), chromosomal marks (cQTLs), and relative fractions of cell types (fQTLs) – may define regulatory relationships within this network. To identify potential regulatory effects from genomic variants for building gene regulatory networks, we also calculated the quantitative trait loci (QTLs). We recently published work identifying an array of these different QTL types by leveraging uniformly processed data from 1866 individual (Wang et al, 2018). These yielded QTLs originating from ~1.3 million, 7,976 and 1,672 SNPs, which were associated with eQTLs, cQTLs, and fQTLs, respectively. Given the large number of samples, this provides a substantially larger dataset of brain-based QTLs than previously available, and the eQTLs identified are consistent with those reported by GTEx. We also evaluated the intersections between these various QTL types (i.e., the frequency with which the SNPs between different QTL types were shared), thereby allowing us to select “multi-QTLs”, which are highly penetrant in that they influence multiple intermediate phenotypes.

We have built statistical frameworks to study relationships between chromatin features and gene expression from bulk sequence data (Cheng et al, 2011; Cheng et al, 2012). These frameworks can be used to predict expression of protein-coding genes and noncoding RNAs, which we demonstrated in worm, fly, mouse, and human. Our models have revealed the positional contribution around genes (upstream or downstream) of distinct chromatin features to predicted expression levels.

We also have extensive experience in conducting network comparisons. For example, we developed many tools for comparative gene regulatory network analyses, including Loregic (Wang et al, 2015), which analyzes regulatory cooperativity; OrthoClust (Yan et al, 2014), which discovers novel human gene functions via clustering cross-species gene co-expression networks; and DREISS (Wang et al, 2016), which analyzes the dynamics of gene regulatory networks using dynamic models.

**Gerstein lab experience with developing a deep-learning model for predicting phenotype from genotype as it relates to psychiatric diseases**

To predict phenotype from genotype (as it relates to psychiatric disorders and aging), we developed an interpretable deep-learning model embedding the GRN. Our model, called Deep Structured Phenotype Network (DSPN; Wang et al, 2018), uses a conditional Deep Boltzmann Machine (cDBM) architecture with multiple layers (including genotype, gene expression, epigenetics and cell fraction layers) and introduces lateral connectivity at the visible layer to embed the GRN and QTL links. DSPN improves disease prediction *by 6-fold compared to additive polygenic risk scores for SCZ*, highlights key genes associated with particular disorders, and allows imputation of missing transcriptome information from genotype data alone. We also developed a rank statistic-based interpretation scheme that allows us to functionally annotate hidden nodes and prioritize variants, genes and hidden nodes relative to disorders.

**Referenced work**

Cheng, C. *et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* **12**, R15, doi:10.1186/gb-2011-12-2-r15 (2011).

Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* **22**, 1658-1667, doi:10.1101/gr.136838.111 (2012).

Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-450, doi:10.1038/nrg2809 (2010).

Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).

Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445-448, doi:10.1038/nature13424 (2014).

Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590, doi:10.1126/science.aaf1204 (2016).

Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80, doi:10.1038/nbt.4038 (2018).

Wang, D. *et al.* Loregic: a method to characterize the cooperative logic of regulatory factors. *PLoS Comput Biol* **11**, e1004132, doi:10.1371/journal.pcbi.1004132 (2015).

Wang, D., He, F., Maslov, S. & Gerstein, M. DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks. *PLoS Comput Biol* **12**, e1005146, doi:10.1371/journal.pcbi.1005146 (2016).

Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, doi:10.1126/science.aat8464 (2018).

Yan, K. K. *et al.* OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol* **15**, R100, doi:10.1186/gb-2014-15-8-r100 (2014).