Gerstein lab experience with integrative analysis of bulk human tissue

The Gerstein lab has considerable experience in analyzing multi-modal genomic and histological datasets of the adult brain. Recent genomic datasets we have used include data from PsychENCODE, BrainSpan, and the Human Brain Atlas. These projects provided systematic and integrative analyses of the transcriptomic, epigenomic, and regulatory features of the human brain at both tissue and single-cell resolution. From these datasets, we have derived substantial insights into human development and disorders.

Deconvolution analysis of bulk and single-cell transcriptomics. To test if the gene expression changes observed at different brain sub-regions are due to changes in proportions of basic cell types, we investigated cell fraction changes across brain disorders, age and developmental stages through combined analysis of gene expression data from 1866 individuals from multiple large datasets including PsychENCODE, adult GTEx, Common Mind Consortium (CMC)[24]. We first used non- negative matrix factorization (NMF) to decompose bulk tissue data and found that the top principal components correlated with cell expression signatures. We then deconvolved the bulk tissue expression using single-cell data via non-negative least squares. Interestingly, differences in the proportions of cell types explained >88% of the cross-population variation observed. In addition, we found that the cell fraction changes were associated with aging, disorders (e.g., increasing excitatory-to-inhibitory ratio, microglia and astrocytes in ASD), and developmental stages (e.g., prenatal astrocytes decreased during postnatal phases) (Fig. 1).

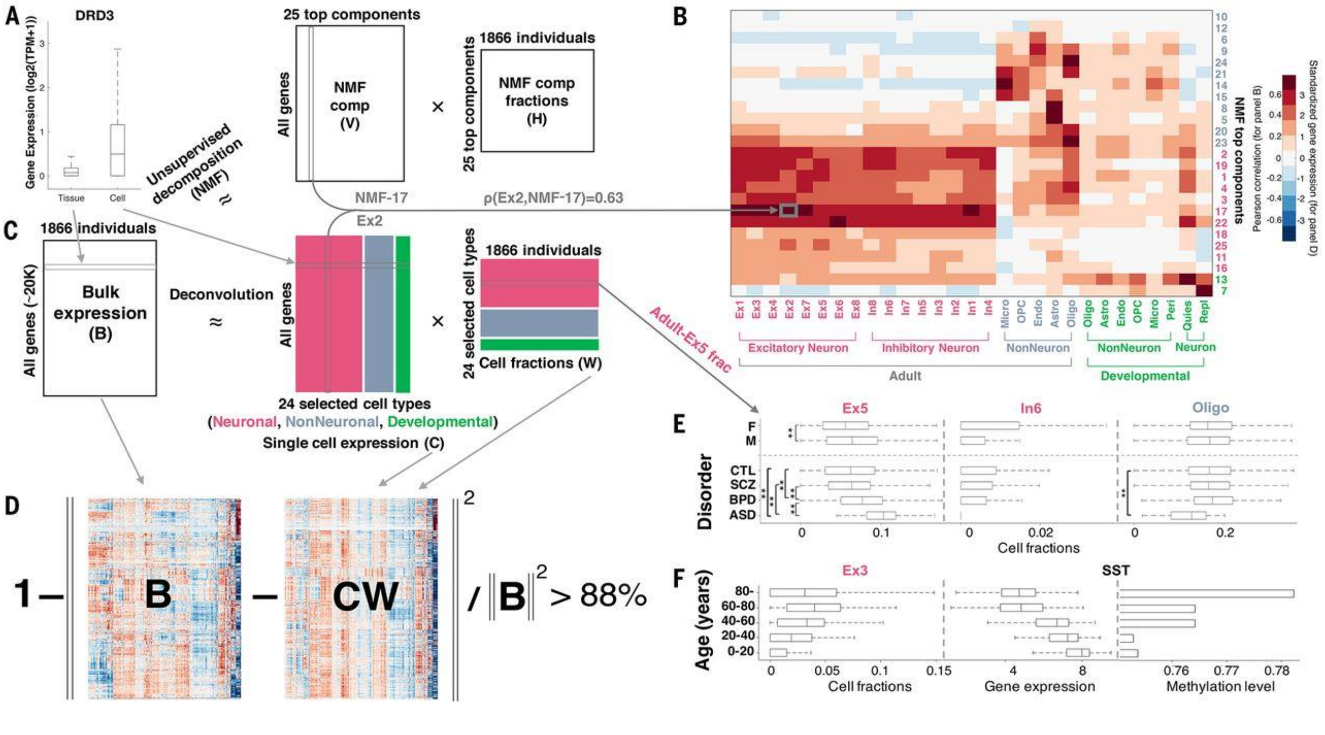

Fig. 1 Deconvolution analysis reveals cell fraction changes across the population. A) Expression variability across single cells sampled from different types of brain cells and across samples from a population. **B)** A heatmap of correlation coefficients in expression between NMF-TCs and single-cell signatures. **C)** The bulk tissue gene expression matrix (B, genes by individuals) can be decomposed by NMF (top) and deconvolved by the single-cell gene expression matrix (bottom; C, genes by cell types) to estimate the cell fractions across individuals (the matrix W). **D)** The estimated cell fractions account for >88% of the bulk tissue expression variation. **E)** Cell fraction changes across genders and brain disorders. **F)** Changing cell fractions (for Ex3), gene expression (for SST), and promoter methylation level (median level, for SST) across age groups.

Gerstein lab experience with studying allele-specific expression

The Gerstein lab developed AlleleSeq (Rozowsky et al, 2011), which uses RNA-seq and ChIP-seq data to detect allelic sites, including those associated with gene expression and TF binding. AlleleSeq also constructs personal diploid genomes. We have spearheaded allele-specific analyses as part of our efforts in several major consortia, including ENCODE and the 1000 Genomes Project. We have also applied AlleleSeq to newer datasets, which have enabled us to annotate the SNP catalog with allelic information. We have also constructed an associated AlleleDB database (Chen et al, 2016). Both AlleleSeq and AlleleDB are widely used by the scientific community.

Gerstein lab experience with integrating large-scale datasets

We have considerable experience with integrating large-scale datasets while working with large consortia. For instance, we have used ENCODE data to develop a customized annotation (for the purpose of studying cancer progression) for genome interpretation by leveraging an array of experimental assays, such as eCLIP, Hi-C, and whole-genome STARR-seq on a number of data-rich ENCODE cell types (Zhang et al, 2020). In yet another study, we integrated ENCODE data to study combinatorial TF binding and co-regulation by constructing a regulatory network (this was carried out by finding the genomic binding information of 119 transcription-related factors in over 450 distinct experiments) (Gerstein et al, 2012). We also used PsychENCODE to construct a comprehensive functional genomic resource and integrative model for the human brain (this was built using an extremely diverse set of data types from 1866 individuals) (Wang et al, 2018).