**Gerstein Lab experience with using machine learning to predict gene expression changes as a result of therapeutic compounds**

In PsychENCODE, we used single-cell brain data to identify cell type-specific DEGs for SCZ, BD, and aging (Emani et al, 2024). A machine learning model that predicts global expression changes as a result of drug exposure would thus be valuable for identifying specific compounds that are effective for treating diseases of the brain. We recently constructed precisely this type of model, and we successfully applied it in other neuro-psychiatric disease contexts. Among other tasks, this model (termed LNCTP) simulates perturbations of select genes (for example, the perturbations directly induced in a drug target gene as a result of drug exposure) and then predicts the resulting changes in the overall expression of other genes.

**Gerstein Lab experience with expression QTLs in distinct cell types**

In recent work, we studied expression variation across genotypes at single-cell resolution to identify scQTLs (Emani et al, 2024) in 17 cell types. We used this set of scQTLs as our “core callset” to enable consistent comparisons with existing datasets (such as GTEx and PsychENCODE bulk data). We identified ~1.4 million scQTLs (Fig. 1A), many of which are unique to a particular single cell type (~47% appear in more than one cell type. 30% of the scQTLs overlap with bulk cis-eQTLs (Wang et al, 2018), and we validated some of our core scQTLs by comparing them with functional elements identified by STARR-seq, mutation STARR-seq, and massively parallel reporter assays (MPRA) (Gaynor-Gillett et al, 2024). As further validation, we examined allele-specific expression at the single-cell level in samples with WGS-based phased variants, and found significant agreement. To address low power in rare cell types (Miller et al, 2018)), we developed a Bayesian linear mixed-effects model to find more scQTLs as an additional callset. We also identified QTLs based on linkage-disequilibrium pruning, regression across pseudotime trajectories, and conditional analysis, in addition to cell type–specific isoform-usage QTLs (iso-QTLs). In all, we identified 330 scQTLs for eGenes related to brain disorders.

We have also performed a series of empirical power analyses to evaluate the sample sizes needed to attain sufficient power for QTL detection. This was performed by randomly subsampling SCORCH data into variously-sized subsets, and then measuring how power scales with sub-sample size. Our preliminary results using this scheme on PsychENCODE data (see Fig. 1B) suggest that ~150 samples may identify a sufficiently large number of scQTLs for downstream studies, such as integration into GRNs.



**Fig. 1: (A) UpSet plot with identified scQTLs (from the core analysis) that are unique to individual cell types (red) or across all cell types (blue). Left histogram summarizes the number of core scQTLs per cell type. Right histogram summarizes the log-scaled total number of Bayesian scQTLs per cell type. (B) Results from empirical power analysis in QTL calculations based on randomly sub-sampling data from the PsychENCODE dataset. Exhibits adapted from Emani et al.\cite{PMID 38781369}**

**Gerstein Lab experience with integrating multiple data modalities to build gene regulatory networks (GRNs) for neurogenomics**

In our recent PsychENCODE work (Emani et al, 2024), we processed single-nuclei multi-omics datasets from the PFC from 388 donors (control and neuropsychiatric disorders) and integrated multiple data modalities (including scQTLs, snATAC-seq, TF-binding sites, and gene coexpression) to build GRNs for each cell type. These GRNs are directed unweighted graphs with edges connecting TFs to target genes. To investigate co-regulation among known disease risk genes, we converted directed GRNs into undirected networks that include connections between targets. We found that SCZ-associated genes form relatively dense subnetworks in neurons, whereas the AD subnetwork is actively co-regulated in microglia and immune cells. We also evaluated the enrichment of disease co-expression modules within the regulons that we identified as part of this analysis, and found that our approach recapitulates disease modules. These GRNs elucidate the interactions between TFs, enhancers, and target genes, offering a dynamic view of gene regulation across cell types in healthy and diseased brain states.

**Gerstein Lab experience with spatial analysis in neurogenomics contexts**

We have experience working with spatial data in related contexts. Cell types are often organized along an ordered lineage or “trajectory” (for example, cell types in the lineage of different cortical layers; see Fig. 2A). In these cases, one limitation of treating cell types as discrete entities along the lineage is that such an approach does not take advantage of the continuous nature of expression changes as one progresses through the lineage. We thus used PsychENCODE data to treat expression changes in a continuous nature through an ordered lineage using trajectory analysis that captures the spatial relationships among cell types in the PFC (Emani et al, 2024). In particular, we embedded cells into lower-dimensional space to build smooth trajectories, and then used these to find patterns of expression variation along the cortical-depth axis to identify genes with significant variation across the layers (Fig. 2B). In the same study, we also developed a regression model that uses a continuous trajectory to identify "dynamic scQTLs" that exhibit changing effect sizes along the trajectory.

****

**Fig. 2: (A) Physiological context in which trajectory analysis can be performed is provided by different cortical layers, wherein different cell types appear in a spatially-ordered lineage. (B) UMAP showing the predicted trajectory for excitatory IT neurons. Panel B adapted from Emani et al.**

**References:**

Emani, Prashant S., et al. "Single-cell genomics and regulatory networks for 388 human brains." Science 384.6698 (2024): eadi5199.

Gaynor-Gillett, Sophia C., et al. "Validation of Enhancer Regions in Primary Human Neural Progenitor Cells using Capture STARR-seq." bioRxiv (2024).

Miller, Kai J., et al. "Cortical activity during motor execution, motor imagery, and imagery-based online feedback." Proceedings of the National Academy of Sciences 107.9 (2010): 4430-4435.

Wang, Daifeng, et al. "Comprehensive functional genomic resource and integrative model for the human brain." Science 362.6420 (2018): eaat8464.