Gerstein's lab experience in Multi-modal modeling of cell-specific and time-dynamic features of Alzheimer's disease

Experience in analyzing the impact of genetic variation

We have extensive experience with quantifying the impact of genetic variants from various layers of information. For instance, we developed a variant-prioritization pipeline named FunSeq that has been widely used to identify disease-causing mutations for further in-depth analyses to understand the mechanisms underlying disease pathogenesis **[1,2]**. We also developed GRAM **[3]**, a generalized model to predict cell-type-specific molecular effects of non-coding variants on their associated genes, and AlleleSeq, a tool for detecting candidate variants associated with allele-specific binding and allele-specific expression **[4-6]**. To probe the functional effects of genetic variation and the mechanistic underpinnings of disease, we previously built integrated models that relate molecular- and cellular-level phenotypes to high-level traits. In our previous work, we developed an interpretable integrated modeling framework for this purpose, within the context of psychiatric genomics **[7]**. Our Deep Structured Phenotype Network (DSPN) framework allowed us to model the joint distribution of all phenotypes of interest conditioned on genetic variation; a joint energy function enabled us to embed prior knowledge in the connectivity of the network and interpret new relationships during and after training. We used a conditional deep Boltzmann machine architecture with multiple layers, including genotype, gene expression, epigenetics, and cell fraction layers, and introduced lateral connectivity at the visible layer to embed the gene regulatory network (GRN) and quantitative trait locus (QTL) linkages. Further, we developed a rank-statistic-based interpretation scheme that allows us to functionally annotate hidden nodes and prioritize them relative to disorders **[8]**. Our model improved disease prediction by 6-fold compared to additive polygenic risk scores for schizophrenia, highlighted key genes for schizophrenia and other disorders, and allowed imputation of missing transcriptome information from genotype alone **[7]**. Finally, we also have extensive experience in extracting latent signatures from gene expression data as biomarkers for asthma **[9],** and in integrating 3D protein structures and dynamics with mutational frequencies to identify cancer driver genes **[10].**

[1] Khurana, Ekta et al. "Integrative annotation of variants from 1092 humans: application to cancer genomics." *Science (New York, N.Y.)* vol. 342,6154 (2013): 1235587. doi:10.1126/science.1235587

[2] Fu, Yao et al. "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer." *Genome biology* vol. 15,10 (2014): 480. doi:10.1186/s13059-014-0480-5

[3] Lou, Shaoke et al. "GRAM: A GeneRAlized Model to predict the molecular effect of a non-coding variant in a cell-type specific manner." *PLoS genetics* vol. 15,8 e1007860. 30 Aug. 2019, doi:10.1371/journal.pgen.1007860

[4] Rozowsky, Joel et al. "AlleleSeq: analysis of allele-specific expression and binding in a network framework." *Molecular systems biology* vol. 7 522. 2 Aug. 2011, doi:10.1038/msb.2011.54

[5] Chen, Jieming et al. "A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals." Nature communications vol. 7 11101. 18 Apr. 2016, doi:10.1038/ncomms11101

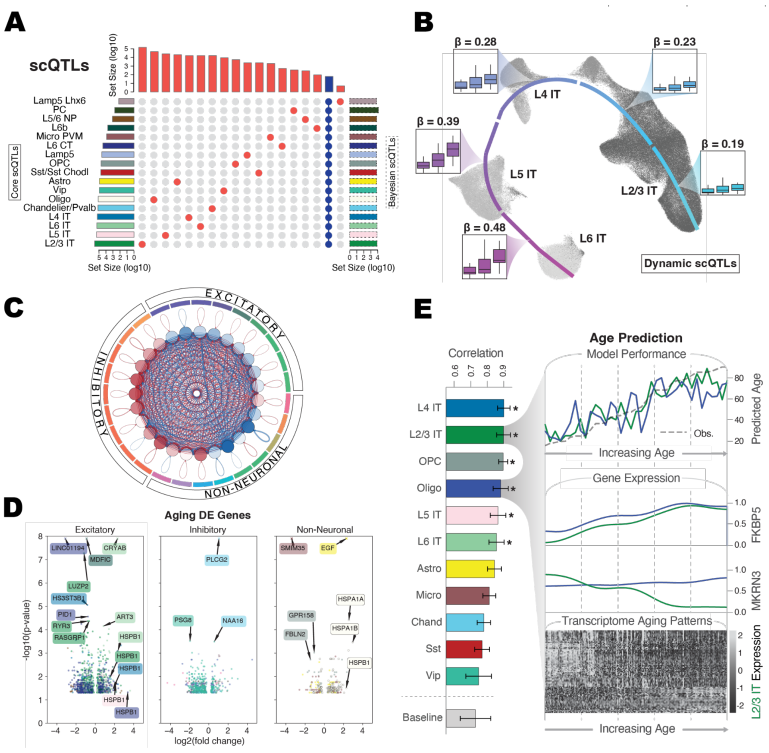
[6] Onuchic, Vitor et al. "Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci." Science (New York, N.Y.) vol. 361,6409 (2018): eaar3146. doi:10.1126/science.aar3146

[7] Wang, Daifeng et al. "Comprehensive functional genomic resource and integrative model for the human brain." Science (New York, N.Y.) vol. 362,6420 (2018): eaat8464. doi:10.1126/science.aat8464

[8] Warrell, Jonathan et al. “Rank Projection Trees for Multilevel Neural Network Interpretation.” arXiv, https://doi.org/10.48550/arXiv.1812.00172
[9] Lou, Shaoke et al. “Latent-space embedding of expression data identifies gene signatures from sputum samples of asthmatic patients.” BMC bioinformatics vol. 21,1 457. 15 Oct. 2020, doi:10.1186/s12859-020-03785-y
[10] Kumar, Sushant et al. “Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures.” Proceedings of the National Academy of Sciences of the United States of America vol. 116,38 (2019): 18962-18970. doi:10.1073/pnas.1901156116

Experience in single-cell characterization of the human brain

In the framework of the PsychENCODE consortium we previously created brainSCOPE (brain Single-Cell Omics for PsychENCODE) **[1]** a uniformly processed single-cell (sc) resource which comprises snRNA-seq, snATAC-seq, and snMultiome data from >2.8 M nuclei of 388 individual brains (182 healthy controls and 206 individuals with neurological disorders). Our cell-type annotation scheme, which harmonizes several published analyses in the prefrontal cortex including the BICCN reference atlas, identified 28 cell types. We used our processed snRNA-seq data in combination with these 28 cell types to identify single-cell *cis*-eQTLs (scQTLs). By employing the same analytical strategy as the GTEx consortium **[2],** we identified a “core callset” of >1.4M single-cell eQTLs, with an average of ~85K *cis*-eQTLs and ~690 eGenes per cell type. To overcome the low statistical power that characterizes the rarer cell types, we complemented this core set of scQTLs with an additional callset derived from a Bayesian linear mixed-effects QTL mapping model (Fig. **1A**). Overall, we identified 330 scQTLs for eGenes related to brain disorders. Some of these variants map to the 17q21.31 locus, previously associated to brain disorders, such as an astrocyte-specific scQTL for the Tau protein gene *MAPT* and a multi-cell type scQTL for the neurodegenerative-disorder risk gene *KANSL1* **[3]**. Furthermore, we developed a Poisson-regression model that incorporates a continuous trajectory and a pseudotime-genotype interaction term, which we used to identify “dynamic scQTLs”, i.e. eQTLs that change effect size along the pseudotime trajectory (**Fig. 2B**) **[4]**.



We also combined our snRNA-seq data with publicly available ligand-receptor pairs to construct a cell-to-cell interaction network **[5]**. Our analyses highlighted three broad ligand-receptor usage patterns that distinguish excitatory, inhibitory, and glial cell types. We also investigated how cell-cell communication patterns are altered in neuropsychiatric disorders. We found that individuals with schizophrenia and bipolar disorder showed notable inter-mixings among the three broad patterns of ligand-receptor usage. For instance, when

comparing individuals with schizophrenia with healthy controls, we found that excitatory neurons received less incoming signaling, while inhibitory neurons received more (Fig. **2C**).

We also used our population-scale single-cell data to systematically assess transcriptomic and epigenetic changes due to aging. First, we identified a list of aging DE genes across cell types (**Fig. 2D**). We found, for instance, that *HSPB1*, which encodes a heat-shock protein and has been previously implicated in longevity, is upregulated in multiple cell types in older individuals **[6-7]**. Additionally, we constructed a model to predict an individual's age from their single-cell expression data (**Fig. 2E**). We generated cell-type specific pseudo-bulk expression matrices derived from snRNA-Seq and used XGBoost to predict the age of each individual. Formally, we define X as the expression matrix, where each row x_i corresponds to a specific cell's transcriptomic profile. y represents the actual age labels, our predictive model aims to minimize the following objective function:

$$L(X, y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_\theta(X_i))^2$$

where N is the number of samples, $f_\theta(\cdot)$ is the model function parameterized by θ. To gauge the efficacy and precision of our models, we implemented multiple evaluation metrics: Pearson correlation, Spearman correlation, mean absolute error, and root mean squared error. To gain more biological insights to the model predictions, we adapted the SHAP module in order to decipher how individual features of the model (e.g., gene expression for a particular cell type) impact accuracy. The model shows that the transcriptomes of six cell types (L2/3 IT, L4 IT, L5 IT, L6 IT, Oligodendrocytes, and OPC) have strong predictive value. It also shows that many individual genes contribute to the model, highlighting broad transcriptome changes in aging. From these, we selected two particularly predictive genes previously associated with aging, *FKBP5* and *MKRN3*, and observed a clear correlation between their expression and aging (**Fig. 2E**) **[8-10].** Finally, we also investigated the effects of age on the epigenome using single-cell candidate cis-regulatory elements (scCREs) to deconvolve bulk chromatin accessibility for 628 individuals into those for specific cell types. The resulting scCRE activity patterns in certain cell types, particularly microglia, cluster individuals into distinct age groups. We further expanded our analysis to highlight how patterns of enriched TF motifs in active scCREs change with age in a cell-type-specific fashion. Some TFs demonstrate consistent patterns across cell types (*FOXO4* and *RXRA)*, while others exhibit more cell-type-specific patterns (*NEUROG1*).

[1] Emani, Prashant S et al. “Single-cell genomics and regulatory networks for 388 human brains.” Science (New York, N.Y.) vol. 384,6698 (2024): eadi5199. doi:10.1126/science.adi5199
[2] GTEx Consortium. “The GTEx Consortium atlas of genetic regulatory effects across human tissues.” Science (New York, N.Y.) vol. 369,6509 (2020): 1318-1330. doi:10.1126/science.aaz1776
[3] Cooper, Yonatan A et al. “Functional regulatory variants implicate distinct transcriptional networks in dementia.” Science (New York, N.Y.) vol. 377,6608 (2022): eabi8654. doi:10.1126/science.abi8654
[4] Nathan, Aparna et al. “Single-cell eQTL models reveal dynamic T cell state dependence of disease loci.” Nature vol. 606,7912 (2022): 120-128. doi:10.1038/s41586-022-04713-1
[5] Jin, Suoqin et al. “Inference and analysis of cell-cell communication using CellChat.” *Nature communications* vol. 12,1 1088. 17 Feb. 2021, doi:10.1038/s41467-021-21246-9
[6] Gomez, Christian R. “Role of heat shock proteins in aging and chronic inflammatory diseases.” GeroScience vol. 43,5 (2021): 2515-2532. doi:10.1007/s11357-021-00394-2

[7] Cross-Disorder Group of the Psychiatric Genomics Consortium. “Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis.” Lancet (London, England) vol. 381,9875 (2013): 1371-1379. doi:10.1016/S0140-6736(12)62129-1
[8] Macedo, Delanie B et al. “Central Precocious Puberty Caused by a Heterozygous Deletion in the MKRN3 Promoter Region.” Neuroendocrinology vol. 107,2 (2018): 127-132. doi:10.1159/000490059
[9] Zannas, Anthony S et al. “Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB-driven inflammation and cardiovascular risk.” Proceedings of the National Academy of Sciences of the United States of America vol. 116,23 (2019): 11370-11379. doi:10.1073/pnas.1816847116
[10] Zannas, Anthony S et al. “Gene-Stress-Epigenetic Regulation of FKBP5: Clinical and Translational Implications.” Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology vol. 41,1 (2016): 261-74. doi:10.1038/npp.2015.235