**Gerstein lab experience with *Interpretable Deep Learning Models to Characterize Cell-type-specific Variant Impact in Alzheimer’s Disease***

We have extensive experience in processing large-scale single cell data. Specifically, we are the lead analysis group in pioneering consortia, including ENCODE, PsychENCODE, and SCORCH, where we have uniformly processed population-scale single-cell transcriptomic data from over 300 postmortem brain regions. Our cell-type-specific gene expression signatures have been widely used in the PsychENCODE to prioritize risk factors in multiple brain disorders.

Our team has developed several tools designed for single cell analysis. For instance, our Forest-Fire method combine iterative label propagation and parallelized Monte-Carlo simulation for scalable single cell data clustering with self-validations, which has been recently accepted by *Nature Communications*. On the ATAC-seq side, we have developed SAILER1 (**Fig. 2**), a deep VAE to learn low-dimensional nonlinear cell representations that define the intrinsic chromatin state, invariant to extrinsic confounding factors like read depth and batch effects. Experimental results on both simulated and real scATAC-seq datasets demonstrate that SAILER learns better and more meaningful representations of cells than other methods.

Our team has performed snRNA-seq from frozen postmortem DLPFC from 120 neurotypical samples, resulting in 1.2 million single-nuclei. The resulting datasets capture the diversity of cell types in the human brain. Data across individuals are highly concordant, with the same cellular populations consistently rediscovered (**Fig 3 A-B**). We discovered meaningful clusters in the ~1.2M nuclei, analyzed jointly across individuals and across phenotypes, enabling us to then study phenotype-associated changes within and across clusters. We also processed ~0.5M nuclei from snATAC-seq data and generated clear clusters using our SAILER pipelines (**Fig. 3C**). In addition, we ran single-cell LDSC analysis on cell type specific peaks and microglia showed the strongest signal to explain AD GWAS summary statistics (**Fig. 3D**).

To test our multiome data model, we downloaded mouse brain SNARE-seq data and compared it with existing methods. From **Fig. 4**, we can see some integration methods severely suffer from this when projecting data from two modalities into one shared latent space. However, embedding generated by SCAN-ART forms tighter clusters (**Fig. 4A**) and achieves the best performance in terms of quantitative results (**Fig. 4C**). The separation of cell types is also demonstrated by marker gene activity score of cells identified by different methods (**Fig. 4B**), where SCAN-ART shows significantly better results compared with other methods.

Our investigator team has been working together for years to build the official ENCODE22,33 and PsychENCODE48 annotation resources via bulk tissue and single-cell multi-omics integration. Specifically, we constructed gene-centric annotations by linking CREs to genes and constructing TRNs. In addition, we have extensive experience in using Hi-C to validate CRE-gene linkages2, and much previous experience with network analysis framework development3-5. We have processed cell-type-specific Hi-C in neural and microglial cells2 to validate our Direct-net CRE-gene linkage predictions.

We applied Direct-net on a GM12878 scATAC-seq dataset to divide CREs into three groups: high confidence (**HF**), low confidence (**LF**), and between HF and LF (**REST**). The HF CREs showed the highest ChIA-PET validation rate (**Fig. 6A1**), larger rare variant enrichment (HF vs. LF: 0.846 vs. 0.838, p=1e-13, **Fig. 6A2**), and stronger H3K27ac signals than the other groups (**Fig. 6A3**). We further benchmarked with Cicero6 and found that Direct-net showed better validation rates (3.4 vs. 2.9 in ChIA-PET, 1.11 vs. 0.93 in Hi-C,**Fig. 6A4-6**). In addition, Direct-net recovered more proximal ChIA-PET and Hi-C connections than Cicero (0.45 vs. 0.36 for ChIA-PET, 0.31 vs. 0.25 for Hi-C, **Fig. 6A5-7**), validating the higher sensitivity of Direct-net.

We applied Direct-net on peripheral blood mononuclear cell (PBMC) sc-multiome data. We found that HF CREs can differentiate cell types, while the REST regions failed with certain mixed cell types and LF regions mixed most cells, which is quantitatively reflected by the Silhouette scores (**Fig. 6B1-4**). In addition, we found that the HF CRE-gene links were better validated by cell-type-matched promoter capture Hi-C data7 than LF links (**Fig. 6B5**), and the HF CREs showed higher rare variant fractions and PhastCons scores8 than the others (**Fig. 6B6-7**). Furthermore, we selected CREs associated with marker genes and found that HF CREs exhibited the strongest H3K27ac signals from matched cell types, while the LF regions exhibited no signals. Moreover, bulk PBMC H3K27ac signals in HF CREs were lower than those from matched cells (**Fig. 6B8-10**), validating Direct-net’s ability to reveal cell-type-specific regulatory mechanisms.

We have extensive experience in quantifying TF and variant impacts in disease studies. For instance, we developed a variant-prioritization pipeline named FunSeq that included an adjustable data context 9,10. This tool has been widely used to identify disease-causing mutations for further in-depth analyses to understand the mechanisms underlying disease pathogenesis. FunSeq links each noncoding mutation to target genes, and prioritizes such variants based on functional annotation, sequence features, conservation, network connectivity, and mutation frequency in diseases. We also developed a generalized model named GRAM to predict cell-type-specific molecular effects of non-coding variants on their associated genes 11. This tool has been used to predict the effects of fine-mapping causal variants from genome-wide association studies. Finally, we developed a variant-scoring framework named RADAR to pinpoint variants associated with RNA binding protein function dysregulation 12. In addition, we developed AlleleSeq, a tool for detecting candidate variants associated with allele-specific binding and allele-specific expression 13-15.

To fully characterize genotype-phenotype association, we harmonized bulk transcriptome and their genotype information for 1866 individuals with large-scale functional assays (e.g., ChIP-seq, Hi-C, and scRNA-seq). First, we created a comprehensive brain regulatory map, including enhancers and their targets, various quantitative-trait loci (e.g., expression, isoform, cell fraction, and chromatin QTLs), and gene regulatory networks. Then, we embedded such regulatory information into a deep-learning model to predict psychiatric phenotypes from genotype and transcriptome. Our DSPN model gives a ~6-fold improvement in prediction over additive polygenic risk scores (**Table 2**). Lastly, it highlights key genes and pathways associated with disorder prediction, including immunological, synaptic, and metabolic pathways, recapitulating *de novo* results from more targeted analyses.

1 Cao, Y. *et al.* SAILER: scalable and accurate invariant representation learning for single-cell ATAC-seq processing and integration. *Bioinformatics* **37**, i317-i326 (2021). <https://doi.org:10.1093/bioinformatics/btab303>

2 Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).

3 Lou, S. *et al.* TopicNet: a framework for measuring transcriptional regulatory network change. *Bioinformatics* **36**, i474-i481 (2020). <https://doi.org:10.1093/bioinformatics/btaa403>

4 Yan, K. K., Wang, D., Xiong, K. & Gerstein, M. Comparing Technological Development and Biological Evolution from a Network Perspective. *Cell Syst* **10**, 219-222 (2020). <https://doi.org:10.1016/j.cels.2020.02.004>

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

6 Pliner, H. A. *et al.* Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell* **71**, 858-871 (2018). <https://doi.org:10.1016/j.molcel.2018.06.044>

7 Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e1319 (2016). <https://doi.org:10.1016/j.cell.2016.09.037>

8 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005). <https://doi.org:10.1101/gr.3715005>

9 Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome biology* **15**, 480 (2014). <https://doi.org:10.1186/s13059-014-0480-5>

10 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013). <https://doi.org:10.1126/science.1235587>

11 Lou, S. *et al.* GRAM: A GeneRAlized Model to predict the molecular effect of a non-coding variant in a cell-type specific manner. *PLoS Genet* **15**, e1007860 (2019). <https://doi.org:10.1371/journal.pgen.1007860>

12 Zhang, J. *et al.* RADAR: annotation and prioritization of variants in the post-transcriptional regulome of RNA-binding proteins. *Genome biology* **21**, 151 (2020). <https://doi.org:10.1186/s13059-020-01979-4>

13 Onuchic, V. *et al.* Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. *Science* **361** (2018). <https://doi.org:10.1126/science.aar3146>

14 Chen, J. *et al.* A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun* **7**, 11101 (2016). <https://doi.org:10.1038/ncomms11101>

15 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011). <https://doi.org:10.1038/msb.2011.54>