Gerstein Lab Experience in Integrative Gene Regulatory Network Construction and Analysis

We have considerable experience in studies detailed the construction of gene regulatory networks by integrating eQTLs and cell-specific regulatory elements to map interactions between transcription factors, enhancers, and target genes, enhancing understanding and predictive modeling of psychiatric and cellular phenotypes (PMIDs: 30545857, 25053837, 38562822, 22125477, 22955619). For a specific example, we constructed detailed GRNs for each cell type by integrating information from the identified eQTLs and the cell-type-specific regulatory elements. These networks elucidate the interactions between transcription factors, enhancers, and target genes, offering a dynamic view of gene regulation across different cell types and conditions. Leveraging the regulatory and expression data, we developed networks depicting cell-to-cell communication within the brain. These networks reveal how cells interact molecularly, shifting the understanding of cellular interactions in both healthy and diseased brain states. Besides, an integrative model Linear Network of Cell Type Phenotypes (LNCTP) was developed to simulate and predict the effects of genetic perturbations on cell-type-specific gene expression. This model aids in prioritizing disease-risk genes and potential drug targets, significantly enhancing the translational impact of their findings. All sequencing data, derived analysis files, and computational codes have been made available through the brainSCOPE resource portal (http://brainscope.psychencode.org/). This portal serves as a valuable tool for the broader scientific community, facilitating further research and validation of the findings [PMID: 38562822].

We developed and published computational tools for analyzing and understanding regulatory networks and identifying hubs and bottlenecks, which are helpful for investigating associations between neuronal development and diseases with gene regulation effects. In the paper "Construction and Analysis of an Integrated Regulatory Network Derived from High-Throughput Sequencing Data," we develop a comprehensive framework for understanding the complex interactions within gene regulatory networks by integrating high-throughput sequencing data across various levels of regulation. The study presents an integrated regulatory network (IRN) that encompasses transcription factors (TFs), microRNAs (miRNAs), and their respective gene targets. This network reveals the complex interplay between different types of regulatory molecules, providing insights into both transcriptional and post-transcriptional regulation [PMID: 22125477]. The integrated network combines data from ChIP-Seq, which identifies TF binding sites, and RNA-Seq, which profiles expression levels, creating a multifaceted view of gene regulation. The paper emphasizes the novel use of this integrated approach to map regulatory interactions in a system-wide manner, marking a significant advance over previous studies that typically focused on one regulatory level at a time. By doing this, our team was able to uncover new insights into how TFs and miRNAs co-regulate genes and participate in complex regulatory circuits that maintain cellular functions [PMID: 22125477].

In another work, we developed advanced computational tools to analyze the human regulatory network using ENCODE project data. They tackled the complex task of understanding how transcription factors (TFs), which regulate gene expressions, interact within a vast network to control genetic activities across different cells and conditions. The research focused on detailing the architecture of this network by identifying and mapping the genomic binding sites of 119 TFs across various cell types. The team utilized ChIP-Seq data to establish where these TFs bind in the genome, leading to the discovery of combinatorial and highly context-specific associations between different TFs. These findings are critical as they show that TF binding is not random but rather highly orchestrated, with specific TF combinations binding at unique locations to execute cellular functions. Our group organized this complex

interaction data into a hierarchical structure that revealed new insights into the regulatory landscape of human genes. They integrated other types of genomic data, such as microRNA (miRNA) interactions and protein interactions, creating a comprehensive meta-network that illustrates how different levels of regulation interact to influence gene expression [PMID: 22955619].

We integrated the genomic elements into a regulatory network. We first processed adult brain Hi-C data, identifying 2735 topologically associating domains (TADs) and approximately 90,000 enhancer-promoter interactions. As expected, the majority of interactions occur within the same TAD, and the genes with more enhancers tend to have higher expression levels. Upon incorporating QTLs, we observed that QTLs linked distally to expressed genes (eGenes) via Hi-C interactions had significantly stronger associations compared to those with SNPs directly within eGene promoters or exons. Additionally, we explored the regulatory connections between transcription factor (TF) activity and target gene expression using elastic net regression. Overall, our work generated a comprehensive regulatory network linking enhancers, TFs, and target genes. This network includes 43,181 proximal and 42,681 distal connections involving 11,573 protein-encoding target genes. The proximal connections TFs to target genes via promoters, whereas the distal connections link TFs to target genes via enhancers [PMID: 25053837]. We also generated potential cell-type-specific regulatory networks. In these networks, several well-known TFs were identified for association with brain development, such as NEUROG1, DLGAP2, and MEF2A for excitatory neurons and GAD1, GAD2, and LHX6 for inhibitory neurons [PMID: 10640277; PMID: 28870203; PMID: 2069816; PMID: 17376969]. These networks not only enhance our understanding of brain development but also serve as a valuable resource for further research. This resource was published in the journal Science in 2018.

In the paper titled "Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights," we pioneer an innovative approach to enhancing the understanding of protein interaction networks through the integration of structural biology with network analysis. They meticulously developed a method to annotate interactions within a consensus yeast interaction network using atomic-resolution data from three-dimensional protein structures. This led to the creation of a Structural Interaction Network (SIN), which classifies interactions as either mutually exclusive or simultaneously possible, based on whether interaction partners bind to the same or different interfaces on a common protein. This methodological innovation provides a more nuanced view of protein networks, distinguishing between different types of protein interactions that were previously undifferentiated. By analyzing these interactions in terms of their structural and functional implications, our team was able to derive deeper insights into the dynamics of protein networks. For instance, they discovered that proteins connected by simultaneously possible interactions often share similar functions and are co-expressed, suggesting stable, possibly permanent, associations. Conversely, mutually exclusive interactions are indicative of more transient interactions within the network. The research elucidates the physical constraints on protein interactions, highlighting how structural properties can influence biological function and evolution. This work significantly advances our understanding of how protein structures shape the behavior of biological networks and provides a valuable framework for predicting protein interactions based on structural data [PMID: 17185604].

Our team has previously developed many tools to visualize important genes, regulatory networks and latent representations of single cells. For instance, we have extensive knowledge in building a web-based platform that contains a wide range of analysis tools for researchers, the NeMO Analytics. NeMO Analytics is an online platform that is well organized and provides

easily accessible landing pages, which currently supports ~600 registered users, hosts more than 1000 datasets, assists 45 genomics-based publications and ~150 data collections. In response to the demand for new visualization tools, we introduced new displays on NeMO Analytics to handle spatial data and patch-seq data, which includes electrophysiological and morphological data alongside gene expression [PMID: 34616075]. For instance, a publication describing the mouse cortex where the figures are linked back to NeMO Analytics demonstrates our support - readers can click on the link and instantly explore the data by gene expression or cell type. The NeMO Analytics supports following six main functions: (1) dataset uploader and curator for uploading new data and visualizing of the data, which includes interactive bar, line, scatter or violin plots; colorized anatomical graphical representations of the data; and ordination plots such as UMAP, tSNE and PCA plots; (2) dataset manager, which is used to group and arrange datasets in profiles, and to specify which individual or groups of datasets are shared; (3) Integration with Epiviz [PMID: 31782758], which is a browser that provides simultaneous visualization of epigenetic information, genome accessibility and the expression results; (4) gene expression browser for displaying gene expression that also provides value-added resources such as functional annotation and link to relevant coding resources; (5) workbench that perform de novo analyses from new single cell data or start with existing data, and perform additional analyses and visualizations. The single cell workbench is largely based on the Scanpy [3] pipeline implemented in a graphical user interface (GUI) and performs better than R-based pipelines; and (6) comparer to analyze expression between any two conditions (e.g., phenotypic groups, timepoints) of the same dataset. This platform has supported several BRAIN initiative publications [PMIDs: 34616075; 34616062; 34616066; 37939194; 31996853].

Another area of our expertise is application of various sophisticated tools to map intricate relationships in the biological systems. These models have been particularly effective in analyzing microbial communities and their metabolic pathways, demonstrating the capability to correlate environmental factors with biological data, which can help understand the metabolic impacts on brain functions and disorders (PMID: 19164758). We also developed "Gene Tracer," an innovative voice-controlled tool designed to enhance the interactive querying and visualization of genomic information. This cloud-based approach not only meets the computational demands of processing large genomic datasets but also guarantees that the system remains responsive and accessible to users from any location. This adaptation to technological advances in cloud computing helps address the specific needs of genomic research and showcases a shift towards multi-cloud strategies to optimize resource use and data management in scientific computing [PMID: 33792640].