Submitter Name: Zhana Duren Submitted Email: zduren@clemson.edu

Inferring gene regulatory networks from single cell multiome data using atlas-scale external data

Qiuyue Yuan[1], Fengge Chang[1], and Zhana Duren[1]

[1]Center for Human Genetics, Department of Genetics and Biochemistry, Clemson University, Greenwood, SC 29646, USA

While differential expression of genes can be observed between addicted and controls, the underlying reasons behind these variations remain elusive. It is plausible that these changes are influenced by drug-induced alterations in the activity of transcription factors (TFs) through mechanisms such as post-translational modifications that can impact the localization and function of TFs, potentially leading to their translocation from the cytoplasm to the nucleus, where they can bind to the DNA and regulate gene expression. Gene Regulatory Networks (GRNs) is a systematic approach for linking the observation of differential expression to change of activity of TFs. Existing methods for GRNs inference rely on gene expression data alone, or on lower resolution bulk data. Despite recent integration of ATAC-seq and RNA-seq data, learning complex mechanisms from limited independent data points still presents a daunting challenge. Here we present LINGER (LIfelong neural Network for GEne Regulation), a machine learning method to infer GRNs from single-cell paired gene expression and chromatin accessibility data. LINGER incorporates both atlas-scale external bulk data across diverse cellular contexts and prior knowledge of TF motifs as a manifold regularization. LINGER achieves 4-7-fold relative increase in accuracy over existing methods and reveals a complex regulatory landscape of genome-wide association studies, enabling enhanced interpretation of disease-associated variants and genes. Following the GRN inference from a reference sc-multiome data, LINGER allows for the estimation of TF activity solely from bulk or single-cell gene expression data, leveraging the abundance of available gene expression data to identify driver regulators from case-control studies.