Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation Summary
Doctoral Program in Computer and Control Engineering (37th Cycle)

Multi-omic and multi-modal single-cell sequencing data analysis for investigating transcriptional regulation and cellular heterogeneity

Candidate
Lorenzo Martini
S293116

PhD Supervisor
Prof. Stefano Di Carlo
Prof. Alessandro Savino

Submitted on:
April 30th, 2025

Abstract

This dissertation explores the incredible cellular heterogeneity by employing single-cell sequencing data. The rapid advancement of Next Generation Sequencing (NGS), particularly single-cell technologies like scRNA-seq and scATAC-seq, has improved the possibility of characterizing biological systems at unprecedented resolution. However, these technologies come with challenges in dealing with their data analysis. High dimensionality, noise, and interpretability issues are inherent to them, and this work contributes to addressing them through model-based, biologically informed computational approaches that bridge multiple omic layers.

The first part of the thesis focuses on the multi-omic single-cell sequencing analysis for decoding transcriptional regulation. Specifically, the combination of scRNA-seq and scATAC-seq hides incredible potential in investigating transcriptional regulation through multiple levels. A central contribution is the Genomic-Annotated Gene Activity Matrix (GAGAM), a biology-informed model-driven approach aiming to construct a gene activity matrix by annotating scATAC-seq peaks with cis-regulatory labels and properly weighting their contributions. GAGAM integrates the different contributions based on their functional role in regulating transcription and obtaining activity profiles per gene. The activity obtained is validated across various datasets through clustering and coherence with scRNA-seq expression data, showing a more remarkable ability to correlate the overall accessibility with the gene expression concerning existing approaches.
This model improves the biological interpretability of accessibility data and highlights the importance of the accessibility of distal cis-regulatory regions in modulating transcription.

Furthermore, this thesis presents the cross-omic analysis of Transcription Factors (TF), correlating TF expression from scRNA-seq with motif enrichment and footprinting derived from scATAC-seq by separately investigating the different GAGAM contributions. This analysis provides insights into cell-type-specific accessibility of TFs revealing distinct activity-expression dynamics and regulatory architectures. More importantly, it shows the importance of distal enhancer regions in carrying relevant information in scATAC-seq data.

Finally, the Gene Regulation Accessibility Integrating GeneHancer (GRAIGH) extends the framework by integrating the GeneHancer database to link enhancer accessibility to specific genes. By generating a matrix of GeneHancer element accessibility across cells, GRAIGH allows unsupervised clustering and marker analysis that outperforms traditional peak-based approaches in specificity and resolution.

The second significant contribution of the thesis explores neuronal heterogeneity in the brain cortex, focusing on electrophysiological diversity among GABAergic neurons. The primary approach presented is the multi-modal integration of single-cell sequencing data with neuron electrophysiological data. On this note, the Neuronal Spike Shapes (NSS) method is introduced as a novel approach to capturing excitability states through electrophysiological waveform features. NSS analysis reveals the presence of distinct intra-subtype clusters that identify different electrophysiological states, which are not distinguishable only from the gene expression analysis. Moreover, it highlights the specific repolarization phase and its correlation with potassium channel gene expression to characterize these excitability states, linking the more static scRNA-seq data with dynamic phenotypical changes.

Finally, the Spike Train Scalograms (STS) framework continues the work on NSS but focuses on the train of spikes rather than the single spike. This deep learning-based pipeline extracts features from spike train spectrograms using continuous wavelet transforms and convolutional neural networks. STS achieves high accuracy in classifying neuron types and provides interpretable saliency maps for biological insight.

Together, these contributions demonstrate the power of combining multi-omic and multi-modal single-cell data with biologically informed computational tools.

Additional Information

This Dissertation illustrates the research activities conducted within the PhD Programme in Computer and Control Engineering, at Politecnico di Torino (Italy), and included in several scientific contributions.
The Programme had a total duration of 3 years (1st November 2021—31st October 2024).

In addition, a part of the research work was conducted at The Norwegian Centre for Molecular Biosciences and Medicine (NCMBM) of the University of Oslo, during a secondment in the period January 2024—June 2024, and continued remotely until the end of the PhD. The collaboration focused on the application of Deep Learning model for the analysis of methylation data in Breast Cancer. Critical health issues significantly impacted the final part of the 2024 and, consequently, the completion of the collaboration, which ultimately did not lead to a paper submission or publication. As a result, that work is not included in the dissertation.

I hereby declare that the contents and structure of this dissertation constitute my own original work and do not compromise in any way the rights of third parties, including those relating to the security of personal data.

Turin, 30th April 2025
Lorenzo Martini