Integrating Spatial Transcriptomics in single cell resolution with Explainable Machine Learning for Enhanced Insights in Lung Cancer Biology.

Haris Manousaki[1,2], Panagiotis Xiropotamos[1,2], Charalampos Sinnis[1,2], Theodore Dalamagas[1,2], Georgios K. Georgakilas[1,2]

1 Information Management Systems Institute, "ATHENA" Research and Innovation Center in Information, Communication and Knowledge Technologies

2 Archimedes Research Unit, Athena Research Center

Introduction

In the rapidly advancing field of transcriptomics, Spatial Transcriptomics (SRT) signifies a transformative advancement, particularly within the scope of Precision Medicine. SRT enables gene expression profiling at the single-cell level and captures the location of transcriptional activity within the under-study tissue. Despite SRT's potential, the field faces challenges similar to its precursor, single-cell RNA sequencing (scRNA-seq), including uncertainties related to procedural parameters and biases in cell labeling due to manual marker-gene annotation [1]. Thus, a central focus of this project is to observe the uncertainty in cell classification and uncover insights into cell-to-cell communication using explainable machine learning.

Methods

This study uses non-small cell lung cancer tissue samples obtained from the CosMx platform repository [2], to train machine learning (ML) models for cell classification by integrating SRT gene expression data with both the conventional marker gene approach, utilizing the Seurat package, and an enhanced version provided by Nanostring. By comparing the classification accuracies of these models, we can quantify the uncertainties associated with marker-gene-based labeling. Additionally, we explore the impact of cell-to-cell interactions by initially identifying boundaries of distinct cell populations, using Shannon's entropy, and subsequently comparing the misclassification rate of the ML models between cells that are localized at the boundaries and at the center of populations.

Results

Significant differences in classification performance were observed. The Seurat-mediated marker gene-based annotation yielded a lower Matthews correlation coefficient (MCC 0.6623) and Precision (0.7327) compared to the Nanostring-based annotation (MCC 0.868, Precision 0.8642), highlighting the limitations of marker gene approaches. Furthermore, areas of high entropy were found to align with regional and cell class boundaries, verifying the impact of cell-to-cell communication in gene expression. This observation unveils the potential of our method to extract biological insight from SRT data and leverage explainable ML to explore the tumor microenvironment.

[1] S. Fang *et al.*, “Computational Approaches and Challenges in Spatial Transcriptomics,” *Genomics Proteomics Bioinformatics*, vol. 21, no. 1, p. 24, Feb. 2023, doi: 10.1016/J.GPB.2022.10.001.

[2] “CosMx SMI NSCLC FFPE Dataset | NanoString.” Accessed: Aug. 06, 2024. [Online]. Available: https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset/