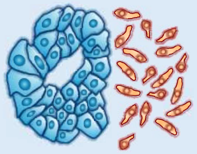
Integrating Deep Learning with Single-Cell Transcriptomics to Resolve Tumor Cell Subtypes

Amanda Warren[1], Dr. Rahul Gomes[2], Dr. Rick J. Jansen[3]
[1]Brigham Young University – Provo, Bioinformatics, [2]University of Wisconsin-Eau Claire, Computer Science,
[3]University of Minnesota, Minneapolis, Masonic Cancer Center

Introduction

Pancreatic cancer is the third leading cause of cancer-related death. Tumor heterogeneity makes treatment challenging, since variation among cancer cell types results in different responses. Classifying tumor cells into known subtypes, such as classical and basal-like, may improve targeted therapy.

Classical Type:
- **Differentiated, epithelial-like**
- **Associated with better prognosis and therapy responsiveness**
- **More structured appearance**



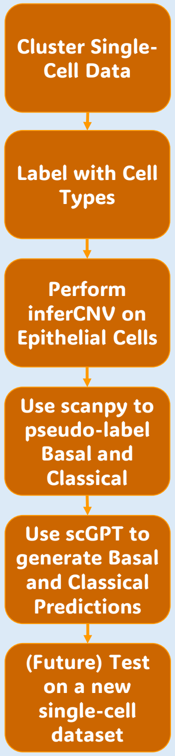
Basal-like Type:
- **Less differentiated, mesenchymal-like**
- **Associated with poorer prognosis, aggressiveness**
- **Disorganized, glandular appearance**

Purpose

This project aims to develop a deep learning pipeline to classify pancreatic cancer cell subtypes in single-cell RNA data.

Our approach provides a more consistent, scalable, and generalizable alternative to manual annotation or clustering methods.

Improved subtype classification could enhance our understanding of intra-tumoral heterogeneity, leading to more treatment options.



Cluster Single-Cell Data

Label with Cell Types

Perform inferCNV on Epithelial Cells

Use scanpy to pseudo-label Basal and Classical

Use scGPT to generate Basal and Classical Predictions

(Future) Test on a new single-cell dataset

What is scGPT?

scGPT is a large generative model pretrained on 33 million human single-cell transcriptomes. Like ChatGPT processes language, scGPT interprets gene expression data. In this project, we fine-tune scGPT using labeled single-cell pancreatic ductal adenocarcinoma (PDAC) datasets to classify malignant epithelial cells into classical and basal-like subtypes.

Methods

Preprocessing

Using the public GEO dataset GSE205013, we tested 45 different quality control parameter combinations.

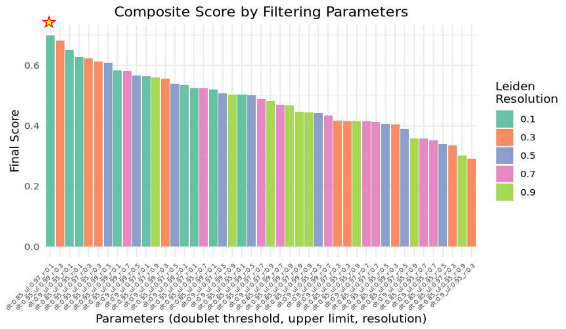


Fig. 1: Composite Score of Filtering Parameters

Clustering

Cells were labeled as epithelial cells or other types through manual curation based on marker gene expression as described in Werba et al. [1].

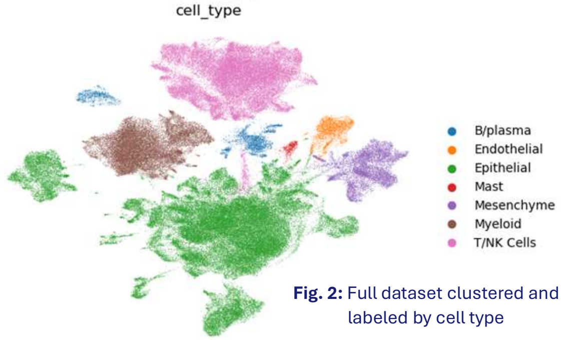


Fig. 2: Full dataset clustered and labeled by cell type

Subtyping of Malignant Cells

We first applied inferCNV to epithelial cells to distinguish malignant from normal populations. To subtype the malignant cells, we used scanpy.tl.score_genes with curated marker genes from Moffitt et al. [2]. This generated expression scores for each subtype by comparing marker gene expression to matched controls. Cells were assigned pseudo-labels based on their highest-scoring subtype.

Downstream Analysis

A scGPT pipeline was developed to fine-tune the pretrained model using scGPT's cell annotation tutorial and the subset of labeled data.

Results

The dataset was split into training, validation, and testing. A total of 44,739 cells were used in training, and 4,972 cells were used to update the metrics as validation. A separate holdout test set of 21,305 cells were used to evaluate model performance.

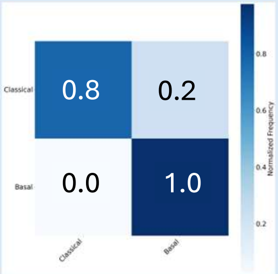


Fig. 4 : scGPT Confusion Matrix

Accuracy	Precision	Recall	F1 Score
0.9209	0.9201	0.8747	0.8939

Table 1: Test Scores

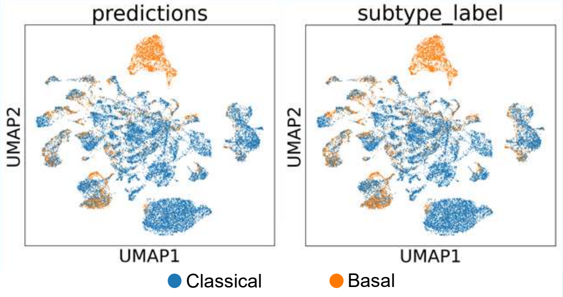


Fig. 5 : UMAP visualization showing alignment between predicted and manually labeled tumor subtypes

Discussion

Future work should evaluate the generalizability of scGPT across PDAC datasets. The classification results can facilitate further studies of tumor cell plasticity or metabolic heterogeneity. Moreover, investigating the relationship between basal-to-classical ratio and clinical outcomes may offer valuable prognostic insight.

Acknowledgements

This research was funded by the National Science Foundation (NSF) Research Experience for Undergraduates (REU) grant OAC- 2447779. The computational resources of this study were provided by the Blugold Center for High-Performance Computing under NSF grant CNS-1920220. Partial funding provided by the Biostatistics Shared Resource of the Masonic Cancer Center, supported in part by National Cancer Institute Cancer Center Support grant P30CA077598 and the National Center for Advancing Translational Sciences of the National Institutes of Health Award Number UM1TR004405.

References

1. Werba G, et al. Nature Communications. 2019;10:1234.
2. Moffitt RA, et al. Nature Genetics. 2015;47(2):116–25.