MSc Data Science & Information Technologies

Bioinformatics - Biomedical Data Science Specialization

Course: Machine Learning in Computational Biology

Professor: Elias Manolakos

MLCB 2024 Final Project Proposal

Authors	*Student ID*
Konstantinos Giatras	7115152300005
Olympia Tsiomou	7115152200035

DEPARTMENT OF INFORMATICS + TELECOMMUNICATIONS

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837

2023-2024

Selected Paper's Full Citation (IEEE)

B. Wang et al., "Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq," Scientific Reports, vol. 11, no. 1, Jan. 2021, doi: 10.1038/s41598-020-80881-2.

Link to the study: `https://www.nature.com/articles/s41598-020-80881-2`

Link to the dataset: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158631`

Selected Paper's Problem Statement and Importance

The paper by Wang et al. (2021) investigates the mechanisms underlying lymph node metastasis in gastric cancer (GC) using single-cell RNA sequencing (scRNA-seq). The authors focus on identifying potential marker genes and evolutionary drivers of metastasis while also exploring the intratumoral heterogeneity of GC. By performing scRNA-seq on samples from primary tumors and paired metastatic lymph nodes of three GC patients, they reveal distinct carcinoma profiles and diverse microenvironmental subsets, providing a comprehensive understanding of lymph node metastasis (1).

The importance of this study lies in the clinical challenges posed by gastric cancer. GC is a leading cause of cancer-related mortality globally, with poor prognosis, especially when metastasis occurs. Traditional bulk-level approaches have masked the roles of subpopulations, limiting understanding of lymph node metastasis. This study identifies potential GC lymph node metastasis marker genes (ERBB2, CLDN11, and CDK12) and cancer evolution-driving genes (FOS and JUN), providing insights into therapeutic targets.

The state-of-the-art research in this area includes studies utilizing scRNA-seq to uncover tumor heterogeneity, immune profiling, and drug response. For instance, Zhao et al. (2021) demonstrated how multiplexed drug perturbation combined with scRNA-seq in acute slice cultures from glioblastoma resections can identify cell type-specific drug responses, thereby facilitating personalized medicine approaches in oncology (2). Additionally, the integration of scRNA-seq with machine learning, as discussed by Qi and Zou (2023), has improved our capacity to predict cellular responses to immune therapies and drug treatments, enhancing targeted treatment strategies (3).

Selected Paper's Data Analysis Methods and Tools

Data Collection and Preparation: The research team collected tumor samples from primary and metastatic sites of three gastric cancer patients. These samples were processed through enzymatic digestion, filtration, and centrifugation to isolate single cells. The RNA from these cells was then reverse transcribed and amplified using modifications of the Smart-seq2 protocol. The resultant libraries were sequenced on a HiSeq2500 platform using a 50 bp single-end mode.

Data Preprocessing: Following sequencing, raw data underwent several preprocessing steps to ensure quality and usability. Using the Trimmomatic tool, adapters were trimmed, and reads with a Phred score below 20 were filtered out to enhance data quality. The remaining reads were aligned to the human genome (hg19 reference) using HiSat2. Gene expression levels were quantified using FeatureCounts, and the raw counts were normalized to transcripts per million (TPM) and log-transformed to stabilize variance and improve the interpretability of data.

Dataset Description: The dataset contains single-cell RNA sequencing (scRNA-seq) data from primary gastric cancer and paired metastatic lymph node tissues of three patients. It includes 21,196 rows representing genes and 95 columns representing single-cell samples. Specifically, there are 65 primary tumor (TT) samples and 30 metastatic lymph node (LN) samples. The first column contains gene names, while the remaining columns (e.g., GC1-TT1, GC1-LN1, GC2-TT1, GC3-LN12) represent expression levels of each gene in different single-cell samples from the three patients. This comprehensive dataset provides insight into the molecular mechanisms underlying gastric cancer metastasis.

Machine Learning and Statistical Analysis: The cleaned and normalized data were then analyzed using various machine learning and statistical techniques:

- **Principal Component Analysis (PCA)** was employed to reduce dimensionality and identify patterns and structures within the gene expression data.
- **Hierarchical Clustering Analysis (HCA)** and **t-Distributed Stochastic Neighbor Embedding (t-SNE)** were used to visualize the data, aiding in the identification of distinct cell populations and expression patterns.
- **Differential Gene Expression Analysis (DEGs)** was performed to identify genes that were significantly differentially expressed between primary tumors and metastatic sites, using fold change and false discovery rate (FDR) criteria.
- **Functional annotations** were derived through Gene Ontology (GO) analysis to understand the biological implications of differentially expressed genes.

Single-Cell Trajectory Analysis: Advanced trajectory analysis techniques, including TSCAN, diffusion map, and monocle2, were used to infer developmental trajectories and lineage relationships among single cells, providing insights into the cellular dynamics and progression from primary to metastatic tumor states.

Validation: The findings from the bioinformatics analyses were validated through immunofluorescence, targeting key proteins such as ERBB2 and CLDN11 to confirm their expression patterns observed in the sequencing data.

Selected Paper's Main Results

The authors of the paper aimed to understand the mechanisms behind lymph node metastasis in gastric cancer using single-cell RNA sequencing (scRNA-seq), on primary gastric cancer tumors and paired metastatic lymph nodes from three patients. Their results revealed distinct carcinoma profiles for each patient, with diverse microenvironmental subsets shared across different patients. They discovered significant intratumoral heterogeneity and identified a subgroup of cells bridging the primary and metastatic groups, indicating the transitional state of cancer during metastasis.

Key findings include the identification of specific gastric cancer lymph node metastasis marker genes, such as ERBB2, CLDN11, and CDK12, and evolutionary driver genes FOS and JUN. Additionally, trajectory analysis demonstrated a distinct pattern of evolution from primary to metastatic states, highlighting the regulatory roles of transcription factors like FOS, JUN, and ZNF256 in cancer progression. These discoveries provide new perspectives on potential targets for precision therapy in gastric cancer.

Selected Paper's Limitations

Several limitations could affect the generalizability and depth of the author's findings. The sample size of the study was relatively small, limited to only three male patients aged between 33 and 67. This narrow demographic scope restricts the study's applicability across different genders and age groups, potentially limiting the ability to generalize findings across the broader gastric cancer population.

Furthermore, the study did not utilize spatial transcriptomics, which could have provided valuable context regarding the spatial distribution of identified subpopulations within the tumor microenvironment, thus enriching the understanding of metastatic progression. Moreover, while the study did employ immunofluorescence to validate the presence of specific marker genes and potential drivers, it lacked functional validation through techniques such as gene knockdown or overexpression studies, which are crucial for confirming the roles of these genes in cancer progression. Additionally, the study's reliance on the Smart-seq2 protocol introduces technical limitations due to dropout events that miss lowly expressed genes, and could benefit from integrating additional sequencing technologies or multi-omics data such as single-cell ATAC-seq and proteomics to provide a more comprehensive understanding of the regulatory mechanisms driving gastric cancer metastasis.

Lastly, the authors utilized PCA and t-SNE for dimensionality reduction and hierarchical clustering to interpret the complex single-cell RNA sequencing data. However, PCA and t-SNE may not adequately capture the complex and non-linear relationships inherent in high-dimensional biological data and hierarchical clustering often depends on the choice of distance metrics and linkage criteria, which can affect the outcome of the analysis. To mitigate these limitations, it would be beneficial to explore different dimensionality reduction and clustering techniques that offer different strengths, such as those that maintain a balance between local and global data structure preservation. Broadening the analytical approach can provide a more robust and comprehensive understanding of the data, ensuring that findings are not artifacts of a specific method but are reflective of underlying biological processes.

Motivation for Additional Analysis - Working Hypothesis

Our team has developed a strategy to enhance the analysis of the single-cell RNA sequencing data by exploring a range of approaches. Initially, we will attempt to replicate the methods used by the original authors to establish a baseline. Following this, we plan to:

- Implement alternative quality control and normalization techniques to examine their impact on the data.

- Explore different clustering methods to better understand the data structure.
- Utilize advanced visualization tools like UMAP to improve our interpretation of the results.

Ultimately, our goal is to demonstrate that our modifications can complement the original findings and provide additional insights into the data.

In this project, we have chosen not to delve into pseudotime analysis as it falls outside the primary focus of this course. Instead, we aim to concentrate on methods that are more directly aligned with machine learning principles, which will allow us to explore and apply clustering, normalization and visualization techniques more relevant to our coursework.

Analysis Outline & Technical Approach

Replication of Original Study Methods: To ensure the reliability of our further analysis, our initial step will be to replicate the analysis pipeline from the original study. We intend to develop our pipeline using R, but we are also prepared to utilize a Python-based environment if specific tasks demand it or provide a more convenient alternative, depending on tool availability and workflow efficiency.

Data Normalization and Quality Control: As we embark on the analysis of our single-cell RNA sequencing data, it is crucial to implement rigorous quality control (QC) measures to ensure the integrity of the data. We plan to utilize the Seurat package, a widely recognized tool in the scRNA-seq analysis, known for its comprehensive suite of QC and data analysis functionalities (4).

Our QC strategy aims to incorporate the following:

- **Doublet Detection:** We plan to employ methods such as Scrublet or DoubletFinder, integrated within Seurat, to identify and exclude doublets from our datasets. Doublets can significantly skew the results and are essential to address, especially in single-cell data.
- **Mitochondrial Content Analysis:** Cells with unusually high mitochondrial gene expression often indicate cell damage or stress; hence, we will filter these cells out to maintain the quality of our analysis.
- **Cell Cycle Effect Adjustment:** Given that cell cycle variations can obscure genuine biological differences between cells, we will apply cell cycle scoring methods available in Seurat to minimize these effects in our downstream analysis.

The scater package (5) may also be a useful tool for quality control implementation. The authors' proposed package (1) or other available packages (6) may be used.

Normalization is a pivotal step in scRNA-seq data processing to allow for accurate comparisons across cells and conditions. We will explore the use of Seurat's normalization methods to address this.

Dimensionality Reduction and Clustering: Our approach will include the use of UMAP for dimensionality reduction, available through Seurat in R (4). UMAP is chosen for its ability to maintain more of the dataset's global structure and its computational efficiency even with larger datasets. It should also be noted that it is considered significantly faster than t-SNE (used by the authors).

For clustering, we plan to apply Gaussian Mixture Models (GMM) using the Mclust package in R, which is effective in optimizing mixture models and automatically determining the optimal number of clusters. GMMs are considered suitable for this analysis due to their ability to model the data points flexibly in a high-dimensional space, accommodating the complex distributions often observed in scRNA-seq data.

Alternatively, should UMAP or the GMM algorithm fail to clearly differentiate cell populations, we could employ other clustering approaches such as spectral clustering or DBSCAN, which might provide clearer separation.

Validation and Extension: Our goal is to validate our methods by comparing our results with those from the original study, assessing how different preprocessing and clustering strategies impact the identification of cellular subtypes and states. In the event of validation difficulties, we will focus on simplifying our analysis, such as reducing the number of parameters in our models or using basic descriptive statistics to draw comparisons between groups, ensuring that our findings are robust and easily interpretable.

Implementation Plan

Our initial efforts will be directed towards quality control and normalization. Our strategy involves one of us focusing on quality control and the other handling normalization. During the clustering phase, to efficiently replicate the algorithms used by the authors and explore additional methods we propose, we will divide these tasks and work concurrently. Ultimately, we will combine our work and jointly proceed with the implementation of the validation analysis.

The proposed methods and work plan are preliminary and subject to change. We anticipate potential challenges in algorithm implementation and in the quality control and filtering of raw expression data, which may require adjustments to our approach. Finally, it should be noted that the code for this paper is not provided.

References

[1] B. Wang, H. Lyu, J. Pei, Z. Huang, D. Wang, and W. Sun, “Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell rna-seq,” *Scientific Reports*, vol. 11, no. 1, Jan. 2021.

[2] W. Zhao *et al.*, “Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell rna-seq,” *Genome Medicine*, vol. 13, no. 1, May 2021.

[3] R. Qi and Q. Zou, “Editorial: Machine learning methods in single-cell immune and drug response prediction,” *Frontiers in Genetics*, vol. 14, Jun 2023.

[4] “Seurat: Tools for single cell genomics,” 2024. [Online]. Available: https://satijalab.org/seurat/

[5] “scater: Single-cell Analysis Toolkit for Gene Expression Data in R,” 2023. [Online]. Available: https://bioconductor.org/packages/release/bioc/html/scater.html

[6] D. J. McCarthy, K. R. Campbell, A. T. L. Lun, and Q. F. Wills, “Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R,” *Bioinformatics*, vol. 33, no. 8, pp. 1179–1186, 01 2017.