PROJECT PROPOSAL

Machine Learning in Computational Biology

Authors

Psallidas Kyriakos 7115152200033 & Alvanakis Spyros 7115152200020

Selected Research

Chung, W., Eum, H. H., Lee, H. O., Lee, K. M., Lee, H. B., Kim, K. T., ... & Park, W. Y. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nature Communications, 8(1), 1-12.

Text Link **Data-set Link**

Contents

1 Selected Research's Focus

The selected research paper focuses into the field of single-cell transcriptome analysis in breast cancer. The main objective of this research is to examine the transcriptome of individuals with breast cancer at a single-cell level and uncover distinctive gene expression patterns that cannot be detected through bulk tumor cell population transcriptome analysis. Additionally, this study aims to explore the range and impact of intratumoral heterogeneity. To achieve these objectives, the researchers seperated cells from the tumour microenvironment into different categories and cell types, namely cancer, immune, and stromal cells. They analyzed pathway activation scores and heterogeneity in cancer cells to identify potential links between breast cancer aggressiveness and molecular subtype-related pathways. Finally, the study also investigated immune cells to explore immune cell-specific gene sets and pathways and their relation to breast cancer sub-type characterization.

2 Importance

Breast cancer is a heterogeneous disease, meaning that tumors can differ in their genetic and molecular makeup. While genomic profiling is commonly employed to describe a bulk tumor in patients, cancer cells show variability within the tumor itself which could impact the effectiveness of a personalized treatment approach. Single-cell RNA sequencing (scRNA-seq) is a powerful technique that allows researchers to examine the gene expression profiles of individual cancer cells within a tumour sample as well as non-cancer cells (e.g. immune) in the tumour micro-environment, allowing for deeper characterization and subsequently precise therapeutic approach.

3 Methods & Results

3.1 Methods

The research carried out an analysis of the transcriptome of 515 cells obtained from 11 patients with breast cancer, encompassing the four molecular subtypes of breast cancer: Luminal A, Luminal B, Triple-negative-breast-cancer (TNBC), and HER2. To ensure data quality, RNA-Seq data underwent **Quality Control Preprocessing** using the **RNA-seQC** and **RSEM algorithms**. Copy number variations (CNVs) were inferred from RNA-seq data by averaging the CNVs across single cells and comparing correlations between genomic CNVs and inferred CNVs using **Pearson's correlation from R functions**, with the expression profiles of normal breast tissues used as a reference. To compare single-cell expression with the pooled sample expression, **Pearson's correlation** and **multiple regression**

analysis were utilized. The study also examined visually the extensive intratumoral heterogeneity through Principal Component Analysis **PCA**.

To eliminate the possibility of non-carcinoma cells interfering with the heterogeneity results, the study aligned single-cell gene expression profiles along the chromosomes as **moving averages** and used **hierarchical clustering** to separate them. The study then validated graphically the separation of the two cell type categories by **PCA**. Additionally, the **ESTIMATE algorithm** was utilized to assign gene expression scores associated with immune or stromal signatures, thereby enabling the separation of non-carcinoma cells into immune or stromal [5]. Having separated the cell categories **PCA** was utilized to graphically showcase the distinct carcinoma characteristics of each patient's tumour and its microenvironment's immune and stromal cells.

Regarding cancer cells, the **R package genefu** was used to assign ER and HER2 scores to each cell, and aggressive cancer gene expression signatures were analyzed via gene enrichment scores for the EMT, stemness and angiogenesis pathways with the **R package GSVA**. The study then utilized the **likelihood ratio test (LRT) of the R package Seurat** to carry out differential gene expression for the predicted subtype of each cancer cell. For TNBC cells, further subtyping into six categories was carried out via the **TBNCtype software** [1, 2].

Finally, regarding immune cells, the study classified them into three groups through **non-negative factorization clustering** with immune cell type-specific gene sets. Additionally, T-cells were **hierarchically clustered** based on their **GSVA enrichment** scores for gene sets for naive T cells, T-cell costimulation, regulatory cytokines and receptors, T-cell exhaustion, and cytotoxicity.

3.2 Results

The analysis of copy number variations (CNVs) revealed significant alterations in triple-negative breast cancer patients, confirming previous reports of extensive genomic instability in this subtype of tumor.

In terms of single-cell expression, the Pearson's correlation analysis between the pooled and individual cell samples showed partial but significant correlations, with a better representation of the tumor population achieved through multiple regression analysis of the transcriptomes of different-sized pools of single cells. Principal component analysis (PCA) indicated a mixed distribution of intra- and interpatient cells for each molecular subtype of breast cancer.

Furthermore, using both PCA and the ESTIMATE algorithm for carcinoma and non-carcinoma cell classification, the analysis found that of 515 single cells, 317 were epithelial breast cancer cells, 175 were tumour-associated immune cells, and 23 were non-carcinoma stromal cells. With the use of R genefu, breast cancer cells

were further classified into molecular subgroups based on ER and HER2 module scores. The R package GSVA illustrated the variable expression of aggressiveness in carcinoma cells.

Triple-negative breast cancer (TNBC) cells demonstrated higher EMT signatures, and both HER2 and TNBC tumor cells expressed high levels of stemness and recurrence signatures. Additionally, the likelihood ratio test (LRT) of the R package Seurat showed that TNBC carcinoma cells expressed variable upregulation of genes in basal pathways, and TNBC tumours exhibited extreme heterogeneity.

Overall, the findings demonstrate significant intra- and intertumoral heterogeneity among the different subtypes of cancer cells. Single-cell transcriptome profiling is a valuable tool for understanding breast cancer and enabling more targeted and effective treatments for patients.

4 Proposed analysis

Given the comprehensive analysis performed by the researchers, as showcased in Section 3, we suggest that our analysis begins with a replication and validation process, following the same methods as described in the original paper. As we move forward with our analysis, we have identified areas where improvements can be made beyond the initial replication as detailed below.

4.1 Extension: Clustering analysis

4.1.1 Reasoning

The paper used hierarchical clustering algorithms for crucial tasks: to distinguish between cancer and non-cancer cells, allocate immune cells in groups, and categorize T cells based on pathway activation. However, the paper only evaluated the clustering method used for immune cell allocation (supplementary material) and did not mention the linkage type used for the other tasks. There were also no internal evaluation criteria provided to assess the effectiveness of the outcomes. We propose a further analysis, beyond replication, of the clustering approach used in the paper to assess its effectiveness in all three tasks, believing that it will result in a more comprehensive explanation.

4.2 Extension: Dimensionality reduction analysis

4.2.1 Reasoning

The paper utilizes the PCA method for dimensionality reduction and data visualization. While PCA is widely used for biological data, it assumes linearity in the data, which may not hold true in cases where the number of dimensions is large

and the linear relationship between genes is not certain. This is evident from the very low variance ($< 10\%$) explained from the first two PCs in the PCA graphs of the paper. Therefore, we also intend to explore other methods for dimensionality reduction analysis and 2D visualization of the data in all three tasks, believing that it will result in a more comprehensive explanation.

5 Analysis outline & Technical implementation

5.1 Replication & Validation

The analysis packages used in this study were implemented in R, but no code is available. To replicate and validate the results, we will use Python, specifically Scanpy [4], which is a leading single-cell analysis library. With this library, we can handle single-cell datasets and analyze pathway activation as accomplished in the paper. However, while some breast cancer-specific algorithms such as **TNBCtype** are available online, others such as **ESTIMATE** are only available in R. To address this, we will consider implementing part of the analysis in R, or bridging R and Python using the r2py library as needed.

5.2 Analysis extension

We suggest two additional algorithms for clustering evaluation: the Leiden clustering algorithm, which is the state-of-the-art for single-cell data clustering [3], and spectral clustering, which is effective for high-dimensional data. Leiden can be accessed via the Scanpy library, while spectral clustering is available via the sci-kit learn library. The optuna library will be used for hyper-parameter optimization. Regarding the evaluation of the algorithms, internal evaluation metrics will be utilized. Specifically the silhouette score, Calinski-Harabasz index and Davies-Bouldin index, available via scikit-learn.

To perform dimensionality reduction analysis, we will utilize the Variational Auto Encoder (VAE) and t-Distributed Stochastic Neighbor Embedding (tSNE) methods, both of which are commonly used for non-linear data. tSNE is a state-of-the-art technique for visualizing high-dimensional data that preserves the local structure of the data. In other words, nearby points in high-dimensional space are likely to be represented by nearby points in the low-dimensional map. On the other hand, VAE is a powerful and flexible technique for representation learning and dimensionality reduction, particularly when dealing with complex non-linear relationships in the data, such as gene expression. These techniques can be implemented using libraries such as scikit-learn and TensorFlow.

6 Implementation plan

We have developed an implementation plan that outlines our approach to this project as illustrated in (figure 1) below. Note that while the white boxes represent the replication of the analysis, the extended analysis (blue boxes) plan is subject to change. If the replication of the main analysis proceeds without major issues, we intend to expand our clustering analysis to include additional methods, such as density-based techniques (e.g., DBSCAN) and probabilistic approaches such as Gaussian Mixture Models (GMM) and/or more hierarchical approaches with different linkage types. Finally, we will leverage the diverse strengths of each team member, particularly in mathematical and biological expertise.

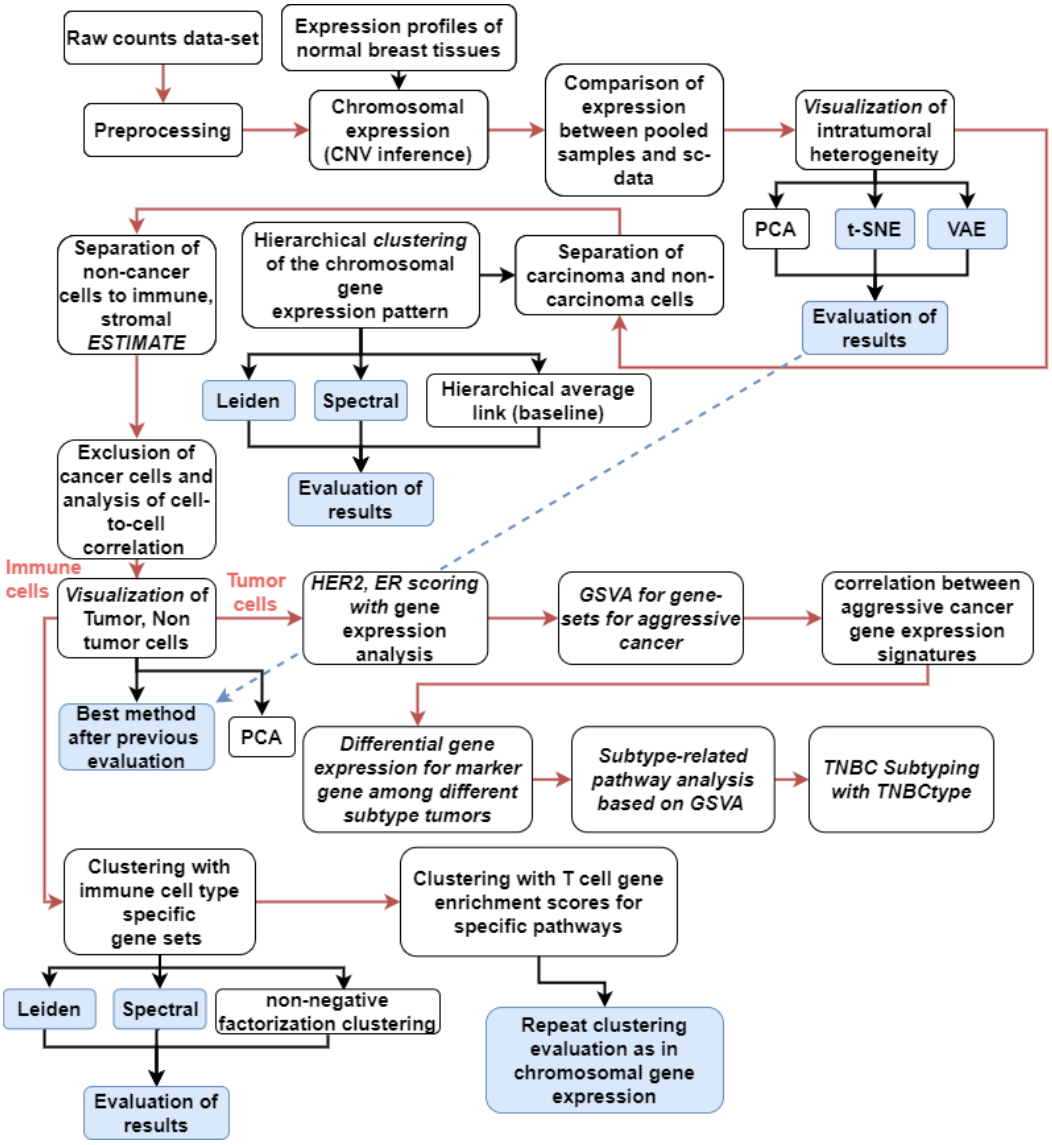

Figure 1: Implementation flowchart — red lines showcase the main path, blue boxes indicate the planned extended analysis

References

[1] Xi Chen, Jiang Li, William H. Gray, Brian D. Lehmann, Joshua A. Bauer, Yu Shyr, and Jennifer A. Pietenpol. TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer. *Cancer Informatics*, 11:CIN.S9983, January 2012.

[2] Brian D. Lehmann, Joshua A. Bauer, Xi Chen, Melinda E. Sanders, A. Bapsi Chakravarthy, Yu Shyr, and Jennifer A. Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation*, 121(7):2750–2767, July 2011.

[3] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

[4] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[5] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W. Laird, Douglas A. Levine, Scott L. Carter, Gad Getz, Katherine Stemke-Hale, Gordon B. Mills, and Roel G.W. Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, 4(1):2612, October 2013.