Studying inter- and intratumor heterogeneity of triple negative breast cancer

by

Anastasia Leshchyk

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science in

Bioinformatics and Computational Biology

April 2019

APPROVED:

Professor Dmitry Korkin, Thesis Advisor

Elizabeth F. Ryder

Professor Elizabeth F. Ryder, Thesis Reader

Professor Amity Manning, Thesis Reader

Professor Dmitry Korkin, Head of Program

Abstract

Triple-negative breast cancer (TNBC) is known for its high inter- and intra-tumor heterogeneity. TNBC is associated with poor survival prognosis due to its aggressiveness and lack of effective therapies. In this project, we are examining single-cell RNA-seq data of primary untreated TNBC tumors obtained from six patients. We apply deep unsupervised single-cell clustering method (DUSC) to reveal new subpopulations of cells sharing common transcriptomic features and proving new biological insights to the TNBC heterogeneity. Our analysis reveals subgroups of cells shared among all patients, determine gene signatures associated with the poor and good patient survival rates, and could be used in identifying new biomarkers. We propose DUSC as a new method to identifying copy-number variation tumor subclones, with the copy numbers strongly correlating with the patient survival prognosis.

Acknowledgment

I want to thank my academic and research advisor Dr. Dmitry Korkin; Korkin lab members, Dr. Liz Ryder, Dr. Randy Paffenroth, Dr. Tanja Dominko, and all my family and friends for their help and support throughout my time at WPI.

Content

List of figures

1. Background

1.1 Introduction

The research we performed was in cancer biology area, addressing the important problem of studying cancer tumor heterogeneity. To find out how cancer occurs and evolves, researchers are investigating the biological differences between normal and cancer cells. The last few decades cancer biology research has revealed new mechanisms of cancer development and progression. However, the understanding of the cancer evolution and its resistance to the therapy remains a challenging task. In our research, we are attempting to identify and to annotate different transcriptomic profiles in cancer cells among triple-negative breast cancer (TNBC) tumors. The TNBC is known for its high degree of heterogeneity within and among the tumors and accounts for approximately 15%-25% of all breast cancer cases [1]. It is associated with poor prognosis and remains a type of cancer hard to treat. Due to its biological features, the existed drug target therapies available or other breast cancer types are not effective for TNBC diagnosed patients.

Studying TNBC cancer heterogeneity within and among different tumors might reveal new cancer cells molecular profiles, pointing to possible strategies for the new effective treatment. Our results might contribute to the new drug target identification and treatment design.

1.2 Cancer

Cancer is a group of diseases, affecting different parts of the human body. It is characterized by the rapid proliferation and the abnormal cell growth beyond their usual boundaries, with possible invasion into neighboring tissues in the body and spreading to other parts of it. Metastasis, which is spreading of cancer cells to other organs, remains the major cause of death from this disease around the world.

According to the data provided by the World Health Organization (WHO), 9.6 million people diagnosed with cancer died in 2018, which puts this disease as second in the list of death-causing conditions globally. Cancer keeps effecting economy due to its increasing cost of treatment. According to the data from 2010, the total annual economic cost of cancer was estimated at approximately US$ 1.16 trillion [9].

The most frequently occurring types of cancers are: lung (2.09 million cases); breast (2.09 million cases); colorectal (1.80 million cases); prostate (1.28 million cases); skin cancer (non-melanoma) (1.04 million cases); stomach (1.03 million cases).

Cancer is caused by the combination of genetic and external factors. The WHO defines three main categories of the cancer-causing external factors: physical carcinogens, (ultraviolet and ionizing radiation); chemical carcinogens, (asbestos, tobacco smoke, aflatoxin), and arsenic (a drinking water contaminant); and biological carcinogens, (infections from certain viruses, bacteria, or parasites).

An adequate cancer diagnostics and prognosis is an important step in effective cancer treatment. Every cancer type follows a unique treatment strategy, frequently, employing a combination of several approaches. The most common treatment strategies include surgery, radiotherapy, and chemotherapy. A recently introduced and promising approach is a drug target therapy, which works by targeting cancer associated with specific genes, proteins, or the tissue environment that contributes to cancer growth and survival. The identification of the effective targets for such therapy remains a challenging problem as cancer cells could develop drug resistance and stop responding to the treatment [10-13].

1.3 Triple-negative breast cancer

Triple-negative breast cancer (TNBC) is a cancer, characterized by loss of estrogen receptor (ER), progesterone receptor (PgR), and human epidermal growth factor receptor 2 (HER2) gene expression. TNBC is a disease associated with poor survival, due to its aggressiveness and lack of effective targeted therapies [14]. Up to 15% of all breast cancer falls into this category which is associated with a high rate of local and systematic recurrence [15]. There are four unique ontologies and differential response to standard-of-care chemotherapy TNBC molecular subtypes: basal like-1 (BL1), basal like-2 (BL2), mesenchymal (MSL) and luminal AR (LAR) [16]. These subtypes differ in age, grade, local and distant disease progression, and histopathology. BL1 subtype has an elevated cell cycle and DNA damage response gene expression. BL2 is enriched in growth factor signaling and myoepithelial markers. MSL is known for upregulated expression in epithelial-mesenchymal-transition, growth-factor pathways and decreased expression of genes involved in proliferation. LAR is defined by luminal gene expression and is triggered by its androgen receptor (AR) [16].

The genetic profile of TNBC is not well studied. The most common mutation in TNBC patients is in TP53 in 62% of basal TNBC and 43% of non-basal TNBC. TNBC-associated frequent mutations include PIK3CA (10.2%), USH2A (9.2%), MYO3A (9.2%), PTEN (7.7%), and RB1 (7.7%). Comparing genomic data with corresponding whole transcriptome data revealed that only 36% of mutations are expressed. Although TP53 and PIK3CA/PTEN somatic mutations appear to be clonally dominant when compared with other pathways, in some cancers their clonal frequencies are incompatible with founder status [17].

Poor TNBC prognosis is also reported to be associated with a tumor microenvironment, usually characterized by higher vascular endothelial growth factor (VEGF), tumor infiltration lymphocytes (TILs) and Tumor-Associated Macrophages (TAM) [18]. The latter play an important role in the immune response to cancer cells, but the mechanism of how the tumor microenvironment controls TAMs and T-cell response is not well studied. TAMs fill out a major leukocyte population infiltrating tumors that originate from circulating blood monocytes, differentiated into macrophages after their relocation into tissues. Later, they usually undergo M1 (classical) or M2 (alternative) activation. In TNBC, TAMs are reported to promote tumor growth and progression by several mechanisms, a few of them include the secretion of inhibitory cytokines, the reduction of effector functions of Tumor Infiltrating Lymphocytes (TILs) and the promotion of Regulatory T-cell (T-reg) [14]. TAMs are reported to, directly and indirectly, modulate anti-programmed cell death 1 PD-1 and anti-PD-ligand(L)1 agents' expression in the tumor environment [14]. Speculating with TAM unique properties, several TAM-associated TNBC-treatment strategies have been proposed. One of them is based on the prevention of TAM differentiation, and their quantity reduction, switching M2-type TAMs into antitumor M1-phenotype and decelerating TAM-associated molecules [14]. However, the high cost of these drugs and the lack of validated predictive biomarkers support the development of strategies aimed to overcome resistance and optimize the efficiency of these approaches. Also, developing an identification strategy of such macrophage's stages can give an insight into optimal treatment solution.

The TNBC intratumor diversity causes metastasis, treatment resistance and, thus, poor survival rate. Cancer evolution within a primary tumor is suggested to be a reason for metastasis development and prevention of the complete elimination of cancer cells. Copy-number variations among cancer subclones are known to reflect the punctuated evolution within a tumor [19]. Merging state-of-the-art computational tools and single cell analysis can greatly contribute to such tumor subclones identification.

1.4 Single-cell RNA-seq technology

High-throughput sequencing methods invented during the last decade, have significantly impacted modern biology. Whole transcriptome data analysis has revealed many novel biological insights and became an essential part of medical research. Unlike RNA-seq (bulk), which averages gene expression patterns across thousands to millions of cells, single-cell RNA-seq methods consider biological differences between cells by capturing the transcripts of isolated single cells and generating sequencing libraries in which the transcripts are mapped to individual cells. Single-cell RNA-seq exposes fundamental biological properties of cell populations and biological systems at an unprecedented resolution [20].

Current single-cell RNA-seq protocols involve the following steps (Fig. 1): 1) isolation of single cell and RNA, 2) reverse transcription (RT), 3) RNA amplification, 4) library generation and 5) sequencing. Novel protocols encapsulate individual cells in droplets in a microfluidic device, where RT is performed, obtaining cDNA from RNAs. Each droplet carries a unique identifier that maps the cDNAs derived from a single cell. When RT is over, the cDNAs from many cells can be mixed for sequencing; transcripts from a particular cell are identified by the unique identifier [21-22].

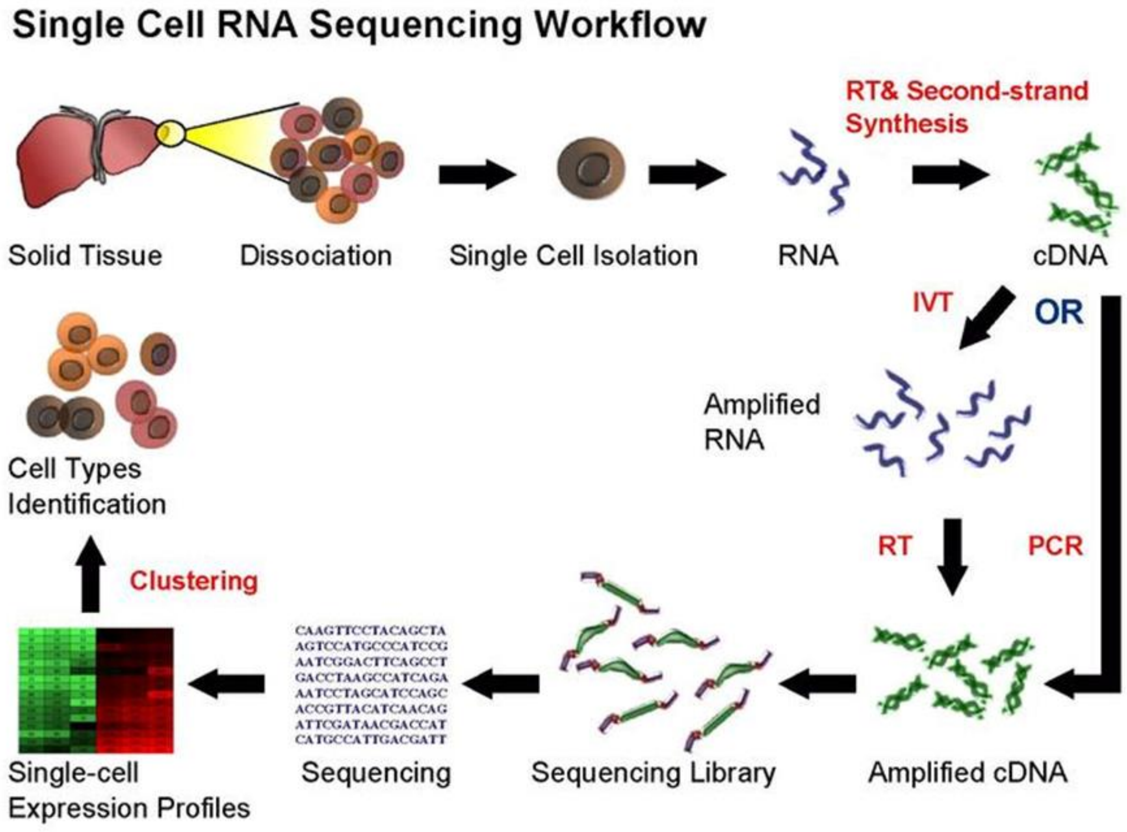


Figure 1. Single-cell RNA-Seq workflow (From [21])**.** The picture demonstrates the major steps in the workflow: 1) extracting the cells from solid tissue, isolation of single cell and their RNAs, 2) performing reverse transcription and second-strand synthesis, 3) amplification of RNAs, 4) library generation and 5) sequencing.

Single-cell RNA-seq meets a few challenges. One of the challenges is in preserving the initial relative abundance of mRNA in a cell. Another challenge is to identify rare transcripts in a cell [23]. The RT is a critical step of the workflow as its efficiency determines how much of the cell’s RNA population will be eventually analyzed. To amplify cDNA, either Polymerase Chain Reaction (PCR) is used or in vitro transcription (IVT). There are a few known single-cell RNA-seq protocols: Tang et al. [24], C1-CAGE [25], SMART-seq [26], STRT[27], Quartz-seq [28], RAGE-seq[29] and CEL-seq [30]. The protocols differ in RT strategy, cDNA synthesis and amplification, and the possibility to accommodate sequence-specific identifiers or to process pooled samples [31].

2. Methods

2.1 Dataset

The dataset contains single-cell RNA-seq data of six tumors of six patients diagnosed with triple-negative breast cancer [1]. The tumors underwent the dissociation followed by the flow-cytometry sorting of single variable cells (Fig. 2A). Isolated cells went through the cDNA preparation and library construction, then through Next Generation Sequence (NGS). After quality control and normalization procedures, the total number of cells was 1189. The cells were analyzed and labeled by clustering and gene-markers approaches. It resulted into identification of 1112 epithelial cells (Fig. 2C), and 244 non-epithelial cells. Also, the authors identified the cell cycle phases and distinguished high cycling cells from the non-cycling (Fig. 2D). Most of the cycling cells (98.5%) were epithelial cells. The authors suggested and determined that the malignant cells reside among epithelial cells.

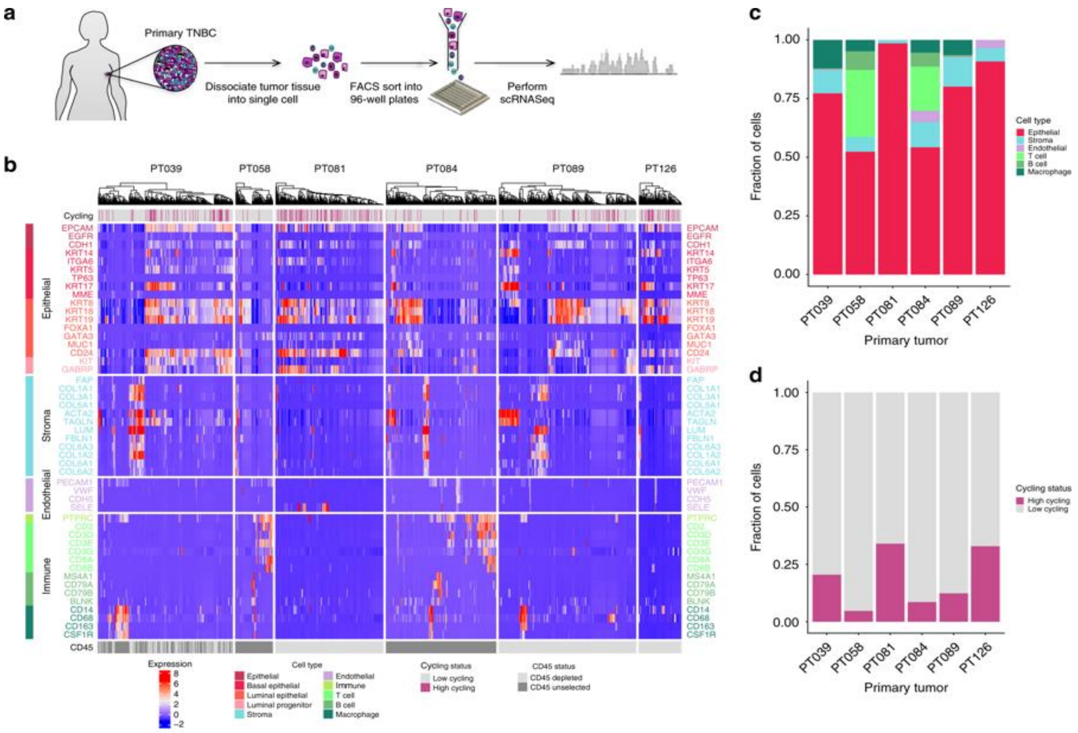


Figure 2. TNBC dataset description (Adopted, from [1]). **a-** The illustration of the steps preceding the collection of single-cell RNA-seq data for the six primary TNBC tumors cells. **b** – Color-coded heatmap displaying the diversity of the passed quality control tumor The columns represent the cells grouped according to the patient identifier, with cycling stage color-coding: high cycling cells are pink and low-cycling cells are gray. The rows demonstrate the expression of the known cell types gene markers. The bottom bar shows the distribution of the cells with the depleted CD45 +. **c-** Bar plots demonstrating the distribution of the assigned cell types among 1112 cells in six primary TNBC tumors. **d** - Bar plots demonstrating the distribution of the assigned low/high-cycling cells among 1112 cells in six primary TNBC tumors.

The dataset could be accessed online at the Gene Expression Omnibus database under the accession code GSE118390.

2.2 Deep unsupervised single-cell clustering analysis

To identify the informative representation of the single-cell transcriptomic data we applied deep unsupervised single-cell clustering method (DUSC) [2]. The method is based on a deep feature learning approach and further clustering of single-cell RNA-seq data. In DUCS, the deep feature learning is accomplished with denoising autoencoders, which leverage the number of latent features.

Autoencoder is a type of neural network, which tries to emulate its input [3]. Autoencoder projects the input data on a different dimensional space and retrieves the data from that dimension onto the original dimension. In the DUSC workflow, the model of the autoencoders is trained during the pre-processing stage. The input is represented as a matrix, where columns are feature vectors with gene expression value, and the rows are the corresponding cell identifiers. The input matrix is preprocessed before the training by removing the columns with all zero values. Then, the columns are normalized by the formula:

$$Norm(x_i) = \frac{x_i - x_{min}}{x_{max} - x_{min}};$$

where x_{max} and x_{min} are the maximum and the minimum feature values across the matrix, and x_i is a feature value in x.

Autoencoder has two parts: an encoder and a decoder. The encoder projects the input data from a higher dimension to a lower dimension according to the formula:

$$y \ = \ s(Wx + b).$$

In DUSC, s(x) is a sigmoid function $s(x) = \frac{1}{1+e^{-1}}$. The decoder part, where the hidden dimension is again projected to the higher dimension, same as the input is described by the formula:

$$z \ = \ s^{'}(W^{'}y + b^{'}) \ .$$

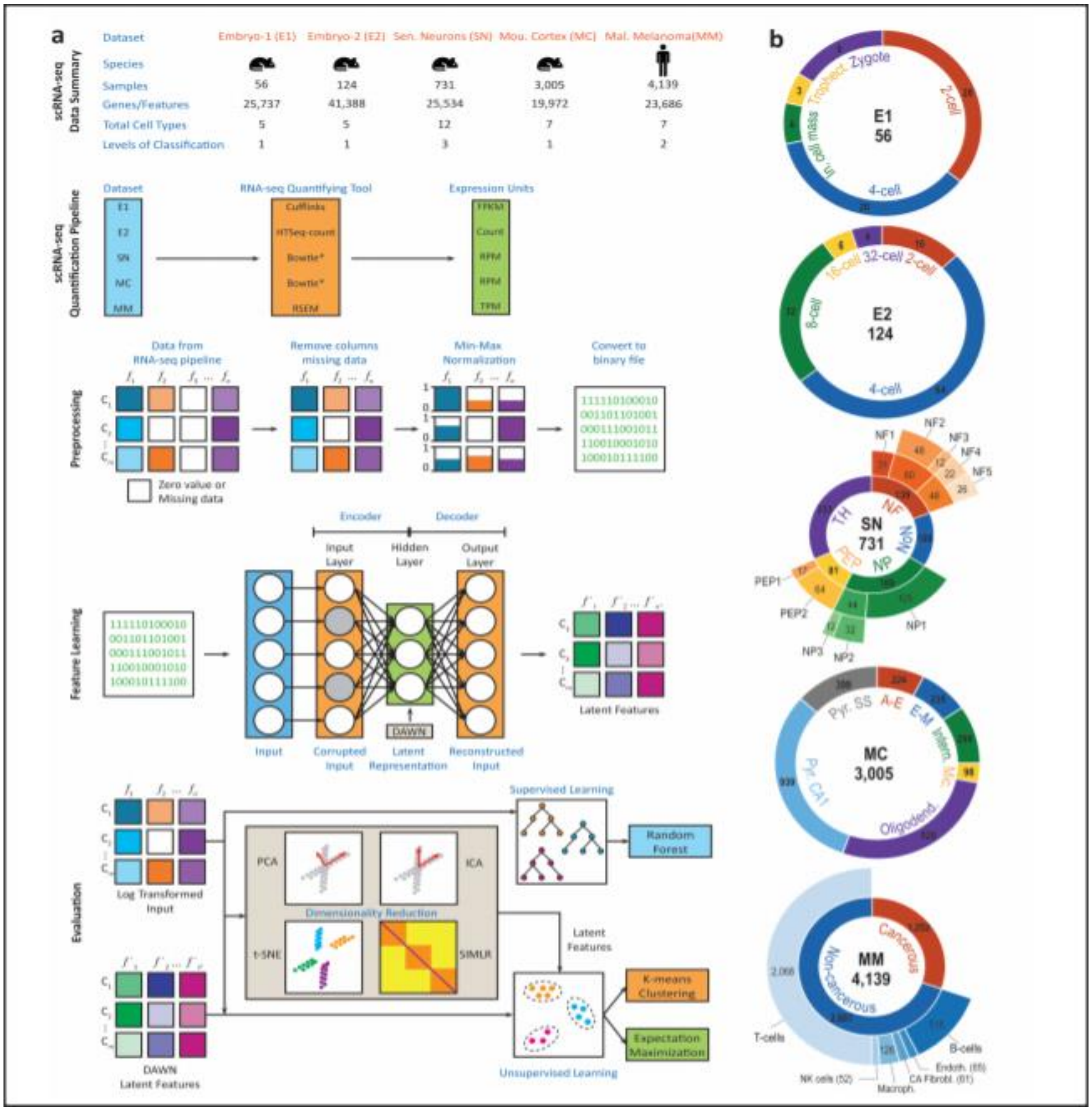
Ideally, the projection is such that it reconstructs the input itself. Generally, in the case of traditional autoencoders, there is a high chance that it learns the identity function to reconstruct the input. This identity although correct is not very useful in applications. Also, it cannot make the autoencoder robust to noise in the input. Denoising autoencoder does exactly that. It trains in noisy input and reconstructs the clean input. Usually, it is done by adding stochastic noise to the input and feeding it to the input of the autoencoder. In the DUSC workflow, the noise is introduced by randomly selecting n features of each input vector x_i and assigning them zero values. The autoencoder is trained to minimize an error defined as $L_H(x, z)$, of the latent features:

$$L_H \ (x, z) = \ - \sum_k^d [x_k \log z_k + (1 - \ x_k \ ) \log(1 - \ z_k \ )],$$

where d is a length of the vector of the feature vector. For the reconstruction loss, the output of the autoencoder is compared to the original, not corrupted input.

The optimization model selects the optimal number of hidden layers and hidden units for the denoising autoencoders in this method. DUSC is implemented using the Theano Python library [33], which supports NVidia CUDA. The authors suggest that this implementation allows fast training of the neural network layers with the increased number of neurons using NVidia GPUs.

The DUSC pipeline contains four major parts (Fig. 3). The first step includes the data quality check and pre-processing for denoising autoencoders training. The second step includes the feature learning using Denoising Autoencoder With Neuronal approximation (DAWN), which consists of training denoising autoencoders and further hyper-parameter optimization. The third step includes previously published unsupervised learning methods: Principle Component Analysis (PCA), Independent Component Analysis (ICA), t-SNE, SIMLR, which are used to generate the compressed dimensions of the input dataset. This step is made for the comparison purpose of autoencoders performance. The fourth step includes reducing the feature representation obtained from each of the previously described methods and passing them to the clustering methods: K-means and expectation-maximization, to reach the clustering accuracy.



Dataset	Embryo-1 (E1)	Embryo-2 (E2)	Sen. Neurons (SN)	Mou. Cortex (MC)	Mal. Melanoma(MM)
Species					
Samples	56	124	731	3,005	4,139
Genes/Features	25,737	41,388	25,534	19,972	23,686
Total Cell Types	5	5	12	7	7
Levels of Classification	1	1	3	1	2

Figure 3. Illustration of the DUSC workflow. (Adopted, from [2]). **A**: The demonstration of DUSC stages: brief description of the datasets used for the validation; RNA-seq quantification pipeline; data preprocessing stage; deep feature learning; evaluation and result comparison with the other unsupervised methods. **B**: A more thorough datasets description, which were used for the method validation: Embryonic Dataset-1 (E1), Embryonic Dataset-2 (E2), Sensory Neurons (SN), Mouse Cortex (MC), and Malignant Melanoma (MM).

2.3 Differential expression gene analysis

Seurat 2.0 performed the differentially expressed gene analysis. Seurat is an R package for the analysis of the single-cell RNA-seq data, which aims to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements [4, 5].

The Seurat workflow starts with the alignment of the input, which is a list of at least two scRNA-seq data sets. The alignment consists of two steps. The first step includes the application of canonical correlation analysis, learning gene correlation between the two datasets (Fig. 4A) and optional identification of the outliers, which cannot be described by the learned function. This step can help to identify outlying cell communities, which do not lie in the intersection of the two compared datasets and could be further analyzed. Then, the two datasets are aligned into a low-dimensional manifold using nonlinear algorithms and further analyzed by, for example, applying clustering methods. The second step of the workflow includes a comparative analysis of the intersection of the two datasets and the original datasets. It reveals changes in population density or gene expression (Fig. 4B). In the original paper, these steps of the workflow were successfully applied and verified on the five different datasets of single-cell RNA-seq experiments.

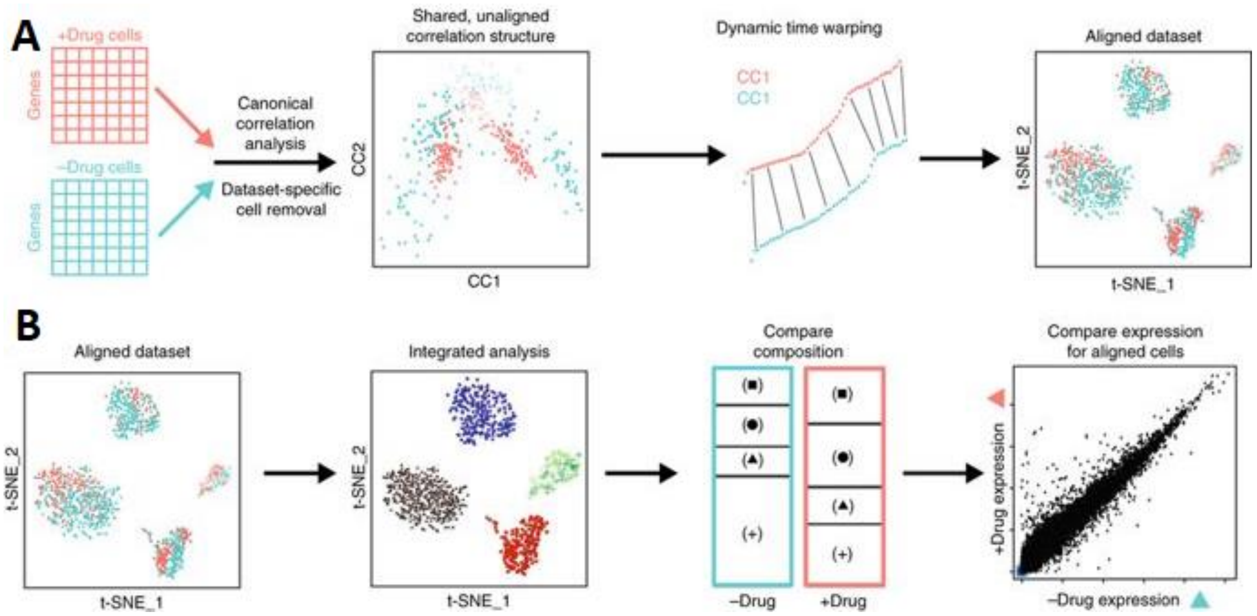


Figure 4. Overview of the Seurat alignment of single-cell RNA-seq datasets (Adopted, from [5]). **A** - The use of canonical correlation analysis (CCA) to reveal a common correlation between two datasets. Further, the cells are projected into a low-dimensional manifold (visualized here in 2D with t-SNE). **B** - After alignment, clustering methods can highlight unique cell types across the datasets, which proceeds to identify shifts in cell type proportion.

We applied Seurat methods on the cell clusters of interest. We obtained the list of differentially expressed genes with pre-calculated adjusted for multiple hypothesis p-value, log fold change, which sign shows the direction of the change and other parameters (Supplementary Table S1-10). Then, we used the adjusted p-value cut-off = 0.05 to select only statistically significant differentially expressed genes among clusters. For each cluster, we obtained lists of up- and down-regulated genes, based on the log fold change value.

2.4 Survival analysis

For the survival analysis, we used the Kaplan-Meier survival curve. The survival curve is a function of time and the event of interest of a subject (patient). Time could be bounded by the time of the subject enrollment in a study, the beginning or the end of the treatment, when the event of interest is reached, or the subject is censored (withdrawing) from the study. This duration of time is known as serial time, and it describes the clinical-course time.

In the Kaplan-Meier survival analysis, each subject is characterized by three parameters: 1) serial time, 2) status at the end of their serial time, and 3) study group [6]. When constructing the survival curve, the serial times of the subjects are arranged from the smallest to the biggest, disregarding when they entered the study. All the subjects within the group are mapped to the curve at its beginning and then marked when withdraw or meet the event of interest. Censoring happens when a subject drops out, is lost to follow-up, or the required data is not available anymore. Alternatively, the study might end before the subject meets the event of interest. Therefore, censoring can occur before the study ends. The serial time duration is bounded by the event of interest, known as an interval and is indicated as a horizontal line on a

Kaplan-Meier curve. All censored subjects are marked on the curve as tick or “star” marks and do not terminate the interval.

In this project, to analyze the prognostic value of a gene transcript, we divided transcripts into two cohort according to the median (or upper/lower quartile) of gene expression. The two groups can be compared in terms of relapse-free survival, overall survival, and distant metastasis-free survival. The curves were generated for each gene by the online tool available at the web address: www.kmplot.com [7]. The background database was established using gene expression data, and survival information of 1,809 patients downloaded from GEO (Affymetrix HGU133A and HGU133+2 microarrays). Using the online tool, for each we obtained the Kaplan-Meier curve, hazard ratio with 95% confidence intervals, log-rank p-value, and False Discovery Rate (FDR) (Fig. 5).

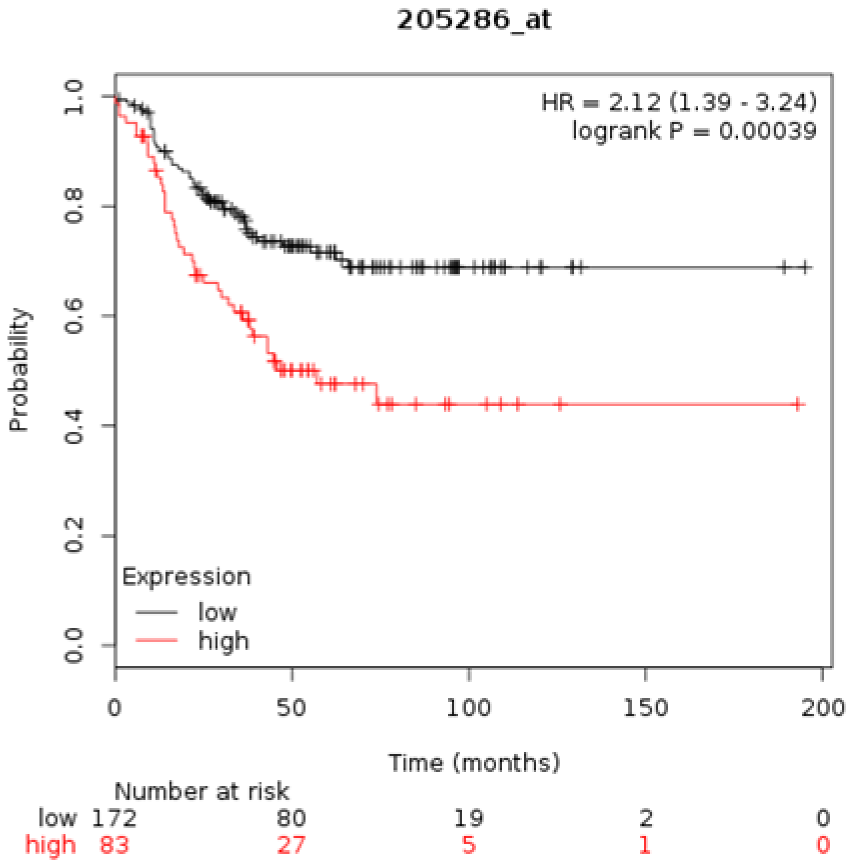


Figure 5. Kaplan-Meier plot for a gene of interest. The two curves are displayed corresponding to the lower- an upregulated gene expression of a gene of interest. The lengths of the horizontal lines along the X-axis of serial times represent the survival duration for that interval, terminated by the event of interest. The thick dashes on the curve are representing censoring data. In the bottom of the graph, there are numbers of patients corresponding to the time values on the x-axis. The color encodes the patients carrying up-(red) or downregulated (black) gene of interest.

We addressed the two groups of patients: triple-negative breast cancer diagnosed patients and any other breast cancer type, patients. We obtained the curves for these two groups of patients and further analyzed them.

2.5 Inferring copy-number variations

Copy-number variations (CNVs) are structural changes in the genome, where the fragment is repeated several times or deleted. There events are called duplication and deletion respectively. There are a variety of computational methods estimating CNVs from single-cell RNA-seq. In this project we use InferCNV [8].

InferCNV is an R package, which deduces somatic large-scale chromosomal copy number alterations, such as gains or deletions of an entire chromosome or its fragment. InferCNV is developed under the approach, which is based on exploring expression intensity of genes across positions of the genome in comparison to “normal” cells. In our analysis, since we do not have any labeled malignant cells, we do not use any “normal” cell reference. Instead, a refence cell cluster is picked by a default and compared to the remaining clusters. As the output of the tool, a heatmap is generated, which illustrates the relative expression intensities across the genome of cells. The visualization and the color-code makes it obvious which regions of the cell cluster genome is over-abundant or less-abundant with CNVs in comparison with the other clusters. In addition, InferCNV builds a dendrogram of cell clusters, defining the cell cluster hierarchy according to their pattern of heterogeneity. In the InferCNV package, there is a possibility to pick residual expression filters to explore minimizing noise and further highlighting the signal supporting CNVs.

3. Results

3.1 Cellular diversity in primary TNBC

The first step of our analysis was the identification of cell populations shared among all tumors. We hypothesized that DUSC identifies the groups of cells sharing molecular subtype or a cell type. On the dataset represented by six primary tumors (Fig. 6A), we applied DUSC and obtained 9 clusters (Fig. 6B). We observed clear separation (Fig. 6C) of clusters containing immune cells (cluster 5), from the rest of the cell clusters. The epithelial cells were distributed among clusters 1, 2, 3, 4, 6, 7 and 8, suggesting their patient-specific nature. The unsupervised UMAP visualization tool on the clusters (Fig. 5D) demonstrated the same clear separation between immune and non-immune cells.

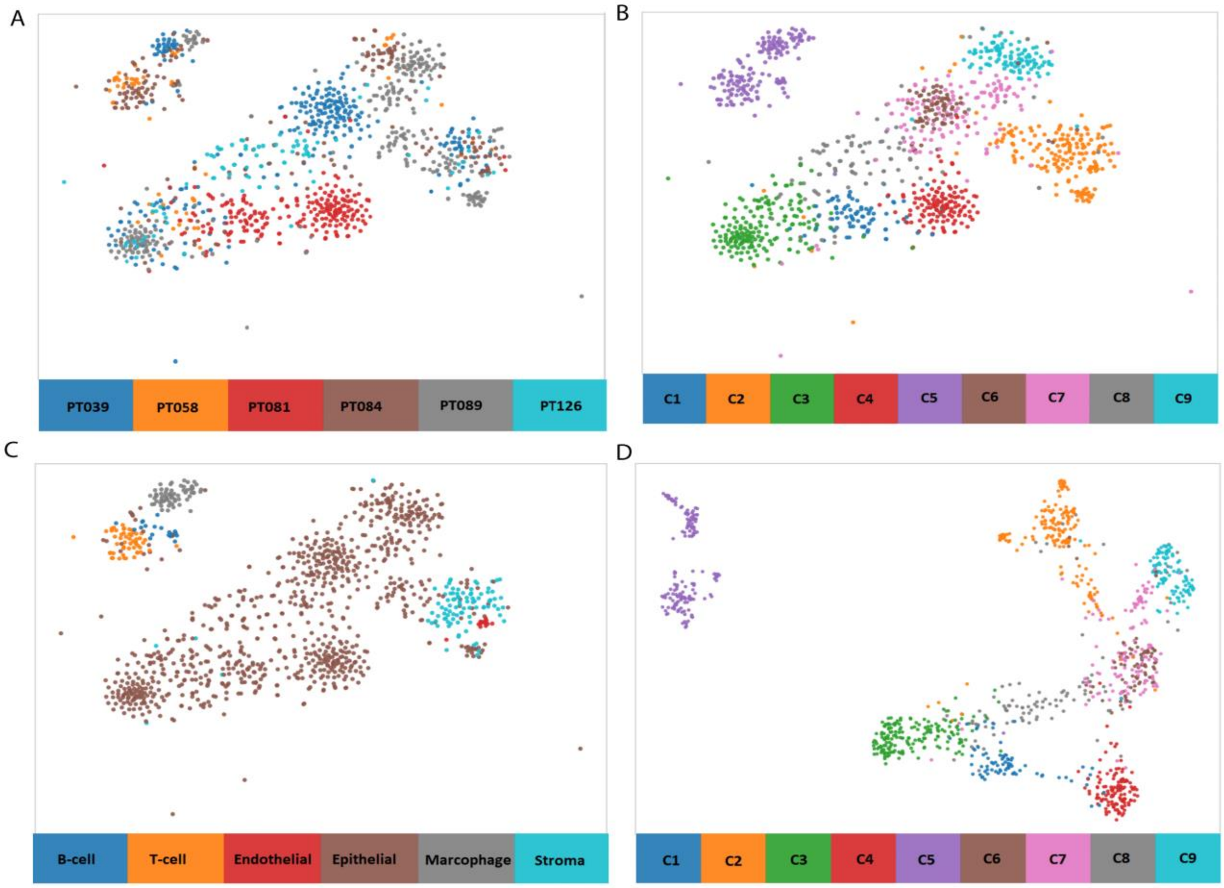


Figure 6. **Cellular heterogeneity of 6 primary TNBC tumors**. **A -** t-SNE plot of tumor cells mapped to their patient's identifier. It shows not clear patient-specific cells separation, which suggests a shared cell group among all patients. **B -** t-SNE plot of cells after DUSC application. The plot demonstrates the formation of 9 clusters. Most of the clusters overlap each other, while cluster 5 forms a distinct sub-group. **C -** t-SNE plot of cell type mapping, which shows that the most cell type group is represented by epithelial cells. The latter forms distinct subgroups. There is an immune cells cluster (T-cells, B-cells, and Macrophages), which is positioned separately and does not overlap with the others. **D -** Application of

unsupervised UMAP visualization tool on obtained by DUSC clusters. The plot demonstrates a gradual transcriptomic transition among the clusters represented by epithelial, stromal, and endothelial cells. The immune cell cluster is isolated and does not communicate with the rest of the cells.

In addition, this analysis revealed a graduate transition among cells clusters, which might suggest a shared transcriptomic profile of the defined cell clusters.

To determine how gene expression signatures differ between these clusters, we analyzed differentially expressed genes in various clusters using Seurat and obtained top 24 up- and down-regulated genes for each of the 9 clusters. The threshold of the number of top regulated genes was determined by the minimum of statistically significant up- and down-regulated genes of one of the clusters. To better understand the functional relevance of these different clusters in cancer and patient outcome, we further performed the survival analysis of the obtained gene cluster signatures using the Kaplan-Meier plots for two cohorts of untreated patients: TNBC-diagnosed and any breast cancer types diagnosed patients. The survival effect profile of the TNBC cluster markers (Fig. 7A) does not reflect the one obtained for all breast cancer patient cohort (Fig 7.B). Cluster 5 (immune cell cluster) (Fig. 7A) demonstrates the increased number of upregulated gene-markers associated with positive survival prognosis, while cluster 7 has an increased number of gene-signatures suggesting poor survival. Clusters 6, 7, 8 (epithelial cell clusters) markers are more likely to negatively affect the patients' survival rate in any breast cancer type diagnosed patients. In contrast to the TNBC cohort, the majority of the cluster 9 (stromal cells) gene-signatures demonstrate good survival prognosis in any breast cancer type diagnosed patients' cohort.

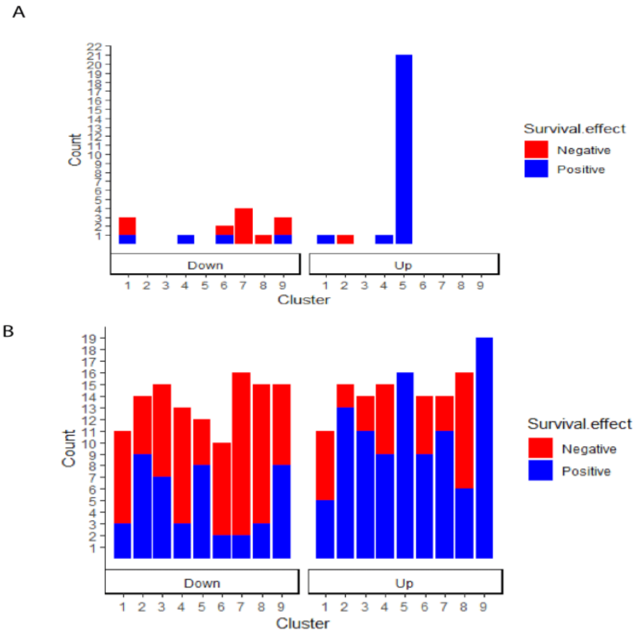


Figure 7. **Survival association of the gene signatures of 9 cell clusters obtained by DUSC**. Y-axis shows the number of genes, X-axis - cluster number and gene-signature regulation level. **A -** Survival effect of top 24 up- and down-regulated gene-signatures of 9 clusters for untreated TNBC-diagnosed patients' cohort. The bar plot shows that the majority of cluster 5 (immune cells cluster) gene-signatures are associated with better survival prognosis. Cluster 7 (epithelial cluster) gene-signatures are associated with poor survival rates. **B -** Survival effect of top 24 up- and down-regulated gene signatures of 9 clusters for any breast cancer type diagnosed patient cohort. Cluster 5 (immune cluster) and cluster 9 (stroma cluster) signatures demonstrate a positive effect on the survival rate, whereas clusters 6, 7, 8 (epithelial clusters) markers are associated with negative survival prognosis.

Gene pathway analysis was used to find out if the differentially expressed genes are associated with a certain biological process or molecular function. We obtained a list of the statistically significant biological processes associated with the differentially expressed genes for each cluster (Supplementary Table S12). Cluster 5 markers are involved in a range of important gene pathways like Ribosome, Spliceosome, RNA transport, Phagosome, Antigen processing, and presentation. Cluster 9 gene-markers affect Drug metabolism - cytochrome P450, Tyrosine metabolism, Antigen processing and presentation, and other important cellular processes.

We further investigated cluster 5 (immune cells cluster). We recursively applied DUSC on cluster 5 cells and obtained two subclusters C5.1 and C5.2 (Fig. 8C). Subcluster C5.1 contains mostly T- and B-cells (Fig. 8A), whereas subcluster 5.2 is mostly macrophage cells populated. The latter is shared among all patients (Fig. 8B), which makes it an interesting subgroup to study. We performed the differentially expressed gene analysis (DEGA) on subclusters C5.1 and C5.2 and obtained the Kaplan-Meier plots for the resulted subcluster gene-signatures. Most of the up-regulated gene-markers of subcluster 5.1 demonstrate positive survival prognosis (Fig 8D), while subcluster C5.2 upregulated genes-signatures show a mixed (suggested neutral) survival prognosis. The downregulated signatures of subcluster C5.2 demonstrate negative survival effect.

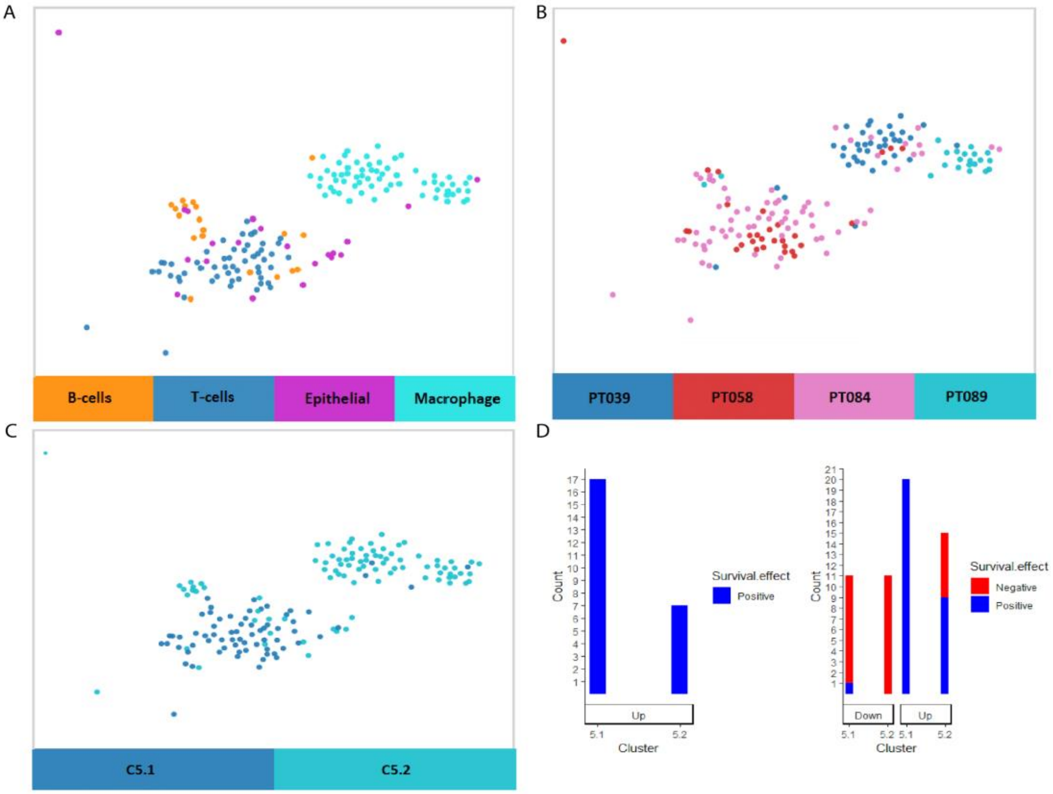


Figure 8. **Cluster 5 (immune cells cluster) intracellular heterogeneity. A -** t-SNE plot of cluster cells mapped to their cell types. Macrophage cells are mostly separated from the rest of the cells, whereas T-cells, B-cells, and epithelial cells are united in one subcluster. **B -** t-SNE plot of cluster 5 cells mapped to their patient-specific identifiers. Subclusters cells are shared among all patients and do not contain the patient' specific cells only. **C -** t-SNE plot of the subclusters shows the overlap between subclusters C5.1 and C5.2. **D –** Survival analysis of two subcluster gene-signatures. In TNBC patient's cohort (left) many of the signatures

of both of the subclusters are associated with the good survival effect. In all breast cancer patients' cohort (right), down-regulated signatures demonstrate poor survival prognosis for subcluster 5.1 signatures (T-cells and B-cells mostly) and for 5.1(macrophages).

Gene pathway enrichment analysis revealed C5.1 subcluster gene-markers' involvement in T cell receptor signaling pathway and complement and coagulation cascades affection and drug metabolism - cytochrome P450 pathway in subcluster C5.2 (Supplementary Table S12).

We further studied cluster 7 (epithelial cell cluster), which is shared among different patients (Fig. 9A). We applied DUSC on the cells and obtained two subclusters 7.1 and 7.2 (Fig. 8B).

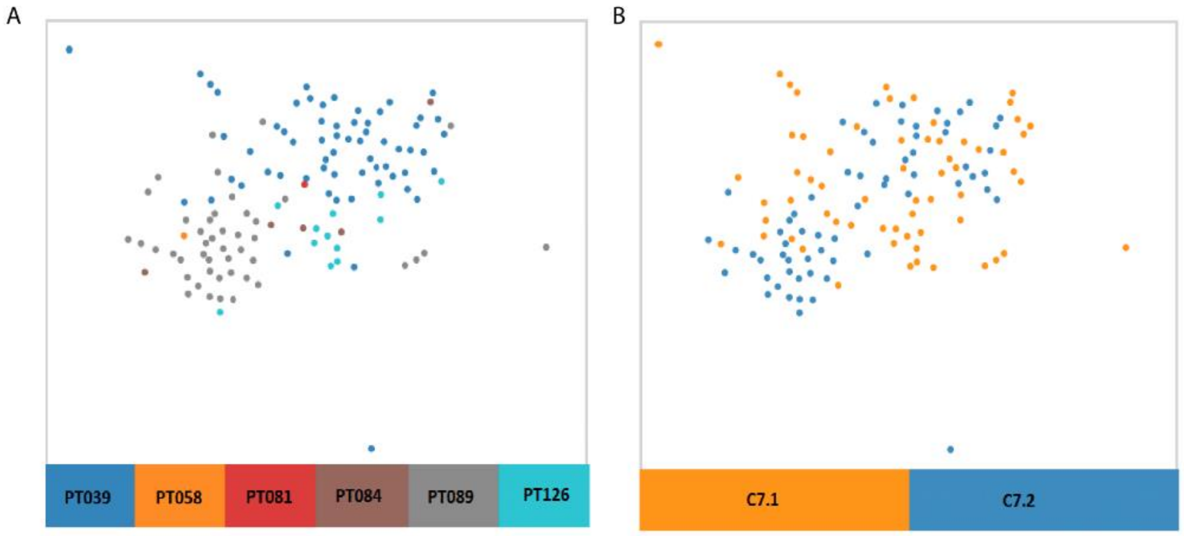


Figure 9. **Cluster 7 (epithelial cluster) intracellular heterogeneity**. **A -** t-SNE plot of cluster 7 cells mapped to patient-specific identifiers. **B -** t-SNE plot of subclusters C7.1 and C7.2 does not demonstrate clear cluster separation.

Further differentially expressed gene analysis did not reveal any statistically significant signatures between two subclusters.

3.2 TNBC intratumor heterogeneity

To study the TNBC heterogeneity within a tumor, we picked two patient's samples as case studies. Patients PT081 and PT039 have been previously reported to contain epithelial cancer cells and cancer subclones, which differ in copy number variations (CNVs). We applied DUSC on PT081 epithelial cells and obtained 4 clusters (Fig. 10A). To check how these clusters reflect the copy-number variation diversity in the tumor, we applied InferCNV computational tool (Fig. 10C), which infers CNVs from the transcriptomic data, and based on that performs hierarchical clustering of the input cell clusters. Sample PT081 CNVs analysis exposes the "main" sub-clone - cluster 4, enriched in CNVs, characterized by a gain in chromosome 1, 8 and 10 and a loss in chromosome 9. In contrast, cluster 3 demonstrates the opposite CNVs pattern, where there is a loss in chromosome 1 and 8. The dendrogram suggests that cluster 1, 2, and 3 originated from the "main" cluster 4, consistently evolving in CNVs. Given that most of the cells in PT081 sample are basal like-1 molecular subtype (Fig. 10B), we applied InferCNVs just for these cells to find out their impact on the tumor CNVs profile.

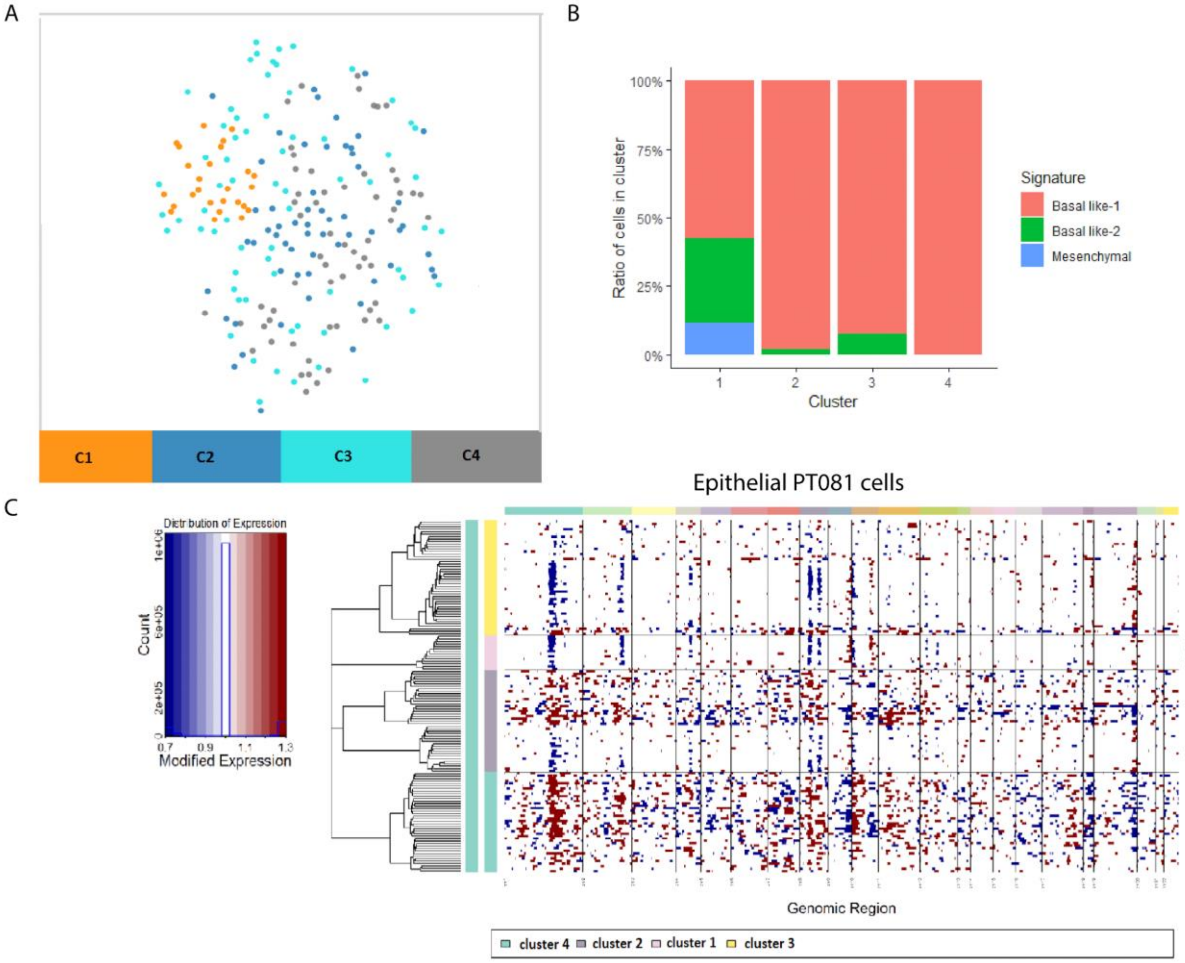


Figure 10. **Patient PT081 intratumor analysis. A** - t-SNE plot of clusters obtained by applying DUSC on epithelial PT081 cells. There are no distinctly separated clusters. **B -** Molecular subtype diversity within a PT081 tumor. The most of the epithelial cells is represented by basal like-1 molecular subtype type. **C**

- InferCNV results for PT081 epithelial cells reveal possible evolution pattern within a tumor. Cluster 4 could be a possible main clone and cluster 2 and three subclones.

Basal like-1 molecular subtype cells reflect the main trends in the tumor CNVs (Fig. 11) suggesting their great impact to the sample CNVs and the clonal evolution within the tumor.

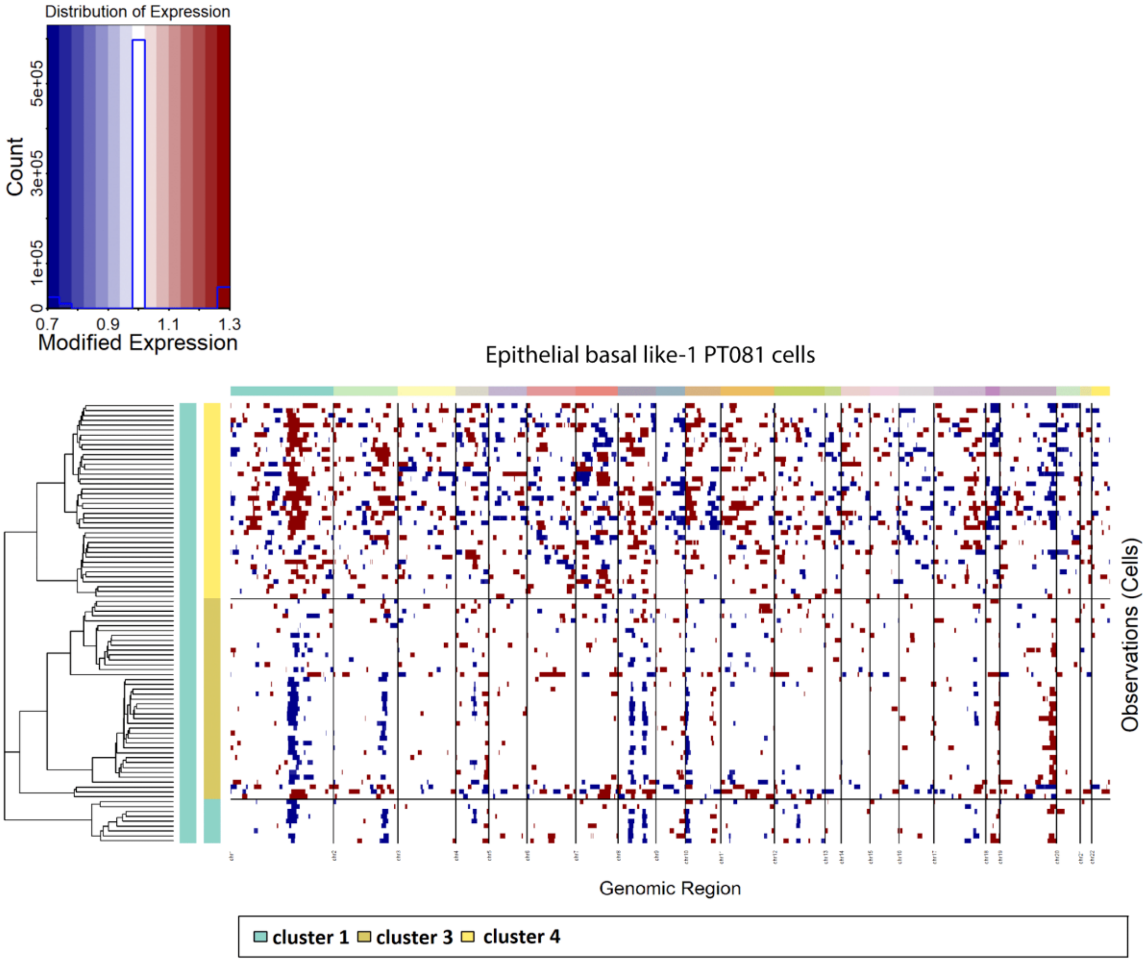


Figure 11. **Patient PT081 basal-like-1 InferCNVs analysis.** The pattern reflects the PT081 epithelial cells inferCNVs results suggesting basal like-1 cells great impact in the clonal diversity.

DUSC identified 3 clusters in PT039 epithelial cells (Fig. 12A), which were not clearly separated on the t-SNE plot. InferCNVs exposed clear CNVs patterns in chromosome 1 and 12 (Fig. 12C): cluster C1 had gains, and cluster C3 had losses in these genomic regions. The cells in clusters 1 and 3 demonstrate very similar CNV patterns yet were separated by DUSC. This suggests that one of these clusters of cells were originated from another with further genetic or epigenetic changes. The cells dendrogram obtained by InferCNV is consistent with this idea. DEGA did not reveal any statistically significant up- or downregulated genes in these clusters. Since the tumor PT039 cells were represented mostly by basal like-1 and mesenchymal molecular subtype (Fig. 12B), we checked the contribution of these cell molecular subtypes to the tumor CNVs. The basal like-1 cells resemble the pattern of all epithelial cells CNVs (Fig. 13A), which suggests their great contribution to the tumor CNVs. PT039 mesenchymal cell clusters looks more similar in CNVs, compared to basal like-1 clusters. This suggest that basal like-1 epithelial cells are more diverse in CNVs than mesenchymal cells (Fig. 13B).

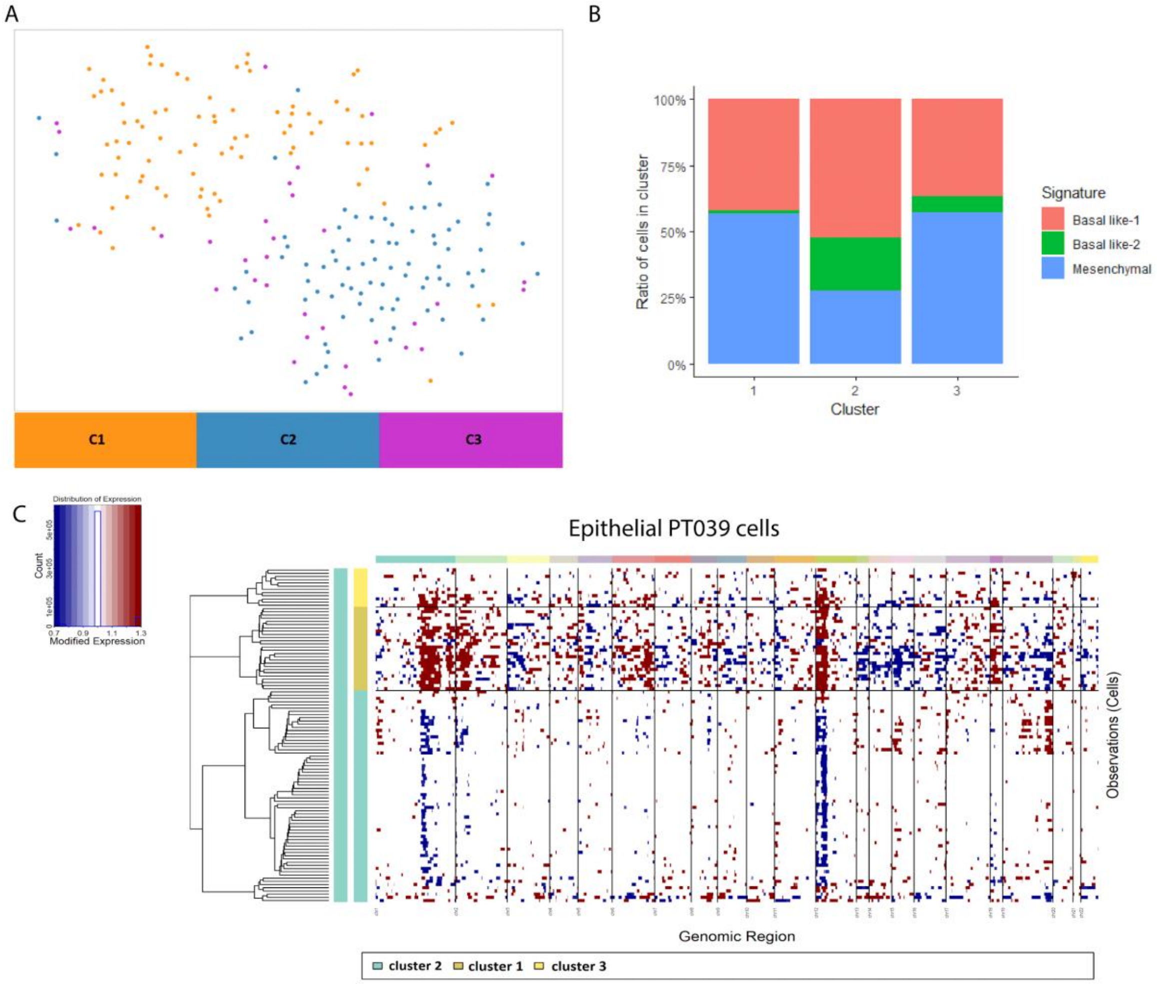


Figure 12. **Patient PT039 intratumor heterogeneity**. **A -** t-SNE plot of clusters obtained by applying DUSC on epithelial cells of tumor PT039. There are no distinctly separated clusters. **B -** Molecular subtype diversity within a PT039 tumor. The tumor mostly consists of basal like-1 and mesenchymal molecular

subtypes. **C** - InferCNV heatmap results for tumor PT039 epithelial cells illustrate possible evolution pattern within a tumor. Cluster C4 could be a possible main clone and cluster 2 and 3 - subclones.

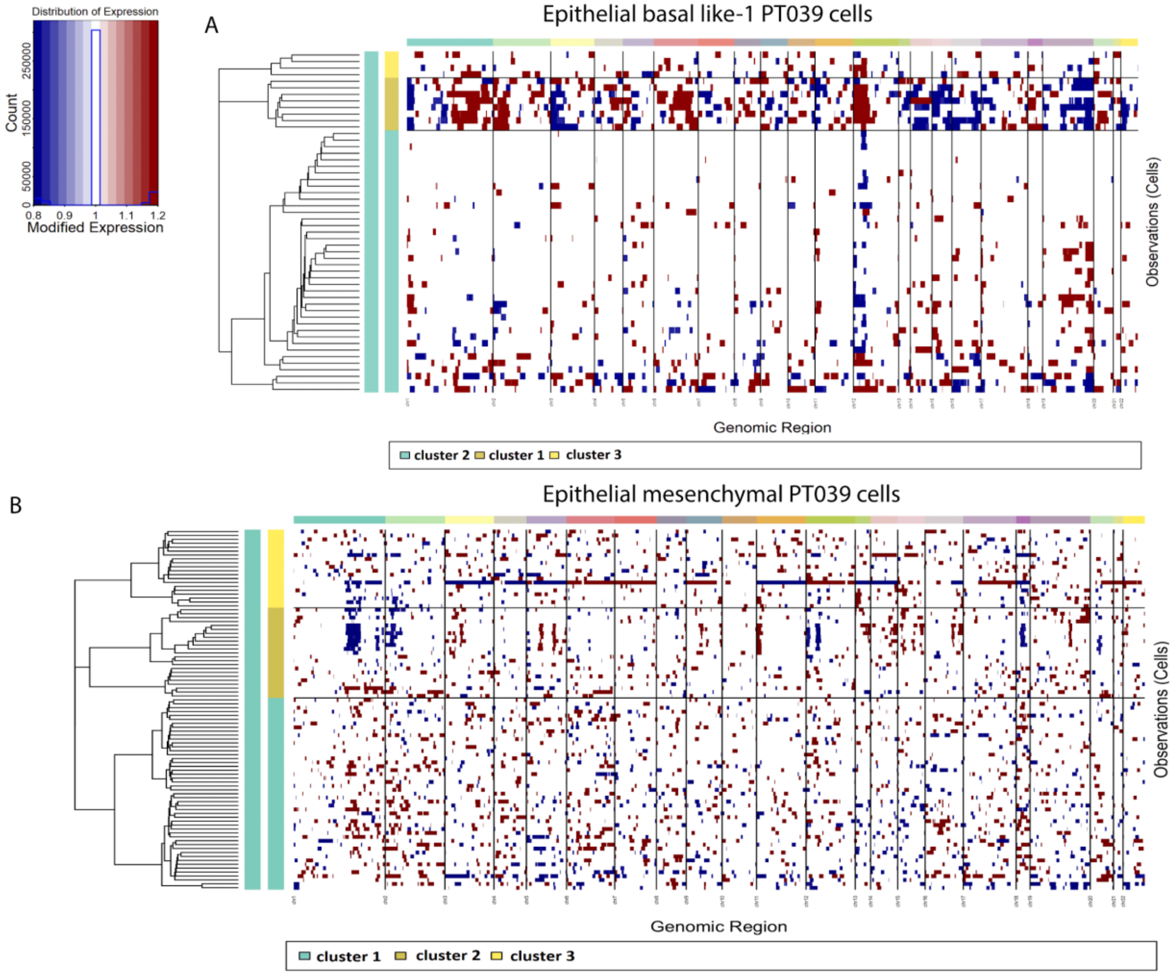


Figure 13. **PT039 mesenchymal and basal-like-1 cells analysis**. **A** - InferCNVs of basal like-1 epithelial cells. **B** - InferCNVs of mesenchymal-epithelial cells. The results suggest a contribution to the CNVs tumor pattern from the basal like-1 cells, rather than mesenchymal cells.

Basal like-1 epithelial cells showed their great impact on CNVs in both PT081 and PT039 tumors. We applied InferCNV (Fig. 14) on all basal like-1 epithelial cells across six tumors. The revealed CNVs pattern (high CNVs in PT081 and PT039) are concordant with the published data [1]. However, InferCNV illustrates high CNVs in PT089 and PT126, which was not reported previously. The obtained result indicates great CNVs in four out of six patients, suggesting that basal like-1 molecular subtype might be more diverse in CNVs than any other TNBC molecular subtype.

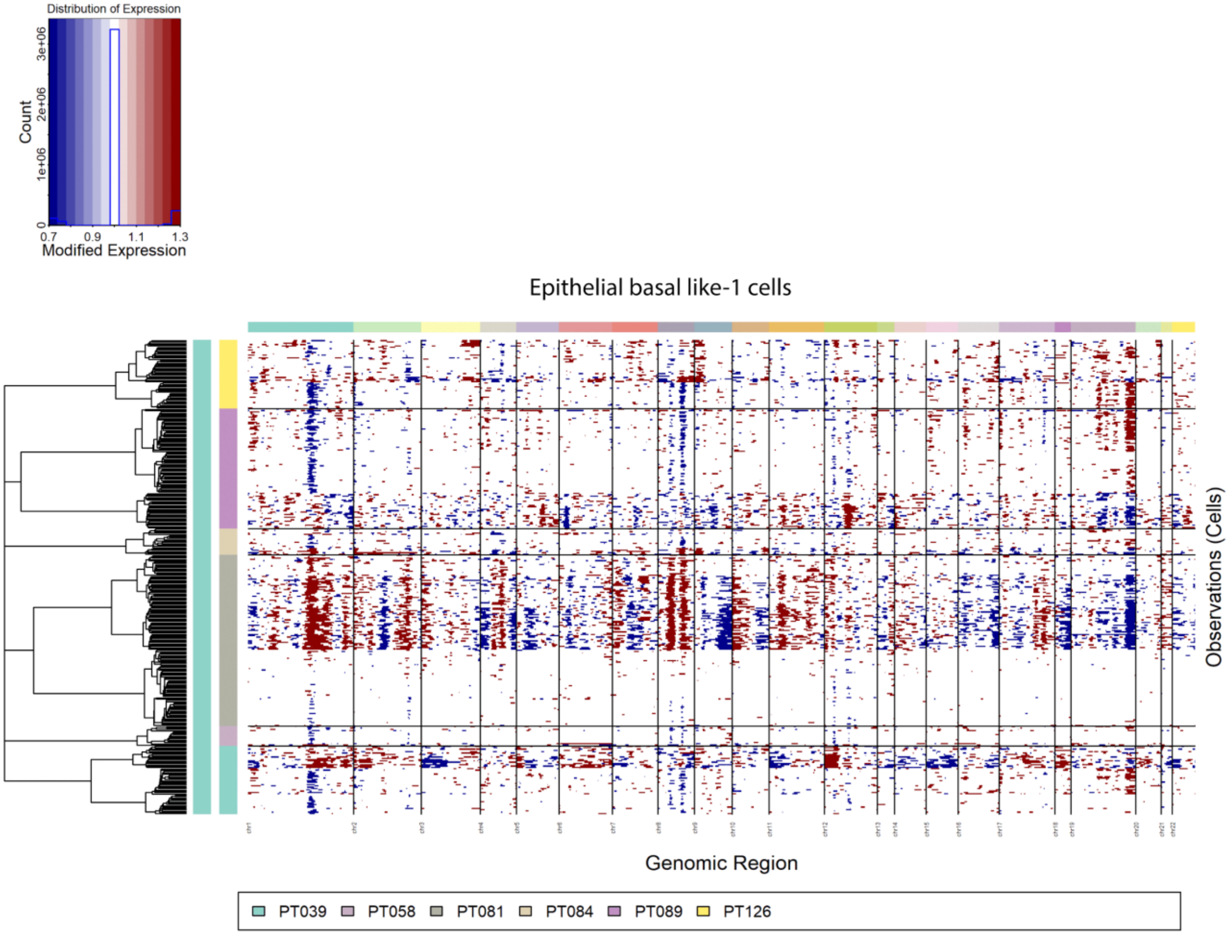


Figure 14. **Epithelial basal like-1 cell analysis across patients.** The CNVs pattern resembles the original paper insights (high CNVs in PT081 and PT039) in addition to the newly revealed high CNVs in PT089 and PT126.

4. Discussion

In this project, we analyze single-cell RNA-seq data of six TNBC tumors. We apply novel unsupervised clustering method DUSC to reveal new subgroups of cells shared among patients, and to determine which transcriptomic profiles might affect the patients' survival outcome. Our results indicate that some immune cells (macrophages, T-cells, B-cells) gene-signatures, could be interpreted as good survival prognosis markers for TNBC patients as well as for any other breast cancer type patients. Macrophages, which formed a distinct subcluster during all patients' cells DUSC analysis, were present in all patients' tumors samples and characterized by the signatures, associated with the survival prognosis. Their genomic signatures, which were involved in tyrosine, glutathione, and drug-cytochrome P450 metabolism could be used as potential biomarkers of the patient's survival prognosis. The drastic difference in gene-signatures survival association between TNBC patients and all breast cancer types patient cohorts might highlight cancer type-specific gene markers. For example, stromal cells gene-signatures positive survival association in any breast cancer type patient cohort might reveal a cell population that is not present in TNB but is in other breast cancer types.

The cells with the large-scale CNVs were hypothesized to be malignant cells and impact the patients' survival rates [1]. The effective detection of such cell subclones could be a potential prognosis metric for the patient outcome. Our results suggest that DUSC captures the subpopulations of basal like-1 epithelial cells sharing CNVs pattern. These findings are consistent with the whole exome sequence data of the tumors and their inferred CNVs obtained for this dataset using a different approach [1]. Besides, PT089 and PT126 are suggested to contain CNVs subclones, which was not reported previously. In addition, TNBC basal like-1 molecular subtype is suggested to be more diverse in CNVs than other molecular subtypes, like mesenchymal.

Future research could be directed into the TNBC tumor microenvironment studying. Since several TAM-associated TNBC-treatment strategies have been recently proposed [14], identification of malignant M2 stage of macrophages and their gene-signatures could be greatly beneficial. These gene-signatures could become potential biomarkers of patient's eligibility for TAM-associated therapies. Our survival analysis did not reveal any poor survival associated macrophages gene-signatures for TNBC patients. This could be due to that all macrophages in our dataset are in the stage M1. However, further experimental or computational analysis should be done to prove this hypothesis.

Collecting more data like increasing the number of analyzed cells would greatly contribute to the CNVs analysis and clonal evolution reconstruction. Some of the patients' tumors were represented just in 200-300 cells, which is not enough for the further CNVs analysis. Adding more cells to the dataset could help to understand the CNVs and their subclonal diversity within a tumor. Since the cells with large-scale CNVs are considered to be cancerous, revealing subclones of such cells, differ in CNVs in various genomic regions, could be beneficial for the survival prognosis of the patients.

In addition, attaching malignant-/non-malignant cells labels to the current dataset could potentially reveal many new biological insights of TNBC cells, ranging in molecular subtypes and cell types and their survival impact.

5. Supplementary materials

We used Seurat 2.0 to obtain a list of DEG. In the tables S1-S9 the output data frames columns are:

- p_val: unadjusted p-value;
- avg_logFC: average log fold change; the sign of the value shows the direction of the change;
- pct.1: percentage of the cells in the first cluster that have some gene expression (non-zero value);
- pct.2: percentage of the cells in the second cluster that have some gene expression (non-zero value);
- p_val_adj: Bonferroni corrected p-value;

Gene	**p_val**	**avg_logFC**	**pct.1**	**pct.2**	**p_val_adj**
HNRNPH1	2.65E-33	3.309946	0.986	0.872	4.33E-29
TRA2A	1.39E-32	2.870889	0.944	0.79	2.26E-28
PPP1CB	8.44E-32	2.837504	0.972	0.815	1.38E-27
PADI2	8.77E-27	1.935371	0.817	0.322	1.43E-22
ACTG1	7.47E-24	-2.65748	0.338	0.823	1.22E-19

PSAP	1.09E-23	-3.10497	0.141	0.754	1.78E-19
CFL1	1.16E-23	-1.19619	0.113	0.761	1.89E-19
RPL10	8.21E-23	-0.95225	0.085	0.755	1.34E-18
HNRNPL	1.07E-22	1.247621	0.859	0.73	1.75E-18
CALM1	2.71E-22	-1.76323	0.07	0.697	4.43E-18
SERF2	8.16E-22	-1.28159	0.127	0.787	1.33E-17
RPL13	8.93E-22	-1.67563	0.197	0.754	1.46E-17
OAZ1	3.43E-21	-1.67416	0.085	0.708	5.60E-17
SYNC	7.91E-21	1.223223	0.662	0.25	1.29E-16
ALDOA	7.99E-21	-2.32677	0.099	0.684	1.30E-16
C1orf56	9.98E-21	2.192791	0.859	0.668	1.63E-16
RPLP0	1.83E-20	-1.34426	0.183	0.762	3.00E-16
CTNNB1	2.37E-20	2.207563	0.887	0.813	3.88E-16
TPM3	6.39E-20	-0.99394	0.465	0.847	1.04E-15
RPL8	1.83E-19	-1.1955	0.211	0.768	2.99E-15
TFAP2C	2.18E-19	2.420334	0.817	0.488	3.57E-15
RPS3	3.15E-19	-1.23766	0.127	0.725	5.15E-15

HLA-C	8.32E-19	-1.55234	0.113	0.702	1.36E-14
SRSF7	1.47E-18	-1.2531	0	0.596	2.41E-14
EEF2	1.64E-18	-1.57623	0.127	0.677	2.67E-14
EEF1G	4.07E-18	-1.50315	0.31	0.762	6.65E-14
GLG1	4.19E-18	-1.30107	0.211	0.721	6.84E-14
SRSF5	7.95E-18	-1.07005	0.056	0.632	1.30E-13
GUK1	1.14E-17	-1.27998	0	0.577	1.86E-13

Table S1. DEG list (fragment) for cluster 1.

Gene	p_val	avg_logFC	pct.1	pct.2	p_val_adj
SERPING1	4.72E-95	2.368978	0.858	0.178	7.72E-91
RBMS3	4.30E-93	1.250729	0.597	0.035	7.02E-89
RARRES2	1.74E-85	1.583086	0.625	0.059	2.84E-81
COL1A2	4.64E-83	3.19924	0.636	0.073	7.58E-79
TIMP3	4.32E-82	2.530081	0.67	0.088	7.05E-78
BGN	1.05E-81	1.671627	0.483	0.021	1.71E-77
SPARC	3.10E-80	1.6329	0.847	0.229	5.06E-76
SPARCL1	9.08E-80	2.566908	0.682	0.104	1.48E-75
COL6A2	2.25E-79	1.319615	0.58	0.054	3.67E-75
CCDC80	2.95E-79	1.890668	0.517	0.032	4.81E-75
LAMA4	3.42E-77	1.573028	0.528	0.04	5.59E-73
C1S	1.05E-75	2.747192	0.744	0.159	1.72E-71
THY1	1.99E-75	1.539258	0.438	0.016	3.25E-71
COL3A1	5.20E-75	3.238868	0.557	0.055	8.50E-71
NNMT	1.99E-73	1.643841	0.722	0.133	3.24E-69
MMP2	5.12E-73	2.405058	0.568	0.062	8.37E-69

C1R	4.35E-72	2.309258	0.739	0.163	7.11E-68
COL6A1	1.39E-70	1.359883	0.568	0.064	2.27E-66
COL1A1	8.61E-70	2.847557	0.682	0.131	1.41E-65
CXCL12	8.00E-69	2.174506	0.5	0.043	1.31E-64
DCN	2.36E-68	3.308089	0.602	0.089	3.86E-64
SERPINF1	2.73E-68	1.559197	0.574	0.076	4.46E-64
CAV1	2.38E-67	1.746373	0.58	0.076	3.88E-63
TIMP1	2.05E-66	1.696699	0.795	0.22	3.34E-62
COL6A3	6.02E-66	2.547719	0.483	0.043	9.83E-62
IGFBP4	2.94E-65	1.099144	0.591	0.081	4.81E-61
MYL9	3.02E-65	0.665549	0.688	0.145	4.94E-61
CFI	3.46E-65	1.889476	0.415	0.022	5.65E-61
ANGPTL2	6.82E-65	0.918793	0.375	0.013	1.11E-60

Table S2. DEG list (framgent) for cluster 2.

Gene	p_val	avg_logFC	pct.1	pct.2	p_val_adj
EIF5A	9.99E-59	-1.2367	0.031	0.906	1.63E-54
H3F3B	9.92E-58	-1.88907	0.359	0.972	1.62E-53
CALR	3.06E-56	-1.61258	0.078	0.922	5.00E-52
PFN1	3.21E-56	-2.29203	0.078	0.891	5.24E-52
RPS6	1.32E-55	-0.89037	0.055	0.897	2.15E-51
RPS18	3.42E-53	-1.81492	0.039	0.852	5.59E-49
SON	8.05E-53	-1.51069	0.039	0.867	1.32E-48
RPL41	2.02E-52	-1.17847	0.086	0.917	3.30E-48
RPL5	3.90E-52	-1.60032	0.07	0.875	6.38E-48
RPS8	2.58E-51	-1.31114	0	0.811	4.22E-47
RPS25	2.63E-51	-1.28366	0.039	0.867	4.29E-47
RPS27	2.70E-51	-0.60079	0.055	0.9	4.42E-47
MYL6	8.15E-51	-2.01842	0.07	0.849	1.33E-46
RPS14	8.31E-51	-1.38222	0.031	0.842	1.36E-46
FTH1	1.34E-50	-0.90039	0.289	0.96	2.19E-46
HNRNPL	3.86E-50	-0.45906	0.023	0.838	6.31E-46
CALM2	4.17E-50	-2.30577	0.219	0.895	6.82E-46
RPS19	5.16E-50	-1.49843	0.062	0.857	8.44E-46
RPL3	6.99E-50	-2.51727	0.109	0.851	1.14E-45
EEF1A1	1.98E-49	-1.32457	0.227	0.935	3.23E-45
PTMA	2.17E-49	-1.22641	0.133	0.917	3.54E-45
CTNNB1	2.54E-49	-2.08749	0.133	0.913	4.15E-45
SRSF6	2.76E-49	-2.53511	0.172	0.881	4.51E-45

Table S3. DEG (fragment) list obtained by Seurat for cluster 3.

Gene	p_val	avg_logFC	pct.1	pct.2	p_val_adj
LTF	9.69E-98	2.7037	0.888	0.123	1.58E-93
CDKN2A	2.27E-79	0.697913	0.816	0.104	3.71E-75
C1orf186	7.50E-72	0.665698	0.544	0.039	1.22E-67
ORM2	4.86E-67	0.62828	0.392	0.012	7.94E-63
ITGB8	7.05E-63	2.365868	0.952	0.414	1.15E-58
C12orf45	1.63E-61	0.776732	0.72	0.129	2.66E-57
PSMB4	2.24E-61	1.648755	0.984	0.571	3.66E-57
SLC35F2	5.80E-61	1.48775	0.84	0.205	9.47E-57
BCL2A1	1.24E-59	1.133044	0.664	0.096	2.03E-55
CCL20	4.41E-59	1.121516	0.528	0.051	7.21E-55
ETNK1	2.93E-55	1.66502	0.984	0.593	4.79E-51
TLK1	4.57E-55	1.19704	0.992	0.522	7.47E-51
TRPS1	7.73E-55	1.29116	0.976	0.475	1.26E-50
YWHAE	1.35E-54	1.277587	1	0.742	2.20E-50
ORM1	6.23E-54	0.326706	0.344	0.014	1.02E-49
PRRT3-AS1	9.96E-54	0.331575	0.408	0.028	1.63E-49
SOX11	1.32E-52	1.042883	0.896	0.263	2.15E-48
LARP4B	6.17E-52	1.440543	0.896	0.35	1.01E-47
GRHL2	7.90E-52	0.655504	0.944	0.323	1.29E-47
AEBP2	2.01E-51	1.355363	0.952	0.442	3.28E-47
ANP32E	3.96E-51	1.537728	0.928	0.461	6.48E-47
BIRC3	3.89E-50	1.763789	0.808	0.263	6.35E-46
UBE2V2	1.03E-49	1.154713	0.888	0.334	1.68E-45

INSR	1.28E-49	1.515857	0.968	0.52	2.10E-45
EIF2S3	4.52E-49	1.257828	0.984	0.75	7.39E-45
TRA2A	2.00E-48	0.773919	1	0.774	3.26E-44
NUDT19	2.17E-48	1.753247	0.976	0.729	3.55E-44
TBL1XR1	3.67E-48	1.55923	0.984	0.629	5.99E-44
DNAJC3	3.79E-48	1.464214	0.952	0.574	6.19E-44

Table S4. DEG list (fragment) list for cluster 4 obtained by Seurat.

Gene	p_val	avg_logFC	pct.1	pct.2	p_val_adj
CD53	1E-139	2.090139	0.769	0.01	1.7E-135
LAPTM5	1.7E-130	2.969218	0.827	0.04	2.8E-126
SRGN	2.5E-122	2.033341	0.929	0.097	4.1E-118
PTPRC	6E-113	2.324262	0.891	0.114	9.9E-109
RGS1	2.5E-111	3.19317	0.731	0.038	4E-107
ITGB2	6.9E-108	2.329949	0.692	0.03	1.1E-103
IL2RG	2.4E-106	2.015151	0.628	0.014	4E-102
CORO1A	7.1E-103	2.258112	0.686	0.036	1.17E-98
EVI2B	1.1E-99	1.445047	0.564	0.006	1.81E-95
SAMSN1	7.04E-98	1.435302	0.564	0.008	1.15E-93
FYB	4.04E-97	1.829954	0.705	0.048	6.6E-93
LCP1	3.54E-94	2.56124	0.853	0.134	5.79E-90
ALOX5AP	6.14E-94	1.200419	0.609	0.022	1E-89
CYTIP	6.07E-93	1.68834	0.526	0.005	9.91E-89
RAC2	3.41E-92	1.733411	0.551	0.012	5.57E-88
HCLS1	1.9E-90	1.677458	0.66	0.044	3.1E-86
IL10RA	6.75E-87	1.680911	0.532	0.013	1.1E-82
CD52	4.81E-86	1.189366	0.526	0.013	7.86E-82

ARHGAP9	2.88E-84	1.220137	0.455	0	4.71E-80
SELPLG	6.3E-83	1.845598	0.519	0.016	1.03E-78
LCP2	2.52E-80	0.438209	0.545	0.023	4.11E-76
C16orf54	1.47E-79	2.046384	0.635	0.06	2.4E-75
CD69	6.93E-79	2.746239	0.564	0.034	1.13E-74
GMFG	1.28E-78	0.872346	0.571	0.034	2.09E-74
DOCK2	1.78E-77	1.321824	0.442	0.004	2.9E-73
ITGA4	2.14E-77	1.262795	0.532	0.025	3.5E-73
PARVG	8.33E-77	0.866198	0.417	0	1.36E-72
GPSM3	1.18E-76	0.361837	0.494	0.016	1.93E-72
ARHGDIB	1.21E-76	2.17519	0.929	0.303	1.98E-72

Table S5. GEG list (fragment) obtained by Seurat for cluster 5.

Gene	p_val	avg_logFC	pct.1	pct.2	p_val_adj
KRT86	2.3E-104	0.396571	0.833	0.048	3.7E-100
KRT81	9.51E-97	2.588612	0.846	0.066	1.55E-92
DSG1	1.01E-95	1.376917	0.808	0.051	1.65E-91
KIF1A	5.35E-91	0.537983	0.577	0.016	8.75E-87
TTYH1	9.21E-88	0.801265	0.705	0.04	1.5E-83
HEY2	1.16E-87	0.664853	0.692	0.037	1.9E-83
UCHL1	1.85E-84	0.804932	0.641	0.032	3.02E-80
FBN3	5.18E-84	0.331529	0.487	0.01	8.46E-80
CDH2	2.6E-82	0.919117	0.692	0.043	4.24E-78
AIF1L	1.25E-81	1.413457	0.885	0.092	2.04E-77
ETV4	6.12E-78	0.316419	0.615	0.031	1E-73
IGSF9	6.13E-75	0.635562	0.654	0.042	1E-70
USP5	4.39E-72	1.098142	0.846	0.102	7.18E-68
NLRP2	1.68E-71	0.621777	0.705	0.058	2.74E-67
DHCR7	4.2E-70	1.323806	0.91	0.122	6.86E-66
PRR4	8.93E-70	0.779228	0.667	0.049	1.46E-65
DBN1	1.12E-68	0.554811	0.782	0.081	1.83E-64

GUCY1A3	1.39E-66	1.432863	0.923	0.141	2.28E-62
HORMAD1	1.73E-66	0.539814	0.718	0.066	2.83E-62
KRT5	6.01E-66	1.003278	0.91	0.128	9.82E-62
SORBS2	9.81E-66	1.035304	0.91	0.136	1.6E-61
CLSTN3	4.5E-65	0.563158	0.91	0.131	7.35E-61
RERG	5.81E-65	1.173296	0.821	0.101	9.49E-61
SCARB1	8.88E-65	0.720235	0.769	0.086	1.45E-60
FAM222B	3.21E-64	0.486058	0.833	0.105	5.24E-60
PTK7	3.48E-64	0.824088	0.833	0.103	5.69E-60
PXDN	5.82E-64	1.483077	0.91	0.153	9.51E-60
AKT3	1.26E-63	0.529611	0.808	0.094	2.05E-59
MUC5B	2.9E-63	0.366972	0.769	0.089	4.74E-59

Table S6. DEG list (fragment) for cluster 6 obtained by Seurat.

Gene	p_val	avg_logFC	pct.1	pct.2	p_val_adj
KIT	2.86E-47	1.474181	0.683	0.148	4.67E-43
TRIM2	1.67E-40	1.253683	0.722	0.191	2.73E-36
NDRG2	1.09E-39	1.352391	0.825	0.301	1.77E-35
KRT15	5.34E-36	1.593123	0.603	0.156	8.73E-32
CLDN4	4.67E-34	1.393095	0.849	0.336	7.63E-30
DSC2	2.89E-32	1.309256	0.817	0.334	4.71E-28
SLC6A14	2.35E-31	1.171813	0.659	0.235	3.84E-27
EHF	4.25E-31	0.965804	0.841	0.33	6.95E-27
TM4SF1	1.08E-30	1.042694	0.889	0.444	1.76E-26
SOX9	1.26E-30	0.57226	0.706	0.232	2.06E-26
FAM60A	4.42E-30	1.001657	0.849	0.393	7.23E-26
PRSS8	5.08E-30	1.042537	0.587	0.176	8.30E-26
GUCY1A3	2.99E-28	0.940687	0.563	0.15	4.89E-24
EPCAM	2.40E-27	0.739339	0.817	0.352	3.93E-23
CHST9	3.05E-26	0.758166	0.349	0.058	4.98E-22
SFRP1	1.63E-25	0.940651	0.643	0.227	2.66E-21

EFNA1	2.78E-25	0.877305	0.786	0.344	4.54E-21
APP	4.17E-25	0.872372	0.698	0.275	6.82E-21
GAS5	1.04E-24	0.584502	0.833	0.428	1.71E-20
SHANK2	1.33E-24	0.452045	0.516	0.136	2.18E-20
DSP	2.48E-24	0.800389	0.81	0.363	4.05E-20
RPS6	3.45E-24	0.767384	0.96	0.773	5.64E-20
TACSTD2	9.02E-24	1.222521	0.81	0.404	1.47E-19
PERP	1.43E-23	0.975681	0.802	0.361	2.33E-19
DSG2	6.80E-23	0.955243	0.897	0.607	1.11E-18
PKP2	1.13E-22	1.194007	0.595	0.23	1.85E-18
KRT23	1.73E-22	1.283966	0.571	0.198	2.82E-18
LRP6	2.02E-22	1.03786	0.603	0.213	3.30E-18
GABRP	4.64E-22	1.201761	0.643	0.259	7.59E-18

Table S7. DEG list (fragment) obtained by Seurat for cluster 7.

Gene	p_val	avg_logFC	pct.1	pct.2	p_val_adj
TMSB10	1.08E-23	1.750295	0.904	0.79	1.77E-19
AZGP1	4.57E-21	1.373789	0.777	0.378	7.46E-17
SRSF6	1.41E-19	-0.83594	0.394	0.834	2.30E-15
KLF6	5.60E-17	-0.93897	0.287	0.783	9.15E-13
RNASE1	8.92E-17	1.630199	0.436	0.129	1.46E-12
CLK1	9.17E-17	-0.85406	0.106	0.607	1.50E-12
TMBIM6	1.64E-16	-0.91279	0.255	0.749	2.67E-12
MTRNR2L2	2.39E-16	1.049068	0.936	0.906	3.91E-12
TOR1AIP2	1.06E-15	-1.21386	0.862	0.925	1.73E-11
DNAJC3	1.11E-15	-1.39944	0.213	0.658	1.81E-11
PPP1CB	1.30E-15	-1.93755	0.436	0.864	2.12E-11
PPP3CA	1.51E-15	-0.6891	0.213	0.699	2.47E-11
STRN3	2.25E-15	-1.22115	0.074	0.524	3.67E-11
R3HDM2	2.37E-15	-0.62217	0.223	0.664	3.87E-11
CALML5	2.46E-15	2.227362	0.479	0.198	4.01E-11
PAFAH1B2	7.34E-15	-1.02663	0.511	0.795	1.20E-10

MCL1	1.16E-14	-0.99164	0.149	0.59	1.89E-10
ANKRD10	2.05E-14	-0.52463	0.245	0.699	3.36E-10
ITM2B	2.81E-14	-0.96051	0.213	0.61	4.59E-10
DDX5	3.04E-14	-0.74931	0.404	0.783	4.96E-10
DDX3X	3.46E-14	-0.58579	0.309	0.746	5.65E-10
ZFP36L2	4.43E-14	-0.47689	0.138	0.588	7.23E-10
MIF	4.64E-14	1.675903	0.777	0.72	7.58E-10
TBL1XR1	5.67E-14	-0.94393	0.298	0.707	9.26E-10
WTAP	6.88E-14	-1.2055	0.298	0.772	1.12E-09
MTRNR2L8	8.22E-14	0.951719	0.915	0.912	1.34E-09
RUFY3	8.40E-14	-0.61904	0.106	0.542	1.37E-09
NOP58	9.27E-14	-1.43715	0.223	0.63	1.51E-09
KRT19	1.09E-13	1.544979	0.723	0.482	1.79E-09

Table S8. DEG list (fragment) for cluster 8 obtained by Seurat.

Gene	p_val	avg_logFC	pct.1	pct.2	p_val_adj
AGR2	3.1E-144	2.045165	0.81	0.022	5.1E-140
ANKRD30A	4.1E-142	2.898919	0.87	0.037	6.7E-138
TTC39A	3.6E-138	1.642882	0.79	0.023	5.9E-134
MLPH	1.3E-131	1.978822	0.88	0.05	2.1E-127
AGR3	5.4E-127	0.894156	0.78	0.031	8.9E-123
TFAP2B	4.1E-126	1.031095	0.71	0.018	6.7E-122
SERPINA5	2.8E-119	1.742475	0.7	0.022	4.5E-115
PIP	7.4E-115	2.898264	0.64	0.016	1.2E-110
TFF3	2.2E-112	0.908269	0.59	0.009	3.6E-108
SPDEF	2.1E-110	1.300739	0.68	0.025	3.4E-106
TFF1	3.6E-106	1.569557	0.48	0	5.9E-102
EFHD1	5.7E-103	1.177751	0.72	0.039	9.4E-99
NEK10	5.4E-99	0.847011	0.55	0.013	8.76E-95
PGR	1.18E-97	0.784987	0.5	0.006	1.92E-93
CCDC74A	1.05E-95	1.300041	0.67	0.036	1.71E-91
REEP6	2.38E-95	0.778097	0.57	0.018	3.9E-91

DNAJC12	4.32E-95	0.656242	0.53	0.012	7.05E-91
C1orf168	1.3E-94	0.628848	0.47	0.004	2.12E-90
ZG16B	5.05E-94	0.590956	0.57	0.019	8.26E-90
CCDC74B	2.32E-90	0.524514	0.48	0.007	3.79E-86
CAPN8	3.45E-90	0.584486	0.49	0.009	5.64E-86
TBX3	3.79E-90	1.336033	0.79	0.074	6.18E-86
SERPINA3	1.18E-88	2.789551	0.89	0.128	1.92E-84
TSPAN1	2.41E-88	0.969208	0.54	0.018	3.93E-84
ESR1	9.84E-84	1.168974	0.58	0.028	1.61E-79
SEC14L2	1.21E-83	1.444746	0.77	0.08	1.97E-79
ADIRF	1.23E-83	0.850953	0.95	0.138	2.01E-79
VAV3	4.06E-83	1.366785	0.84	0.1	6.62E-79
DHCR24	1.48E-82	2.045319	0.85	0.117	2.42E-78

Table S9. DEG list (fragment) obtained by Seurat for cluster 9.

Gene pathway enrichment analysis was performed using the R package GAGE [32].

5.1_down	5.1_up	5.2_down	5.2_up
Complement and coagulation cascades ECM-receptor interaction Phagosome Protein digestion and absorption Focal adhesion	T cell receptor signaling pathway Jak-STAT signaling pathway Phosphatidylinositol signaling system Natural killer cell mediated cytotoxicity Fc epsilon RI signaling pathway	Drug metabolism - cytochrome P450 Tyrosine metabolism Glutathione metabolism Valine, leucine, and isoleucine degradation Fatty acid metabolism	Drug metabolism - cytochrome P450 Tyrosine metabolism Glutathione metabolism Valine, leucine and isoleucine degradation Fatty acid metabolism

Table S10. Gene enrichment pathway analysis for cluster 5 (immune cells cluster). The column names indicate the number of cluster and the type of DEG gene regulation. The pathways are listed in the decreasing manner of their p-values.

1_down	Lysosome Cell adhesion molecules (CAMs) Glycolysis / Gluconeogenesis Antigen processing and presentation Leukocyte transendothelial migration
1_up	Oxidative phosphorylation Ribosome biogenesis in eukaryotes Spliceosome Ribosome Proteasome
2_down	Oxidative phosphorylation Antigen processing and presentation Purine metabolism NOD-like receptor signaling pathway Proteasome
2_down	ECM-receptor interaction Focal adhesion Protein digestion and absorption Complement and coagulation cascades TGF-beta signaling pathway

3_down	Spliceosome Protein processing in endoplasmic reticulum RNA transport Ribosome Proteasome
3_up	Apoptosis Jak-STAT signaling pathway Bile secretion Adipocytokine signaling pathway Tryptophan metabolism
4_down	Ribosome Lysosome Complement and coagulation cascades Antigen processing and presentation Jak-STAT signaling pathway
4_up	Proteasome Ubiquitin mediated proteolysis RNA transport Ribosome biogenesis in eukaryotes Spliceosome
5_down	Ribosome Spliceosome RNA transport Ribosome biogenesis in eukaryotes Valine, leucine and isoleucine degradation

5_up	Antigen processing and presentation Phagosome Lysosome Osteoclast differentiation Natural killer cell mediated cytotoxicity
6_down	Ribosome Antigen processing and presentation NOD-like receptor signaling pathway Oxidative phosphorylation RIG-I-like receptor signaling pathway
6_up	Glycolysis / Gluconeogenesis Protein processing in endoplasmic reticulum Gap junction Adherens junction Aminoacyl-tRNA biosynthesis
7_down	Apoptosis Antigen processing and presentation NOD-like receptor signaling pathway Phagosome Lysosome
7_up	Ribosome Spliceosome Adherens junction

	hsa03013 RNA transport junction
8_down	Jak-STAT signaling pathway Osteoclast differentiation Apoptosis NOD-like receptor signaling pathway Focal adhesion
8_up	Ribosome Oxidative phosphorylation hsa03040 Spliceosome Cardiac muscle contraction Proteasome
9_down	Antigen processing and presentation NOD-like receptor signaling pathway Phagosome Focal adhesion Apoptosis
9_up	Drug metabolism - cytochrome P450 Tyrosine metabolism Glutathione metabolism Valine, leucine and isoleucine degradation Fatty acid metabolism

Table S12. Gene pathway enrichment analysis for 9 clusters (up- and down-regulated DEG) obtained by DUSC. The column names indicate the number of a cluster and the type of DEG gene regulation. The pathways are listed in the decreasing manner of their p-values.

References

[1] Karaayvaz, M., Cristea, S., Gillespie, S.M., Patel, A.P., Mylvaganam, R., Luo, C.C., Specht, M.C., Bernstein, B.E., Michor, F. and Ellisen, L.W., 2018. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature communications*, *9*(1), p.3588.

[2] A Hybrid Deep Clustering Approach for Robust Cell Type Profiling Using Single-cell RNA-seq Data Suhas Srinivasan, Nathan T Johnson, Dmitry Korkin, bioRxiv 511626; doi: https://doi.org/10.1101/511626

[3] Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.A., 2008, July. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103). ACM.

[4] Comprehensive integration of single cell data, Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck III, Marlon Stoeckius, Peter Smibert, Rahul Satija

bioRxiv 460147; doi: https://doi.org/10.1101/460147

[5] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R., 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, *36*(5), p.411.

[6] Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*, *143*(3), 331–336. doi:10.1016/j.otohns.2010.05.007

[7] Győrffy, B., Lanczky, A., Eklund, A.C., Denkert, C., Budczies, J., Li, Q. and Szallasi, Z., 2010. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast cancer research and treatment*, *123*(3), pp.725-731.

[8] inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV

[9] Stewart BW, Wild CP, editors. World cancer report 2014

International Agency for Research on Cancer; 2014.

[10] Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 *International Agency for Research on Cancer*; 2013.

[11] GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016 Oct; 388 (10053):1659-1724.

[12] Plummer M, de Martel C, Vignat J, Ferlay J, Bray F, Franceschi S. Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob Health.* 2016 Sep;4(9):e609-16. doi: 10.1016/S2214-109X(16)30143-7.

[13] Global Initiative for Cancer Registry Development. International Agency for Research on Cancer. Lyon: France.

[14] Santoni, M., Romagnoli, E., Saladino, T., Foghini, L., Guarino, S., Capponi, M., Giannini, M., Cognigni, P.D., Ferrara, G. and Battelli, N., 2018. Triple negative breast cancer: key role of tumor-associated macrophages in regulating the activity of anti-PD-1/PD-L1 agents. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, *1869*(1), pp.78-84.

[15] TV Ajithkumar, HM Hatcher, in Specialist Training in Oncology, 2011).

[16] Lehmann, B.D., Jovanović, B., Chen, X., Estrada, M.V., Johnson, K.N., Shyr, Y., Moses, H.L., Sanders, M.E. and Pietenpol, J.A., 2016. Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. *PloS one*, *11*(6), p.e0157368.

[17] Chapter 26 - Molecular Biology of Breast Cancer - Philippe Aftimos MD Hatem A. Azim Jr. MD, PhD Christos Sotiriou MD, PhD https://doi.org/10.1016/B978-0-12-802761-5.00026-2

[18] Takenaka, M.C., Gabriely, G., Rothhammer, V., Mascanfroni, I.D., Wheeler, M.A., Chao, C.C., Gutiérrez-Vázquez, C., Kenison, J., Tjon, E.C., Barroso, A. and Vandeventer, T., 2019. Control of tumor-associated macrophages and T cells in glioblastoma via AHR and CD39. *Nature neuroscience*, p.1.

[19] Gao, R., Davis, A., McDonald, T.O., Sei, E., Shi, X., Wang, Y., Tsai, P.C., Casasent, A., Waters, J., Zhang, H. and Meric-Bernstam, F., 2016. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature genetics*, *48*(10), p.1119.

[20] Olsen, T.K. and Baryawno, N., 2018. Introduction to Single-Cell RNA Sequencing. *Current protocols in molecular biology*, *122*(1), p.e57.

[21] Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A. and Kirschner, M.W., 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), pp.1187-1201.

[22] Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M. and Tirosh, I., 2015. AR 517 Bialas, N. Kamitaki, EM Martersteck, JJ Trombetta, DA Weitz, JR Sanes, AK 518 Shalek, A. Regev, and SA McCarroll. Highly parallel genome-wide expression profiling 519 of individual cells using nanoliter droplets. *Cell*, *161*(5), pp.1202-1214.

[23] Hebenstreit, D., 2012. Methods, challenges and potentials of single cell RNA-seq. *Biology*, *1*(3), pp.658-667.

[24] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A. and Lao, K., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, *6*(5), p.377.

[25] Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P. and Linnarsson, S., 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research*, *21*(7), pp.1160-1167.

[26] Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C. and Schroth, G.P., 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*, *30*(8), p.777.

[27] Hashimshony, T., Wagner, F., Sher, N. and Yanai, I., 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports*, *2*(3), pp.666-673.

[28] Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J.M., Blackburn, J., Barton, K., Roden, D., Luciani, F., Phan, T., Junankar, S. and Jackson, K., 2018. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *bioRxiv*, p.424945.

[29] Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T. and Ueda, H.R., 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology*, *14*(4), p.3097.

[30] Kouno, T., Moody, J., Kwon, A.T.J., Shibayama, Y., Kato, S., Huang, Y., Böttcher, M., Motakis, E., Mendez, M., Severin, J. and Luginbühl, J., 2019. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nature communications*, *10*(1), p.360.

[31] Dal Molin, A. and Di Camillo, B., 2018. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. *Briefings in bioinformatics*.

[32] Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D. and Woolf, P.J., 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, *10*(1), p.161.