UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Molecular Biology

TESI DI LAUREA

Implementation and use of cancer expression signatures from high resolution transcriptomics data

Relatore: Prof.ssa Enrica Calura
Dipartimento di Biologia

Correlatore: Dott.ssa Stefania Pirrotta
Dipartimento di Biologia

Laureanda: Martina Aere

ANNO ACCADEMICO 2023/2024

A nonno Emilio

Abstract

High-resolution transcriptomic sequencing techniques are crucial to gain an in-depth understanding of tumour biology, contributing to discovering new diagnostic and therapeutic insights in oncology. In particular, the single-cell transcriptomic technology allows for the analysis of individual cells separately, providing a detailed picture of cellular heterogeneity within the tumour. Additionally, it enables the identification of cellular subtypes, crucial in understanding tumours, as cellular diversity can impact disease progression and treatment response.
Here, I contributed to developing a new version of *signifinder*, an R package that streamlines the computation of signature scores from transcriptomic data, making it faster and more user-friendly. For the first time, I integrated different public cancer expression signatures from single-cell RNA sequencing data in the package. Thus, with this change, *signifinder* allows the users to compute scores of cancer expression signatures from high-resolution transcriptomic sequencing techniques in addition to the already-existing signatures from bulk data.
I also provide an example of how to use *signifinder* with a single-cell RNA sequencing dataset and how to interpret the signature scores.

Contents

Chapter 1

Introduction

1.1 The problem of tumour heterogeneity

Tumour heterogeneity represents one of the most significant challenges in the treatment efficacy of malignant tumours. The conversion from non-malignant to malignant cells occurs through various events that alter the nature of the cells. An accumulation of alterations that lead to cell proliferation, angiogenesis, evasion of cell death mechanisms, and uncontrolled growth can trigger this process. The carcinogenic nature can, thus, be described as stochastic, and even after the transformation of cells from healthy to malignant, a dynamic and continuous evolution of the tumour is observed, which does not follow a fixed mechanism: the genetic, transcriptomic, epigenetic, and phenotypic changes that occur are different for each patient. This incessant development leads to the generation of molecular heterogeneity of cancer, which consists of characterising various expression signatures in different individuals who are affected by the same tumour or within the cancer cells of the same individual. Different levels of sensitivity to therapies can thus manifest.

Two types of tumour heterogeneity can be distinguished: intertumoural heterogeneity and intratumoural heterogeneity. Intertumoural heterogeneity

is defined when there is a difference between patients with the same histological type of cancer, due to patient-specific factors. Intratumoural heterogeneity can be identified among the tumour cells of a single patient: spatial intratumoural heterogeneity describes the distribution of genetically diverse tumour subpopulations across different disease sites or within a single disease site. Another form of intratumoural heterogeneity can be temporal, meaning that the genetic variations in the tumour occur over time. That is one of the reasons why, despite good responses to therapy at the beginning, it is very common for cancer to develop resistance to treatment over time (Fig. 1.1).

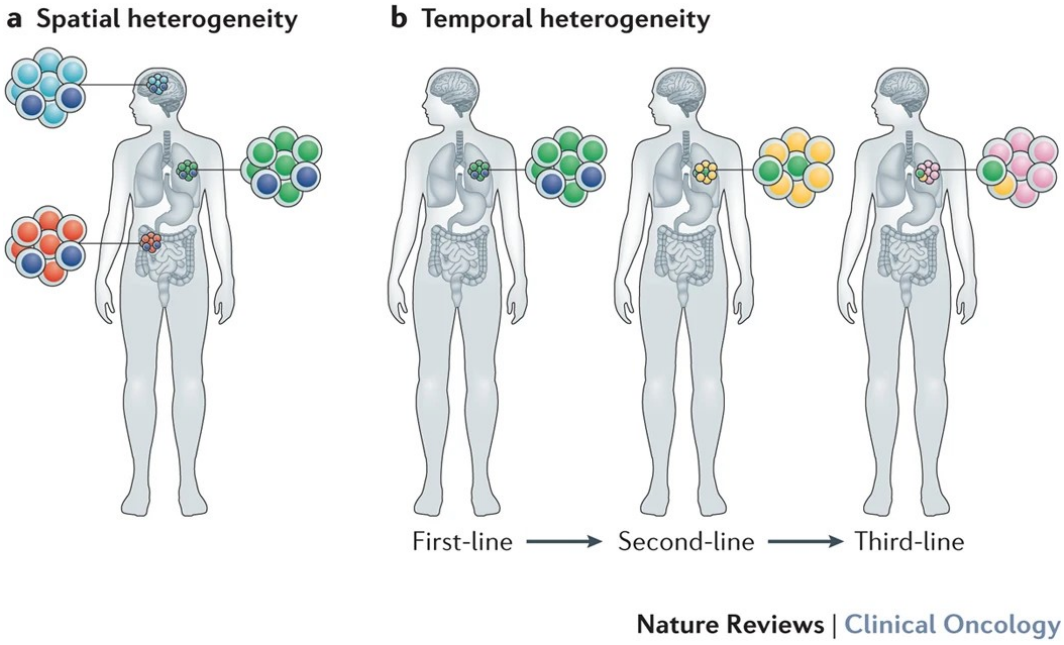

Figure 1.1: Spatial and temporal intratumoural heterogeneity (Dagogo et al., Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15, 81–94 (2018)).

The causes of intratumoral heterogeneity are multiple. To begin with, genomic instability is one of these causes: this can range from the alteration of a single base, such as a substitution, to the duplication of entire genomic portions. This problem can result from exogenous mutagenic factors such as UV radiation, risky behaviours like smoking, or even from anomalies in endogenous processes such as mistakes in DNA replication or repair.

Furthermore, it should be considered that exposure to treatments such as chemotherapy could partly increase the mutational spectrum of a tumour and create further genomic instability or decrease it by killing the cancer cells that respond better to therapy, thus promoting the selection of cells resistant to

treatment [1].

To address the complexity of tumour heterogeneity, researchers have developed approaches such as gene expression signatures.

1.2 What is a cancer gene expression signature?

Gene expression signatures represent a crucial tool in understanding and managing cancer. When faced with a disease as complex and heterogeneous as cancer, it is essential to be able to answer fundamental questions about the nature of the tumour, whether it is possible to predict how the tumour may progress, and whether the patient will respond positively to treatment. Gene expression signatures offer a powerful way to address these questions. [2]

The cancer gene expression signatures are developed by studies in which specific cancer processes or states are analysed. Some examples of these processes are the epithelial-to-mesenchymal transition or the interferon response; in some studies, instead, the researchers classify the tumour into different cellular states such as astrocyte-like, oligodendrocyte progenitor cell-like or neural progenitor cell-like in glioblastoma. The result of these analyses is the identification of a specific set of genes, which are responsible for the tumour process or state of interest. The combination of those genes' expression, the score, is unique and it is a sign of the presence of that biological function or cellular type behaviour. The collection of genes is called "cancer gene expression signature". In addition, they offer the opportunity to interrogate other datasets of different tumour samples to assess the amount to which a gene pattern identifying the signature is present.

Sometimes these signatures also have a prognostic influence and they can be used for instance to predict the patient's response to a drug or therapy, avoiding ineffective treatments. Moreover, gene signatures can be exploited to identify new therapeutic targets [2].

New cancer gene expression signatures are detected and generated using RNA sequencing technologies such as bulk RNA sequencing, single-cell RNA sequencing, and spatial transcriptomics [3].

This study will focus on public signatures generated from single-cell RNA sequencing data: this technology has paved the way for the discovery of previously unknown cell types and subtypes, to better understand intratumour heterogeneity, or to better characterise cellular subpopulations in particular biological responses. This is why it is crucial to develop new versions of tools

which can be able to exploit public signatures based on high-resolution technologies.

1.3 Single-cell RNA sequencing technology

Starting from the past two decades, RNA sequencing (RNA-seq) technology has begun catching on and become increasingly prevalent in molecular biology. RNA-seq is especially exploited for the analysis of differential gene expression (DGE), which allows users to measure quantitative changes in expression levels between experimental groups. In recent years, more advanced RNA sequencing techniques have been developed, allowing for higher resolution in expression-level analysis. In particular, single-cell RNA sequencing (scRNA-seq) has emerged as the state-of-the-art approach to address the problem of heterogeneity and complexity of RNA transcripts within individual cells, thereby revealing the composition of cell types and functions within tissues and tumour organs [4].

In general, there are two important categories of single-cell RNA sequencing methods: the full-length scRNA-seq, so cells are physically separated and then sequenced; and the tag-based scRNA-seq, so cells are tagged with barcodes and then their data are separated computationally(Fig. 1.2) [5].

A platform that uses the full-length scRNA-seq technology is the SMART-seq2. The process begins with the isolation of single cells using techniques like FACS (fluorescence-activated cell sorting) or microfluidic systems. After that, the cells are lysed, and their RNA is reverse transcribed and converted into cDNA.

A mechanism called "switching mechanism at the 5' end of RNA template" is employed to add a specific adapter sequence to the 5' end of the cDNA, improving the efficiency of capturing full-length transcripts.

The cDNA is then amplified by PCR without the need for tags, as in SMART-seq2 each cell is treated individually. At this point, the cDNA is prepared for sequencing by adding specific adapters to the ends of the fragments [6].

After the sequencing process and its validation with specific tools, an output is obtained. This is a matrix that represents the number of counts for each gene in each cell.

The second technology is the tag-based scRNA-seq, like 10x Genomics droplet-based single cell sequencing.

The main feature of this technique is that the single cells are placed into little droplets together with some beads that have known barcodes attached.

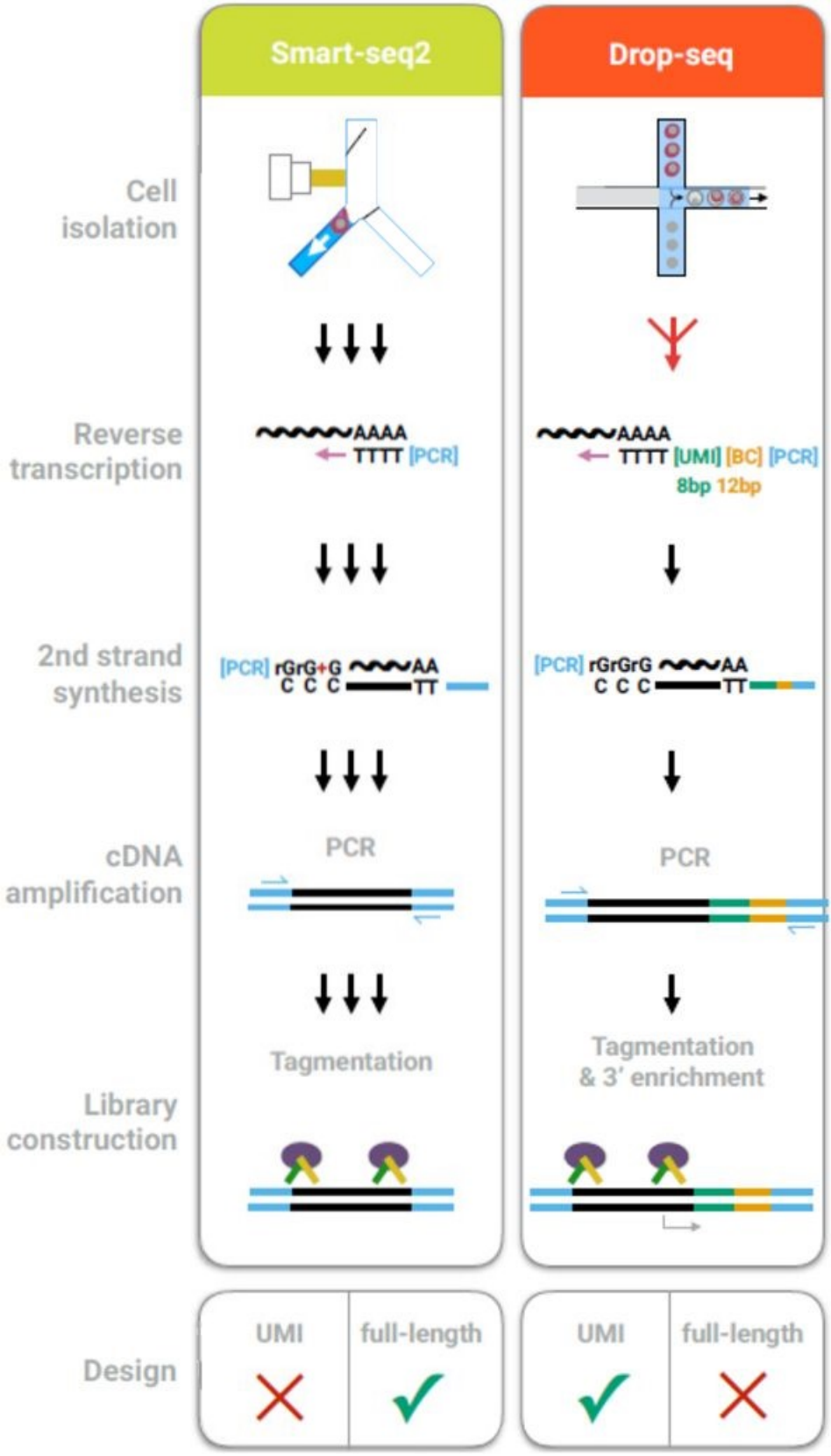

Figure 1.2: SMARTseq2 and Drop-seq differences in library preparation steps (Ziegenhain et al. "Comparative Analysis of Single-Cell RNA Sequencing Methods". Molecular Cell 65.4 (Feb. 2017), 631– 643.e4).

So, the beads are inserted through some microfluidics with the sample cells, and they are partitioned together into the droplet. Since they are suspended in oil, they do not recombine. So, they can be processed individually in downstream reactions. At the end of the process, a library is created in which reads have some identifiers attached to them, the most important ones are the already mentioned 10x barcodes, which are strings of nucleotides that can identify from which cell the sequence comes from; and the UMIs (unique molecular identifiers): they are another type of barcode that is not cell specific, but it gets amplified in PCR step and allows to differentiate between PCR duplicates and actual gene copies [7].

After the read processing with quantitative and alignment tools, a count matrix as output is obtained.

The count matrix is a table where rows represent genes and columns represent individual cells, with each cell containing a value indicating the number of transcripts (or reads) detected for a specific gene in that cell. This matrix is crucial for downstream analysis as it provides the quantitative data needed to understand gene expression patterns at the single-cell level.

After the creation of the output, the matrix exhibits several issues, including data quality: some cells may have an abnormal number of detected genes or could be of low quality due to technical artefacts during sample preparation or sequencing; and the presence of a sparse matrix: gene expression matrices in scRNA-seq often contain many zeros due to dropout events, where some transcripts are not detected in all cells. To address these problems the count matrix has to be filtered to remove low-quality genes and cells. This involves discarding genes that are expressed in very few cells and cells that have an abnormal number of detected genes. This helps to ensure that the analysis focuses on the most informative data, essential for managing the challenges posed by scRNA-seq technology: its large datasets require significant computational resources, and interpreting them is complex, especially in tumour samples with high intra-sample cellular and transcriptomic heterogeneity [8].

In this scenario gene expression signatures are powerful tools to solve the data complexity and simplify interpretations.

1.4 Gene expression signatures and the *signifinder* package

Gene expression signatures have the potential to reveal ongoing cancer activities and guide therapeutic decisions, but they face several significant challenges. Despite numerous gene expression-based prognostic signatures being

reported in the literature, very few have been adopted in clinical practice. The variability between patients and the complex relationships between tumours and their microenvironment further complicate their application.

Calculating signature scores is a crucial task for identifying signatures in new datasets or performing analyses. However, each signature has a specific method of score calculation recommended by its developers, which cannot be universally applied to other signatures. Additionally, there is a lack of tools that standardise these computations, highlighting the need for a package that can compile public gene expression signatures and automate their score calculation [3].

signifinder is an open-source R package, part of the gene expression data structures within the Bioconductor project.

Its purpose is to make public cancer gene expression signatures more reproducible and user-friendly by automating the calculation of their scores, while preserving the computation method provided by the signature developers. It also offers the possibility to interpret and compare the scores through visualisation methods.

For this reason, *signifinder* curates a comprehensive collection of signatures from the literature, adhering to stringent selection criteria. These criteria include the utilisation of cancer samples, availability of transcriptional data, clarity and coherence of gene lists and the type of input expression data required, and an unambiguous description of score calculation methods.

An R function for each signature, built according to the methodology elucidated in the corresponding research articles, allows the computation of the scores. Users can easily input, as an argument of the signature function, normalised expression data derived from bulk, single-cell, or spatial transcriptomics cancer samples. The input can be provided in the format of SummarizedExperiment, SingleCellExperiment, or SpatialExperiment. These are data structures usable in R that act as containers for biological data. The SummarizedExperiment is used for bulk RNA sequencing data and includes gene expression matrices and related annotations [9]. The SingleCellExperiment is used for scRNA-seq data and, in addition to the expression matrices, contains specific metadata related to biological or technical characteristics of individual cells [10]. The SpatialExperiment is used to represent spatial data obtained from molecular or spatial imaging techniques to study, for example, the distribution of gene expression spots in tumour tissues [11].

The package allows the calculation regardless of the abundance metrics

required to compute the score, because *signifinder* has internal functions which automatically transform the normalised expression counts of the input into the necessary metrics, contributing to standardise the scoring process [3].

signifinder is also equipped with all necessary data components for each signature and facilitates automated data transformations whenever necessary to be compliant with the signature's original requirements. The functions return signature scores into a section of a SummarizedExperiment, a SingleCellExperiment or a SpatialExperiment.

In addition, *signifinder* offers diverse visualisation tools for inspecting scores, enabling users to gain comprehensive insights into the underlying data. Furthermore, the package incorporates features to identify the top contributor genes, furnishing users with valuable insights into the molecular drivers underpinning observed expression patterns.(Fig. 1.3) [3].

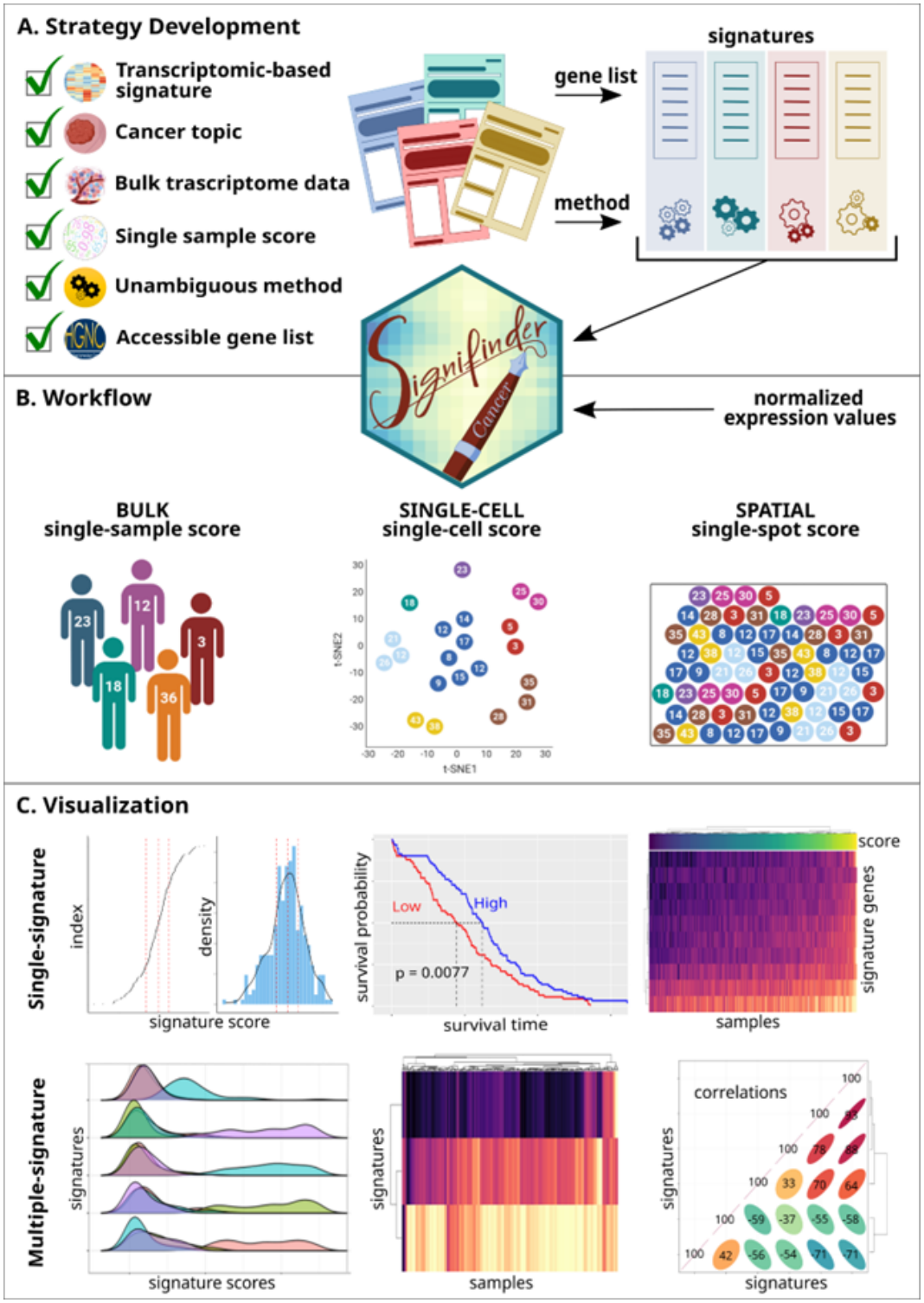

Figure 1.3: *signifinder* implementation and workflow.(Pirrotta et al., signifinder enables the identification of tumor cell states and cancer expression signatures in bulk, single-cell and spatial transcriptomic data. Mar. 2023)

Chapter 2

The aim of the thesis

The goal of this thesis project is to enrich the *signifinder* package with new cancer gene expression signatures generated from single-cell RNAseq, thus improving the power of resolution on the biological complexity of the tumour.

The project is divided into three tasks. The first task involved researching new cancer gene expression signatures from high-resolution data in the literature. To do this, I used the PubMed database and searched for terms like "cancer", "gene expression signature", "single-cell RNA sequencing", "modules" or "hallmark" in the Advanced Search. In addition, I filtered the results by date to find recent articles on this topic.

The second task was building a function in the programming language R for each signature. This function computes the signature's score following the method specified in the article. I then add this function to the *signifinder* package using Git and GitLab.

The third task was presenting a case study that demonstrates the applicability of the new single cell *signifinder* functions.

Chapter 3

Materials and Methods

3.1 Collection of public gene expression signatures

As already stated in the thesis objective, the first step of the project is to create a collection of public gene expression signatures from single cell RNA sequencing data.

The signatures must be selected based on the following criteria:

1. The signatures must be related to a cancer topic;
2. They must be gene expression signatures, therefore related to transcriptional data;
3. The type of expression in input must be clear, in my case, the data they originate from must be single-cell RNA sequencing;
4. The list of genes characterising the signature must be clear and easily accessible. Hence, the list must be provided through a table in an Excel file or a PDF file.
   Additionally, the genes should mostly be in official nomenclature or should have nomenclature that can be translated into the official one. For example, the gene might have an ENSEMBL ID but easily translatable into SYMBOL.
5. Finally, the method for calculating the signature score must be described clearly and in detail, without ambiguity, within the paper. Even better if the scoring method code is also made available [3].

To this purpose, PubMed was used: a database of biomedical literature from MEDLINE, life science journals, and online books, curated by the NCBI (National Center for Biotechnology Information). The search was conducted

using various keywords such as “cancer”, “gene expression signature”, “single-cell RNA sequencing”, “modules”, and “hallmark”, and the advanced search mode was also employed.

The advanced search on PubMed allows the use of keywords as queries and linking them with logical connectors such as AND or OR. It is also possible to apply filters such as publication date to select only the most recent papers or the type of paper to consult.

After selecting the papers of interest and ensuring they met the parameters indicated above, it is necessary to summarise the characteristics of the signature and its score. To do this, I used a table (Table 3.1) with the following entries:

- *signature*: The signature is named using a two-part name. The first is a term that can quickly identify what the signature characterises. If present, the same name provided by the author in the paper is used; otherwise, an appropriate name is assigned. The second part is the name of the first author of the article. This way, it is possible to recognise when two signatures with the same characteristics from different papers, with two different scoring computation methods, are included in the package. For example, the name EMT_Barkley highlights a signature that characterises an epithelial-to-mesenchymal transition in the cells of the dataset where it is expressed (EMT, Epithelial-to-Mesenchymal Transition), and “Barkley” is the name of the author of the paper from which it was taken.

- *scoreLabel*: the name given to one or more scores calculated for a specific signature.

- *functionName*: the name used to identify the function that calculates the signature score in the *signifinder* package.

- *topic*: the field to which the signature belongs, such as “immune system” or “epithelial-to-mesenchymal transition”.

- *tumour*: the tumour in which the signature score can be calculated.

- *tissue*: the specific tissue of a certain signature.

- *cellType*: the cell type in which the signature can be calculated.

- *developedWith*: the data from which the signatures were developed; in my case, they are all signatures derived from scRNA-seq data.

- *usableInput*: the type of data that can be used with the signature.

- *transformationStep*: often, to calculate the score, it is necessary to first transform the normalised expression values of the dataset of interest. This information is contained in the score calculation method described in the article.
- *author*: it indicates the first author of the article from which the signature is taken.
- *reference*: the article from which the signature is taken.
- *description*: a description of what the signature score is and how it can be interpreted.

These characteristics are then reported in the *signifinder* package in the "signatureTable" function, so that they can be accessible to users who can consult it to find the signatures to calculate that best fit their dataset.

signature	scoreLabel	functionName	topic	tumour	tissue	cellType	developedWith	usableInput	transformationStep	author	reference
EMT_Barkley	EMT_Barkley_cEMT, EMT_Barkley_pEMT	EMTSign	epithelial to mesenchymal	pan-cancer	pan-tissue	malignant	sc	sequencing	normCounts	Barkley	[12]
Hypoxia_Barkley	Hypoxia_Barkley	hypoxiaSign	hypoxia	pan-cancer	pan-tissue	malignant	sc	sequencing	normCounts	Barkley	[12]
CellCycle_Barkley	CellCycle_Barkley	cellCycleSign	cell cycle	pan-cancer	pan-tissue	malignant	sc	sequencing	normCounts	Barkley	[12]
State_Neftel	State_Neftel_MES1, State_Neftel_MES2, State_Neftel_AC, State_Neftel_OPC, State_Neftel_NPC1, State_Neftel_NPC2	stateSign	glioblastoma cellular states	glioblastoma	brain	malignant	sc	sequencing	$log_2(TPM/10+1)$	Neftel	[13]
State_Barkley	State_Barkley_Alveolar, State_Barkley_Basal, State_Barkley_Squamous, State_Barkley_Glandular, State_Barkley_Ciliated, State_Barkley_AC, State_Barkley_OPC, State_Barkley_NPC	stateSign	pan-cancer cellular states	pan-cancer	pan-tissue	malignant	sc	sequencing	normCounts	Barkley	[12]
State_Tirosh	State_Tirosh_MITF, State_Tirosh_AXL	stateSign	metastatic melanoma cellular states	metastatic melanoma	spleen, subcutaneous, intramuscolar, skin	malignant	sc	sequencing	$log_2(TPM/10+1)$	Tirosh	[14]
SCSubtype_Wu	SCSubtype_Wu_Basal, SCSubtype_Wu_Her2E, SCSubtype_Wu_LumA, SCSubtype_Wu_LumB	SCSubtypeSign	breast cancer cellular states	breast cancer	breast	malignant	sc	sequencing	$log_2(normCounts+1)$	Wu	[15]
Stress_Barkley	Stress_Barkley	stressSign	stress response	pan-cancer	pan-tissue	malignant	sc	sequencing	normCounts	Barkley	[12]
Interferon_Barkley	Interferon_Barkley	interferonSign	interferon response	pan-cancer	pan-tissue	malignant	sc	sequencing	normCounts	Barkley	[12]
Oxphos_Barkley	Oxphos_Barkley	oxphosSign	oxidative phosphorylation	pan-cancer	pan-tissue	malignant	sc	sequencing	normCounts	Barkley	[12]
Metal_Barkley	Metal_Barkley	metalSign	metal response	pan-cancer	pan-tissue	malignant	sc	sequencing	normCounts	Barkley	[12]

Table 3.1: Signatures based of scRNA-seq table, the descriptions are present in Table A.1 in Appendix.

3.2 The R language

R is a programming language and environment that is open-source and designed for statistical analysis and graph visualisation. It has numerous advantages, which is why it was chosen to create the *signifinder* package and conduct analyses in the presented case study. R is a versatile programming language that offers a huge number of statistical techniques and visualisation tools, making it ideal for filtering and optimising input data for the package and integrating analyses performed using *signifinder*. It also provides a lot of packages, particularly for biological studies built within the Bioconductor project. Furthermore, R is user-friendly and the RStudio interface makes the analysis easier.

R is also considered as a powerful tool for constructing tables and graphs suitable for visualising data, as it is shown in *signifinder*. Lastly, it benefits from a large and active community of users and developers who contribute to its development, ensuring up-to-date documentation and various online tutorials. Therefore, R is the perfect programming language for building the functions of the signature scores and to conduct the analysis of the case-study [16].

3.3 Git and GitLab

A R package is an extension of the statistical programming language R, containing a collection of standardised codes, data, and documentation that can be installed by R users.
However, developing an R package is not easy, especially when multiple developers are working simultaneously to modify or update the data and information within it.
It is also fundamental to consider that making changes can often lead to errors, such as in a code, which could compromise the functioning of the package if done directly on the main body without first testing it locally.
Therefore, the use of a development environment becomes necessary. This environment allows cloning a copy of the package locally, designing, implementing, and testing the changes while keeping track of them. It also provides the ability to revert in case the changes lead to errors, or it allows multiple developers to work simultaneously on the same file.

Git is an open-source distributed version control system (VCS), that tracks changes during the R package development process. It builds a self-contained repository around a set of folders and files, storing metadata about the files and changes made to them in a hidden folder within the root folder. In particular, Git allows cloning a repository from a remote location to one's local machine, enabling the creation of branches, which are copies of the master repository,

where the developers can make changes without risking modification of the principal body. Git also allows updating the main repository stored locally with a specific command: *pull*, saving a new version of the just made changes with a name and description of the modifications; this operation is called a *commit*. Finally, all commits made locally are uploaded to the remote location with the *push* command [17].

The coordination of changes made by developers occurs through an official repository stored on GitLab, an open-source DevOps platform based on Git. Thus, it is a platform that utilises a software development methodology focused on communication, collaboration, and integration among developers.

In other words, GitLab serves as a front-end for Git, providing a robust interface for Git repositories stored on its server. Users create an account on GitLab and associate an SSH (Secure Shell) key with their profile, which is an identifying key allowing secure communication between two machines working remotely connected to a public network. This enables working locally and pushing changes remotely in an encrypted manner.

Users can then create or be invited to participate in Git repositories stored on the GitLab server. GitLab offers additional features, including an issue tracker, merge request tools, and an interface for reviewing changes before their integration into the main branch of the repository [17].

3.4 The procedure to add a new signature in *signifinder* by using R, Git and GitLab

Git and GitLab were essential in integrating the functions for calculating signature scores into the *signifinder* package. Firstly, it is necessary to activate the SSH key in the Git Bash command shell using the commands $ssh - agent\ bash$ and $ssh - add\ key\_name$. Then, clone the package repository locally by entering the command *git clone* [*url*]. The [*url*] represents the URL link of the repository in GitLab.

To keep the main repository unchanged, a branch of the repository is usually created with the command *git branch branch_name*. This procedure allows modifications to the branch, and after ensuring everything works, it can be merged into the main repository using the command *git merge branch_name* (Fig. 3.1) [17].

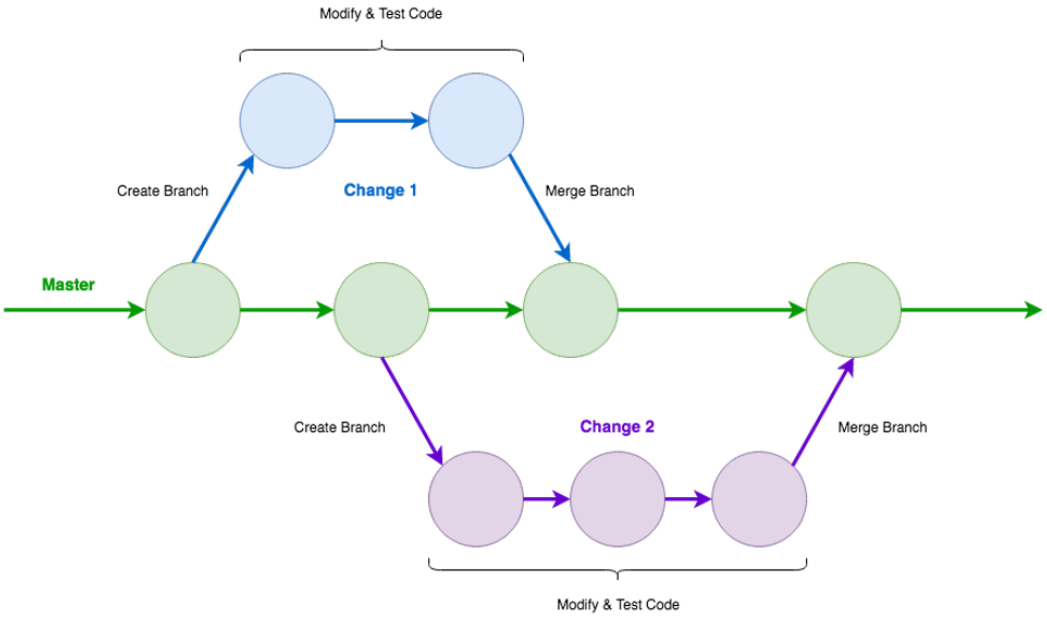

Figure 3.1: Development branches in Git. (Engwall and Roe, Git and GitLab in Library Website Change Management Workflows, The Code4Lib Journal 48 (May 2020))

Additionally, the SSH key can be associated with RStudio to facilitate content modifications on the branch.

The procedure to add a new cancer gene expression signature in *signifinder* generally includes the modification of multiple files in the original package repository as described below. The first file to modify is the "SignatureFunction.R" file, which contains the collection of all the signature functions present in *signifinder*. Here, the function for the score of the new signature is appended to the existing ones, or an existing signature is adapted to include the newest signature. Additionally, a description of the new signature function for the documentation manual is always provided at the beginning of each function, written using *roxygen2*, an R package that provides an easy and standard way to document functions in the package [18].

Another file that has to be updated implementing a new signature is the "UtilityFunction.R" file, which contains other functions essential for the complete working of the package. Here, the score label of the new signature function is saved in the "SignatureNames" vector. Also, the ".GetGenes" function, retrieving the genes characterising the signatures necessary for score calculation, is modified: the code lines for retrieving the genes of the new signature are inserted here.

The ".GetGenes" function operates thanks to a project called "signaturedata_signifinder". This project contains an internal laboratory procedure to prepare *signifinder*'s signature data. In particular, signature genes are retrieved from online resources found in the literature through a code in the "signatureData.rmd" file, which subsequently inserts them into a data frame. Therefore, a script to retrieve and save the genes of the new signature in a data frame is attached to the code. At the beginning, the script should provide a description with the name of the new signature, the function label for the score, and the tissue where the signature was detected. Here, the link to the literature article and the online resource from where the genes were downloaded are inserted. Then, the script consists of several steps. Firstly, the signature genes are downloaded from a URL. Secondly, the downloaded data is read depending on the type of file of the genes. At this point, a data frame is created with two columns: "SYMBOLS", containing the gene symbols, and "class", indicating the tumour states or subtypes to which each signature belongs.

Afterwards, the gene names' presence is checked in "AnnotationDbi", ensuring that other aliases are not employed and that the gene symbol is unambiguous, meaning it does not have more than one "entrez". This procedure is run using an internal project function called "checkSignGenes". Subsequently, the generally few genes that do not pass the automatic translation are manually checked, in case the correct symbol is inserted, and genes with an ambiguous symbol are deleted. Then, the gene symbols are updated in the data frame. Finally, the data frame is saved in a .rda file named "sysdata", which is imported from "signaturedata_signifinder" into the *signifinder* package.

After all these steps, tests are created in the "test-SignatureFunctions.R" file in *signifinder*. These tests ensure that the computation of the score function of the new signature works correctly returning a SingleCellExperiment, for instance, with the names of the states or subtypes present in "colData", or that the number of score values is equal to the number of cells in the analysed dataset. Finally, the command $roxygen2 :: roxygenise()$ is executed to run the function of the *roxygen2* package and update the documentation manual including the new parts [18].

Chapter 4

Results and Discussion

4.1 The public cancer gene expression signatures

I collected some public gene expression signatures of 11 different topics. These signatures can be divided in 2 macro-categories: the signatures targeted to infer specific cellular processes and signatures that are more related to infer cell subtype identity. All the signatures that I implemented and that are presented in the following paragraphs derive from single-cell RNA sequencing data and their scores are supposed to be calculated only in malignant cells (Table 4.1) [12, 13, 14, 15].

signature name	**tumor**	**tissue**	**reference**	**macro-category**
Epithelial-to-mesenchymal transition signature	pan-cancer	pan-tissue	[12]	Cellular process
Hypoxia signature	pan-cancer	pan-tissue	[12]	Cellular process
Cell cycle signature	pan-cancer	pan-tissue	[12]	Cellular process
Stress response signature	pan-cancer	pan-tissue	[12]	Cellular process
Interferon response signature	pan-cancer	pan-tissue	[12]	Cellular process
Oxidative phosphorylation signature	pan-cancer	pan-tissue	[12]	Cellular process
Metal response signature	pan-cancer	pan-tissue	[12]	Cellular process
Pan-cancer cellular states signatures	pan-cancer	pan-tissue	[12]	Cell identity
Glioblastoma cellular states signatures	glioblastoma	brain	[13]	Cell identity
Metastatic melanoma cellular states signatures	metastatic melanoma	spleen, subcutaneous, intramuscolar, skin	[14]	Cell identity
Breast cancer cellular subtypes signatures	breast cancer	breast	[15]	Cell identity

Table 4.1: The collection of scRNA-seq-derived signatures with their features.

4.1.1 Epithelial-to-Mesenchymal transition pan-cancer signature

The first signature is constructed to infer the epithelial-to-mesenchymal transition. This process is characterised by the loss of epithelial cells' features like their shape or their tight arrangement given by cell-cell junctions; and the acquisition of mesenchymal features like a more elongated shape and more motility. Generally, in cancer cells, the epithelial-to-mesenchymal transition (EMT), and therefore the mesenchymal phenotype, is connected with tumour progression, metastasis formation. The EMT_Barkley signature scores can be computed across all cancer types and all tissues, so I refer to this signature as a "pan-cancer" one. It is composed of two modules: pEMT and cEMT.

The cEMT module represents a group of genes which can detect the cells in which a complete epithelial-to-mesenchymal transition happens.

The pEMT module, or rather "partial epithelial to mesenchymal transition", represents a group of expressed genes that can detect a not finished EMT. In effect, this signature does not have the canonical mesenchymal markers like collagen genes.

When expressed, both EMT_Barkley_cEMT and EMT_Barkley_pEMT indicate two pathways exhibiting mesenchymal differentiation behaviours in tumour cells, influencing phenotypic properties such as migration or drug resistance [12].

4.1.2 Hypoxia signature

Hypoxia is the condition in which the cells activate some survival mechanisms due to an environment poor of oxygen. In tumours, this condition is quite common: when there is rapid cancer growth, the level of oxygen required for development exceeds the oxygen supplied by the blood vessels. This leads to a state of hypoxia, which causes the activation of pathways for adaptation and survival under these adverse conditions.

The Hypoxia_Barkley signature has a set of genes whose expression in a tumour sample detects a cellular response to hypoxia. Its score can be calculated in all cancers and tissues [12].

4.1.3 Oxidative phosphorylation signature

There is a signature related to another metabolic process: the oxidative phosphorylation (OXPHOS). This process plays a crucial role in the energy metabolism of tumour cells and their ability to adapt and survive in adverse environments,

thereby influencing tumour progression and response to treatments. OXPHOS is responsible for the production of the majority of ATP. A high energy supply in tumour cells could therefore indicate that the cancer is using a reprogrammed metabolism that supports its proliferation and survival [19].

The Oxphos_Barkley signature score can detect an oxidative phosphorylation response in all cancer samples [12].

4.1.4 Cell cycle signature

The presence of actively cycling cells within a tumour is a significant factor in tumour growth and progression. The main influences of actively cycling cells on tumour dynamics are rapid proliferation due to uncontrolled cell division; genetic instability caused by DNA replication errors that can occur during cell division; and heterogeneity, as different cells are in different phases of the cell cycle, leading to variations in gene expression, metabolism, and susceptibility to treatments [20].

The CellCycle_Barkley signature score can capture the subset of cancer cells in any tumour that is cycling [12].

4.1.5 Stress response signature

In the presence of tumours, cells can undergo various types of stress due to an environment characterised by nutrient deficiency, low oxygen levels, or exposure to drugs and treatments. This induces cells to activate pathways to oppose this hostile condition.

The Stress_Barkley signature score can be computed in all cancers and it identifies the expression of a group of genes involved in a response to a stress condition [12].

4.1.6 Interferon response signature

In tumours, interferons play a significant role by activating a series of mechanisms that influence the tumour microenvironment and associated immune responses. For instance, IFNα and IFNβ activate signalling pathways that induce the expression of genes affecting cancer cell growth, proliferation, and survival. They can exert anti-proliferative effects by promoting mechanisms such as apoptosis or inhibiting angiogenesis. However, immune response in tumours is complex, and chronic exposure to interferons may trigger mechanisms of tumour resistance [21].

The Interferon_Barkley signature contains both canonical interferon-stimulated genes and components of antigen presentation and its score can be calculated for each pan-cancer study [12].

4.1.7 Metal response signature

The Metal_Barkley signature is composed of a set of metallothionein genes and its score can be computed for all cancers. It may have a role in proliferation and drug resistance [12]. In effect, the metal response is linked to the expression of metallothioneins, which are crucial proteins in the regulation of metal homeostasis in the body. Specifically, they regulate zinc and copper, which are important for cellular proliferation and differentiation. They are cofactors for enzymes involved in DNA replication and protein synthesis. Metallothioneins also act as free radical scavengers, further contributing to the survival of cancer cells. Their role in protecting cells from oxidative stress also makes them a cause of tumour resistance to certain drugs [22].

Below I am going to present the signatures associated with cell identity and cell subtypes.

4.1.8 Pan-cancer cellular states signature

The State_Barkley signature is composed of 8 modules which identify different cellular types in cancer.

The first 4 modules are related to epithelial cell type markers: alveolar, squamous, basal and glandular cell modules. As the name suggests, they mirror the corresponding cellular type. Then, another module related to cellular identity is proposed, that is the ciliated module, which is made of cilium-related genes.

At the end, 3 neurological cancer-specific modules have been identified: the astrocyte (AC)-like module; oligodendrocyte progenitor cell (OPC)-like module and the neural progenitor cell (NPC)-like module. They are all linked to cell types related to the nervous system [12].

All the above signatures are developed by Barkley et al in 2022, their scores are included between 0 and 1 values. A signature is expressed if its score in the cell is higher than 0.5 [12].

4.1.9 Glioblastoma cellular states signature

This signature was developed by Neftel et al. in 2019, and it is characterised by six main cellular states of glioblastoma. This is very important because the

main therapeutic failures are due to glioblastoma heterogeneity, so this new signature came very useful.

The signature explains four main cellular states that mirror four neural cell types, anyway two of them can be divided into two subgroups: mesenchymal 1-like, mesenchymal 2-like, astrocyte-like, oligodendrocyte progenitor cell-like, neural progenitor cell 1-like and neural progenitor cell 2-like states.

The first state which can diverge into two modules is associated with high expression of mesenchymal related genes: MES1 (mesenchymal 1-like) module and MES2 (mesenchymal 2-like) module.

The MES2 module is interesting because it is related to hypoxia-response, stress and glycolytic genes; suggesting that in some tumours the mesenchymal state is linked to hypoxia and the increasing glycolysis. This is why this module is defined as an hipoxia-dependent signature.

On the other hand, the MES1 is just related to mesenchymal genes, so it is called hypoxia-independent signature.

The other four modules are connected to neurodevelopmental genes; in particular, the astrocyte-like module (AC), includes astrocytic markers; the oligodendrocyte progenitor cells-like (OPC) has markers related to the oligodendroglial lineage.

The neural progenitor cells (NPC)-like is the other state which can be subdivided into two modules: the NPC1 and the NPC2.

The NPC1 includes OPC-related genes and reflects the potential of NPCs to differentiate toward OPCs. The NPC2, instead, mirrors the potential of NPCs to differentiate toward neurons as it includes neuronal lineage genes.

The State_Neftel signature returns six cellular states scores, the higher is the score computed for each module in each cell, the higher that state is present in that cell [13].

4.1.10 Metastatic melanoma cellular states signature

The State_Tirosh signature was developed by Tirosh et al. in 2016, it represents two transcriptional cellular states present in metastatic melanoma: MITF-high and AXL-high.

The first module is called MITF, because it is composed of the MITF gene (microphthalmia-associated transcription factor gene) and other MITF target genes. The MITF gene is the most important melanocyte transcriptional regulator and melanoma lineage survival oncogene.

The second module is called AXL, it is negatively correlated with the MITF module. This module is composed by a set of genes in which there is also AXL, that is linked to the resistance to various targeted therapies and the NGFR marker, that is a putative melanoma cancer stem cell marker. In the Tirosh

et al.(2016) study, the expression of this module was found also in treatment-naive patients, indicating the presence of dormant drug-resistant population in some cell lines.

The signature cellular states scores can be computed, the higher is the score in a cell, the higher is the presence of that module [14].

4.1.11 Breast cancer cellular subtypes signature

This signature was created by Wu et al. in 2022. The study investigates heterogeneity at the subtype level in breast cancer. Specifically, the aim was to find a method compatible with the PAM50 signature, which classifies breast cancer into different molecular subtypes based on bulk transcriptomics profiling, starting from scRNA-seq data. Thus, the SCSubtype signature was constructed. It classifies cancer cells into four different breast cancer subtypes: Basal, Her2E, LumA, and LumB [15].

LumA (luminal A) and LumB (luminal B) are two subtypes that both express estrogen and progesterone receptors but have different levels of cell proliferation. LumA is characterised by a low level of the cell proliferation marker Ki-67 and the absence of human epidermal growth factor receptor 2 (HER2). In contrast, LumB may not express progesterone receptors, it has a higher level of proliferation than LumA and may express HER2.

Her2E is characterised by the expression of HER2 and the general absence of estrogen and progesterone receptors.

The basal-like subtype is so named because it expresses genes characteristic of the basal myoepithelial cells of the breast. Basal-like tumours are negative for estrogen receptors (ER), progesterone receptors (PR) and HER2, and they exhibit a high level of expression of genes associated with cell proliferation. A cell is assigned to a subtype if it has the highest score for that subtype [23].

4.2 Computation of the scores

Each signature can be computed thanks to a precise method indicated by the authors in their paper.

4.2.1 The scoring method of the cellular processes signatures and the pan-cancer cellular states signatures

All the signatures related to the cellular processes and the pan-cancer cellular states signature have the same scoring method [12].

To calculate the score, it is necessary to have a SingleCellExperiment of a tumour sample that includes a normalised matrix of expression values for each gene in each cell.

For each signature, a thousand random lists of genes with similar expression levels have to be generated [12].

To do this, I initially calculated the mean values for each gene in the dataset, removing genes with null values. I then ordered them in ascending order.

```
data.avg <- sort(rowMeans(x = dataset,
                          na.rm = TRUE))
```

To each mean value, I added random variables that follow a normal distribution as indicated in the article. This allowed me to obtain a scale of genes ordered by their mean value. After that, I divided the gene list into 25 bins, that is, 25 groups with a roughly equal number of genes based on the mean expression value of each gene.

```
data.cut <- cut_number(
    x = data.avg + rnorm(n = length(data.avg))/1e
       +30,
    n = 25, labels = FALSE, right = FALSE)
names(x = data.cut) <- names(x = data.avg)
binned <- split(names(data.cut), data.cut)
```

Next, through a for loop, I took each gene from each module of the signature. If the gene existed in the studied dataset, I checked which bin it was in, randomly selected a gene from the same bin in the dataset, and inserted the gene into an empty list called "new". If the gene was not present in the list, I moved on to another gene in the module. After inserting the random gene into the empty list, I recreated the same bin without the selected gene to avoid using the same gene for that list. At this point, I repeated the same process a thousand times to create a thousand random gene lists for each module.

```
rand <- lapply(names(mod), function(m){
    lapply(seq_len(1000), function(i){
        used <- vector()
        unused <- binned
        for (g in mod[[m]]){
            pool <- data.cut[g]
            if (!(is.na(pool))) {
                new <- sample(unused[[pool]], 1)
                used <- c(used, new)
                unused[[pool]] <- setdiff(unused
                   [[pool]], new)}}
```

```
            used}) })
    names(rand) <- names(mod)
```

At this point, I took each random gene list for each module and calculated the mean expression value for each cell using only the genes from each list.

```
    ra <- sapply(rand[[m]], function(i){
            colMeans(dataset[i, ], na.rm = TRUE) })
```

Next, I followed the same procedure for the gene list characterising the module.

```
    re <- colMeans(
        dataset[rownames(dataset) %in% mod[[m]],],
        na.rm = TRUE)
```

At this point, I took only the random geneset values greater than the module geneset value, called p, and calculated $-log_{10}(p)$ [12].

```
    p <- -log10(rowMeans(ra >= re))
```

I applied the score only to malignant cells, assigning a null value to the others.

```
    s <- rep(NA, n)
        s[isMalignant] <- p
```

If the logarithm yielded an infinite value, it was assigned a value of 1.

Finally, I used the "scale" function to rescale the score to a value between 0 and 1 [12].

```
    scores[is.infinite(scores)] <- 4
    scores <- scores/4
    scores <- rescale(scores)
```

In conclusion, I have entered the code within the 'UtilityFunction' file in the *signifinder* package. For each Barkley signature, I created a function in 'SignatureFunction' of *signifinder*, that calls the function just described and calculates the score on the specific modules of each function. Finally, the computed scores are saved within the SingleCellExperiment object of the dataset that the user is studying.

4.2.2 The scoring method of glioblastoma and metastatic melanoma cellular states signatures

Glioblastoma signature by Neftel et al. and the metastatic melanoma signature by Tirosh et al. have the same algorithm [13, 14].

Firstly, an initial condition must be satisfied, without which it is impossible to calculate the score: the study dataset must have at least 3000 genes for the glioblastoma signature and 2500 genes for the metastatic melanoma signature. That is because, as we will see later, to compute the signatures, it is necessary to create lists one hundred times larger than the number of genes in the dataset divided into 30 or 25 bins, respectively; so there must be at least the number of bins multiplied by 100 [13, 14]. For simplicity, I will show the code for the Neftel signature calculated with 30 bins; the code for the Tirosh signature is available in the appendix (Code A.1).

```
if(nrow(dataset)<3000){stop(
"dataset must have at least 3000 genes to compute
    the signature")}
```

Both signatures require the transformation of the normalised gene counts for each cell into transcripts per kilobase million (TPM). I performed this step using an internal function of the *signifinder* package called "dataTransformation" [13].

```
dataset <- .dataTransformation(
    dataset, datasetm, "TPM", hgReference,
       nametype)
datasetm_n <- as.matrix(
    assays(dataset)[["TPM"]])
```

The next step involves calculating the expression level for each gene i in each cell j using the formula $E_{i,j} = log_2(TPM_{i,j}/10 + 1)$ [13]:

```
exp_lev <- log2(datasetm_n/10+1)
```

At this point, I calculated the relative expression for each value in the dataset, which is given by the difference between the gene expression level in the cell and the average expression of the same gene in all cells of the dataset:

```
    rel_exp <- exp_lev - rowMeans(exp_lev,
                                  na.rm = TRUE)
```

Subsequently, I computed the aggregate expression of each gene given by the average of the log2-transformed values of the gene. I then sorted the gene means and divided them into 30 bins, with this method I have 30 groups of genes with similar expression values [13]. Finally, I retained only the gene names within each bin, thus obtaining 30 lists of genes with similar expression levels:

```
    agg_exp <- log2(rowMeans(datasetm_n,
                             na.rm = TRUE)+1)
    ea_bin <- split(
    sort(agg_exp, na.last = TRUE), factor(
        sort(round(x = rank(agg_exp) %% 30,
                   digits = 0))))
    ea_bin <- lapply(ea_bin, function(x){names(x)})
```

Once the 30 bins were obtained, I proceeded to calculate the score, which is given by the difference between the relative expression of the genes in each module of the signature and the relative expression of the genes in a control sample created based on the genes in the studied dataset. To construct the control sample G^{cont}, I took the genes of each module of the signature one by one and checked if they were present within a bin. If the gene was not there, I moved on to the next gene; if it was present, I saved the index of the bin containing the gene in an empty vector u. I then used this index to determine the bin with the gene of interest and selected 100 random genes from the same bin [13]. In this way, I created a control sample one hundred times larger than the gene set of the signature, as stated in the two studies.

```
    scores <- as.data.frame(
                    lapply(sign_list, function(x){
         Gcont <- unlist(lapply(x, function(y){
           u <- NULL
           for (i in seq_along(ea_bin)) {
            if (y %in% ea_bin[[i]]) {
                u <- i
                break}}
        sample(ea_bin[[u]][!(ea_bin[[u]] %in% x)],
                100)}))
```

After this, I assign a null value to all non-malignant cells in the dataset and assign the score to the cancer cells given by the difference between the relative expressions, as previously mentioned: $SC_j(i) = average[Er(G_j, i)] - average[Er(G_j^{cont}, i)]$, where $SC_j(i)$ is the score of gene set j in cell i; $Er(G_j, i)$ is the relative expression of the gene set of the signature for each cell, and $Er(G_j^{cont}, i)$ is the relative expression of the control gene set for each cell [13].

```
score <- rep(NA, ncol(dataset))
SC <- colMeans(
    rel_exp[x,], na.rm = TRUE)-colMeans(
            rel_exp[Gcont,], na.rm = TRUE)
score[isMalignant] <- SC
```

Both signatures, along with the Barkley cellular states signature, have been incorporated into a single function in the "SignatureFunction" file of *signifinder*: "stateSign". This function can calculate the signature of one of the three authors by changing the input parameters.

To calculate the score, it is necessary to provide the "dataset", i.e. the SingleCellExperiment object of the dataset being studied; the "nametype", which is the type of nomenclature used for the genes in the dataset: the three accepted parameters are "SYMBOL", "ENSEMBL", and "ENTREZ ID"; the "author", to identify which of the three signatures to compute among "Barkley", "Neftel", and "Tirosh"; "whichAssay" is a parameter indicating which normalised matrix present in the "assays" of the SingleCellExperiment to use. Finally, it is necessary to provide a boolean vector to the "isMalignant" parameter, indicating which cells in the dataset are malignant and which are not.

4.2.3 The scoring method of breast cancer subtypes signature

To compute the score of the breast cancer cellular subtypes, I constructed the function "SCSubtypeSign" which calculates the average read counts for each of the four modules for each cell [15]. To do this, I first transformed the expression values of the dataset with a log2 transformation:

```
datasetm_n <- log2(datasetm + 1)
```

Secondly, I assigned a null value to all the cells in the dataset and saved it in a variable s. For each gene set of each subtype, if the gene list has a length

greater than one, I then calculated the mean of the expression counts in each cell of the dataset corresponding to the genes of the signature. Conversely, if only one gene of a given subtype is present in the study dataset, the score will be given only by that value [15]. At this point, I entered the scores into the malignant cells of the dataset present in *s*.

```
scores <- as.data.frame(lapply(sign_list,
   function(x) {
         s <- rep(NA, ncol(datasetm))
         if (length(x)>1) {score colMeans(
            datasetm_n[x,], na.rm = TRUE)}
         else {score <- datasetm_n[x,]}
         s[isMalignant] <- score
         s }))
```

Finally, I entered this code too, into the function within the "SignatureFunction" file of *signifinder*, as I did with the previous signatures.

4.3 The case-study analysis

4.3.1 Case-study data

To test some of the public signatures I collected, I used *signifinder* on four single cell RNA sequencing data from a patient with a diagnosis of stage IV high grade serous ovarian cancer (HGSOC). The dataset is part of a collection of samples from an AIRC-funded project called SHOwMEOVC: "Spatiotemporal HeterOgeneity in Multicellular Ecosystems of OVarian Cancer at single-cell resolution." This project is dedicated to studying the temporal, spatial, and inter-patient heterogeneity of the tumour microenvironment (TME) of HGSOC. It involves the collection and comparison of samples from different metastatic sites in the abdominal cavity of the same patient, using scRNA-seq technology [24].

The data chosen for my case study includes four samples from four pelvic sites of a patient. The first sample is from a primary tumour, specifically in the ovary, before the patient underwent neoadjuvant chemotherapy. The other three samples were taken from metastatic sites: the omentum, paracolic guts, and the peritoneal site. These three samples were collected after neoadjuvant chemotherapy. This sampling approach allows us to understand how cancer develops post-therapy [24].

My analysis started after the quality control, read alignment and quantification and the construction of the four count matrices. Samples were provided

to me in the SingleCellExperiment data format. SingleCellExperiment data format is a class of objects in R that allows the storage and management of single-cell gene expression data. The *SingleCellExperiment* package, which contains this class, is built around the Bioconductor project's structures and aims to facilitate the analysis and visualisation of scRNA-seq data. The SingleCellExperiment object is organised into different compartments: for example, it contains "colData," which consists of metadata related to the cells, such as the cell type, labelled as "consensusTME" in my samples. It is also the container where the *signifinder* score matrices will be saved.

Another interesting compartment is "assays", which contains one or more matrices of expression data. In my case, the SingleCellExperiment contains the raw counts matrix "counts", the normalised data matrix "logcounts", and the matrix of counts transformed into TPM: "tpm".

The study stated that the raw expression counts, after filtering, were normalised using the deconvolution method. First, the cells were separated into preliminary clusters based on gene expression using the "quickCluster" function from the *scran* R package. At this point, the values were rescaled to compare the different clusters and were normalised by dividing each count by the appropriate size factor. Finally, a logarithmic transformation was applied using the "logNormCounts" function [24].

4.3.2 Computation of the signature scores with *signifinder* of each sample

I will now present an analysis of the four samples taken from a patient with HGSOV. The dataset is used to test how *signifinder* and the public scRNA-seq signatures, I have included, can be used.

The calculation of scores for each dataset was done using the "multipleSign" function of *signifinder* [3]. This function allows for the computation of multiple signatures simultaneously. I chose all the signatures available in the package that were developed from both bulk RNA sequencing and scRNA-seq, which can be calculated in all tissue types. In the ovary sample, I also included all the signatures that can be calculated in the ovary.

I built the following code for computing the signatures in the ovary sample dataset. For the omentum, paracolic guts, and peritoneum samples, the code is the same; the only difference is the parameter indicating the tissue in which the signature scores can be calculated: I only included "pan-tissue" (all tissues) and not "ovary".

```
library(signifinder)
sce44 <- readRDS("../sce44_consensusTME.rds")
sce <- multipleSign(sce44, nametype ="SYMBOL",
        inputType = c("sc", "rnaseq"),
        whichAssay = "logcounts",
        tissue = c("ovary", "pan-tissue"),
        isMalignant = sce44@colData@listData$
           consensusTME == "Cancer_cells")
```

Firstly, the *signifinder* package is loaded into the R workspace and the SingleCellExperiment, which is in a .rds file, is read using the "readRDS" function. As previously stated, the signature scores are calculated with the "multipleSign" function and will be saved in a numeric vector within the "colData" section of the SingleCellExperiment. Once the data is saved, it is possible to visualise the scores using some functions in *signifinder*.

4.3.3 Visualisation with *signifinder* R package

First of all, I can evaluate the quality of the signatures calculated in the studied datasets and how well they fit the dataset using the "EvaluationSignPlot" function [3].

In the graphs created for the datasets corresponding to the ovary, omentum, paracolic gutters and peritoneum sites, some signatures were omitted, such as those related to neurological cancers from Barkley's "stateSign" function: State_Barkley_NPC, State_Barkley_OPC, and State_Barkley_AC.

The "EvaluationSignPlot" provides a multipanel graph representing various characteristics and functionalities. I will discuss the first panel later. In the second panel, the package shows the percentage of genes used for score calculation of each signature in the different samples. What can be observed from the four plots, is that the thresholds for each gene are all above 80%, and most are above 95%. The percentages of genes used are therefore good for all signatures.

In the third panel, a box plot shows the percentage of zero values of the signature genes for the cells in the dataset. This plot can help detect dropout events: the loss of mRNA transcripts during sequencing or the low gene expression in a cell, leading to numerous zero values.

In the fourth section, *signifinder* presents a graph showing the correlation between the score value and the number of zeros in blue, and the score value and the total expression value in the cell in pink.

Finally, in the first panel, the goodness percentage of the signature can be found, which combines the results of the three previously described panels, providing a value that summarises the quality of the signature in the sample [3].

What can be observed for the ovary sample dataset is that ten scRNA-seq based signatures out of thirteen have a goodness percentage above 68%, from which it can be deduced that they fit this dataset very well (Fig. 4.1).

The percentages related to this type of signature for the other three samples are slightly less reliable. In particular, the goodness percentages are more scattered, although for each dataset, more than half have a value above 60% (Fig. 4.2, 4.3, 4.4).

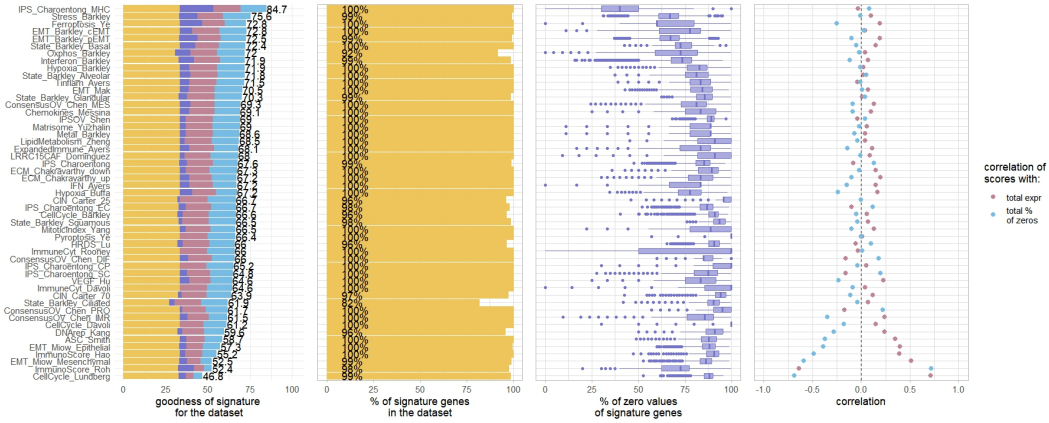

Figure 4.1: Signature evaluation plot for the ovary sample.

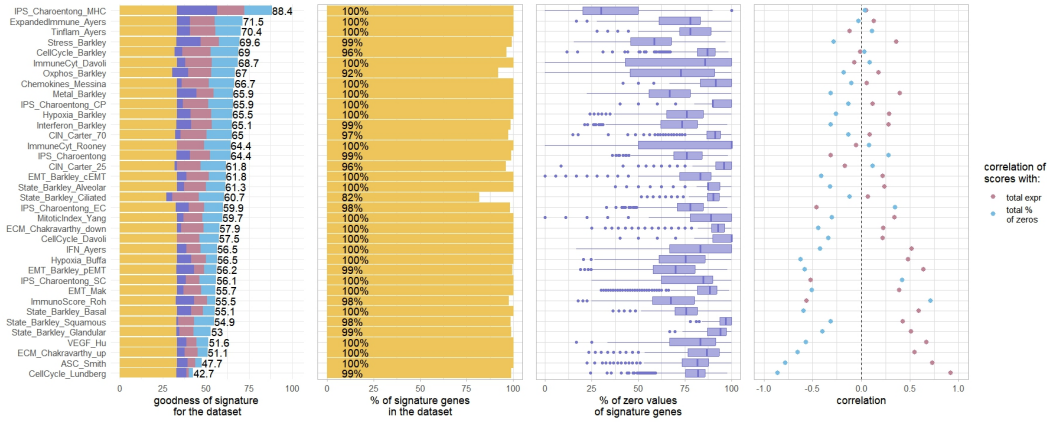

Figure 4.2: Signature evaluation plot for the omentum sample.

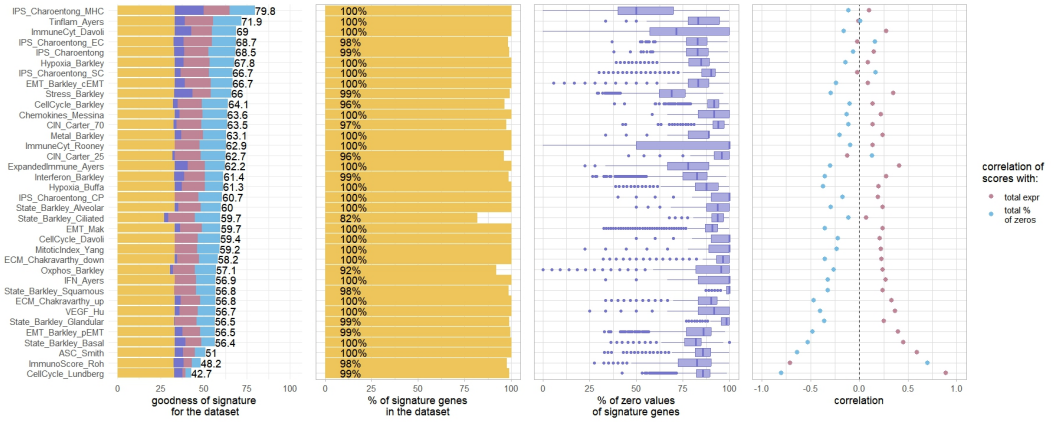

Figure 4.3: Signature evaluation plot for the paracolic guts sample.

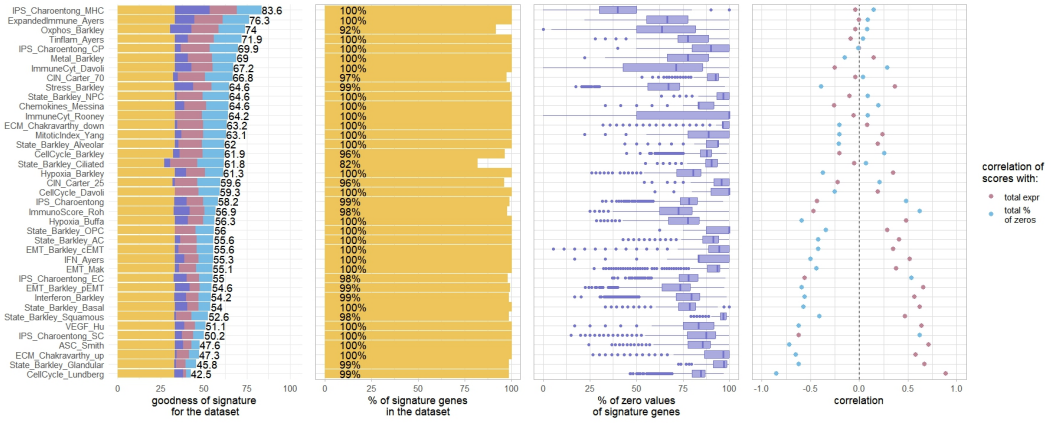

Figure 4.4: Signature evaluation plot for the peritoneum sample.

Among the explorative plots in *signifinder* we have the pairwise correlation plot among signatures through the "correlationSignPlot" function [3]. The correlation matrix is composed of all the correlation coefficients between all the possible pairs of signatures: the more negative the coefficient, the more the values of the two signatures in all the cells anti-correlate; the more positive the coefficient, the more the values of the two signatures in all the cells correlate. The coefficient values are also visually represented by an intense red colour for positively correlated signatures, gradually changing to an intense blue for anti-correlated signatures.

With the correlation matrix, it is possible to observe clusters of correlation between the signatures. In the appendix the description of the corresponding signature in the Table A.1.

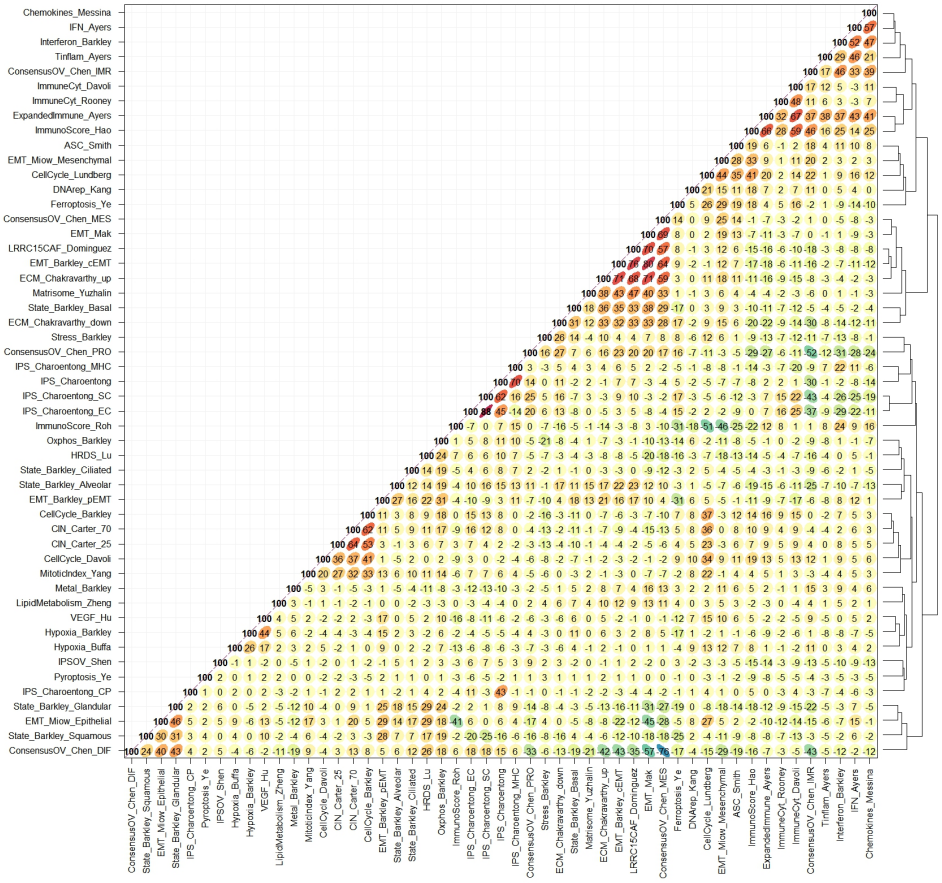

Figure 4.5: Signature correlation plot for the ovary sample.

In figure 4.5, the correlation matrix of the ovary sample dataset is presented. Compared to the other matrices, it predominantly has less evident correlation coefficients.

In particular, a cluster of signatures with positive correlation can be detected: ConsensusOV_Chen_MES, EMT_Mak, LRRC15CAF_Dominguez, EMT_Barkley_cEMT and EMC_Chakravarthy_up. These signatures are also negatively correlated with ConsensusOV_Chen_DIF.

Secondly, another cluster of positively correlated signatures can be identified: CellCycle_Barkley, CIN_Carter_70, CIN_Carter_25 and CellCycle_Davoli.

Although no signatures based on scRNA-seq technology are present, it is possible to observe the relationship between some signatures linked to the immune response: IPS_Charoentong is positively correlated with IPS_Charoentong_MHC, IPS_Charoentong_SC, and IPS_Charoentong_EC, the latter two are also positively correlated with each other.

In the other three datasets, we predominantly have a greater number of

pairwise correlations, with larger clusters of signatures.

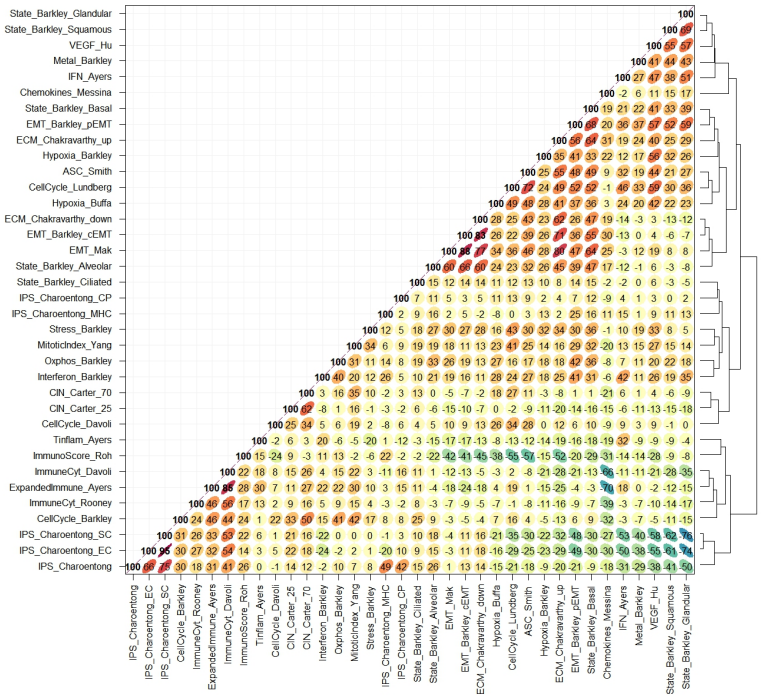

Figure 4.6: Signature correlation plot for the omentum sample.

In the omentum sample dataset (Fig. 4.6), five positively correlated clusters can be identified:

1. ExpandedImmune_Ayers, ImmuneCyt_Rooney, IPS_Charoentong_SC, IPS_Charoentong_EC, IPS_Charoentong and ImmuneCyt_Davoli. These signatures are all related to the immune response.

2. EMT_Barkley_cEMT, State_Barkley_Alveolar, EMT_Barkley_pEMT, ASC_Smith, CellCycle_Lundberg, EMT_Mak, EMC_Chakravarthy_up, State_Barkley_Basal and EMC_Chakravarthy_down.

3. State_Barkley_Glandular, State_Barkley_Squamous, Metal_Barkley, EMT_Barkley_pEMT, State_Barkley_Basal, Hypoxia_Barkley, Hypoxia_Buffa, EMC_Chakravarthy_up, ASC_Smith, VEGF_Hu, CellCycle_Lundberg and IFN_Ayers.

4. The signature CellCycle_Barkley is linked to the signatures ExpandedImmune_Ayers, ImmuneCyt_Davoli, Oxphos_Barkley and MitoticIndex_Yang.

The correlation matrix also shows a negative correlation between IPS_Charoentong_SC and IPS_Charoentong_EC and the signatures State_Barkley_Glandular, State_Barkley_Squamous, VEGF_Hu, Metal_Barkley, IFN_Ayers and EMT_Barkley_pEMT.

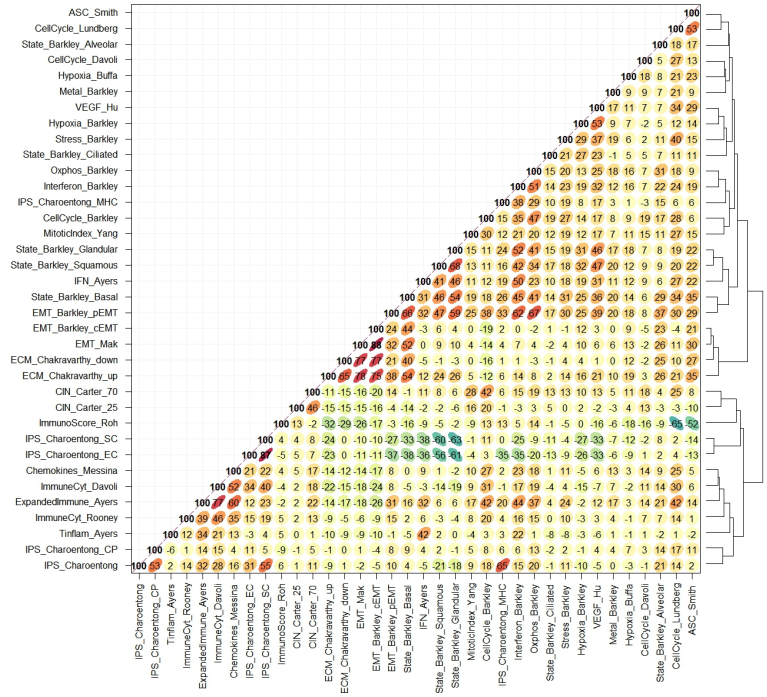

Figure 4.7: Signature correlation plot for the paracolic guts sample.

In the matrix referring to the paracolic gutters sample (Fig. 4.7), it is also evident that IPS_Charoentong_SC and IPS_Charoentong_EC are negatively correlated with EMT_Barkley_pEMT, IFN_Ayers, State_Barkley_Glandular, State_Barkley_Squamous and State_Barkley_Basal.

Additionally, two positively correlated clusters are evident: the first concerns EMT_Barkley_cEMT, EMT_Mak, EMC_Chakravarthy_down and EMC_Chakravarthy_up; the second is related to EMT_Barkley_pEMT, State_Barkley_Basal, State_Barkley_Glandular, State_Barkley_Squamous, Interferon_Barkley and Oxphos_Barkley.

In the correlation plot of the peritoneum site sample (Fig. 4.8), it can be observed that IPS_Charoentong_SC, IPS_Charoentong_EC and the general signature IPS_Charoentong are positively correlated with each other and negatively correlated with Hypoxia_Barkley, Stress_Barkley, IFN_Ayers, Interferon_Barkley, State_Barkley_Squamous, State_Barkley_Glandular, EMT_Barkley_pEMT, State_Barkley_Basal, Hypoxia_Buffa, VEGF_Hu, CellCycle_Lundberg, ASC_Smith and EMC_Chakravarthy_up, which also tend to be positively correlated with each other.

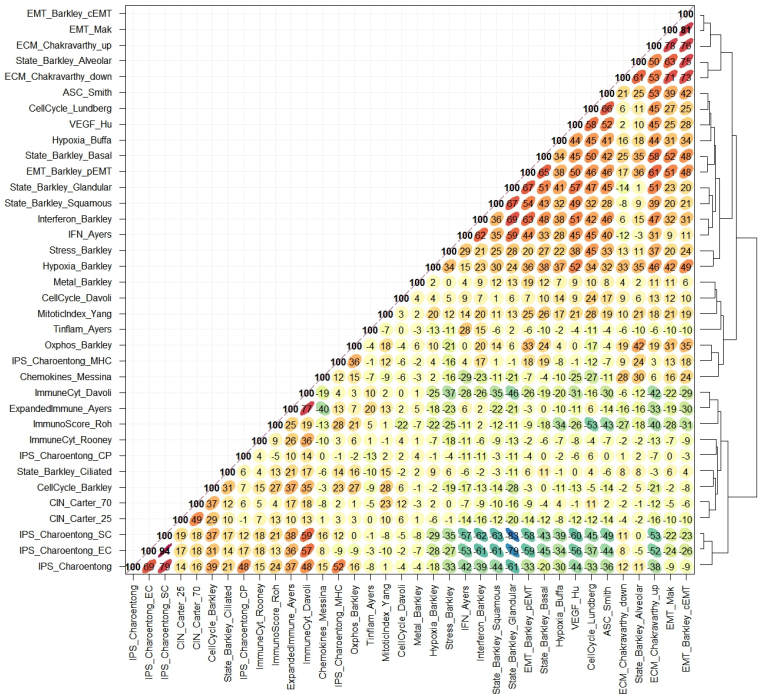

Figure 4.8: Signature correlation plot for the peritoneum sample.

Finally, we can observe that, in this sample too, EMT_Barkley_cEMT, EMT_Mak, EMC_Chakravarthy_down and State_Barkley_Alveolar are positively correlated.

In conclusion, I can observe that in the three metastatic samples there is a recurrence of correlated clusters, such as some epithelial-to-mesenchymal transition signatures, some immune response-related signatures and some signatures related to a stress condition, like hypoxia, VEGF, stress or metal response.

Additionally, these signatures which are linked to a condition of stress, are anti-correlated with the immune-related signatures: this means that the tumour microenvironment's stress conditions may suppress immune activity, potentially impacting the efficacy of the therapy in these groups of cells. So, understanding these correlations and anti-correlations can provide valuable insights into the tumour biology.

To observe how the expression of the above-mentioned signatures is distributed across cells, I decided to use t-SNE plots. t-SNE, or t-distributed stochastic neighbour embedding, works using a dimensionality reduction algorithm and helps to represent high-dimensional data in a two-dimensional scatter plot [25].

To create and export the t-SNE plots as images, I used the R packages *scater*, *ggplot2*, and *gridExtra*.

Regarding the four HGSOC samples, the t-SNE plots could be exploited to confirm what has been defined with the correlation plots.

For example, the ConsensusOV_Chen_DIF signature indicates a differentiated subtype of the tumour with high expression of HGSOC markers. This subtype is usually associated with the longest progression-free survival (PFS), a measure indicating the period of time in which the patient survives without the disease aggravating [3]. Malignant cells classified as DIF have lower expression of expression related to epithelial-to-mesenchymal transition, as seen in the scores of the EMT signatures by Mak and even more clearly in the cEMT by Barkley or a mesenchymal state captured by the ConsensusOV_Chen_MES subtype, in effect, in the correlation matrix these signatures were anti-correlated to ConsensusOV_Chen_DIF. Therefore, we can identify cellular clusters, where mesenchymal tumour state expression prevails, indicating cells more prone to dissemination, especially if compared to cells where the differentiated subtype prevails (Fig. 4.9).

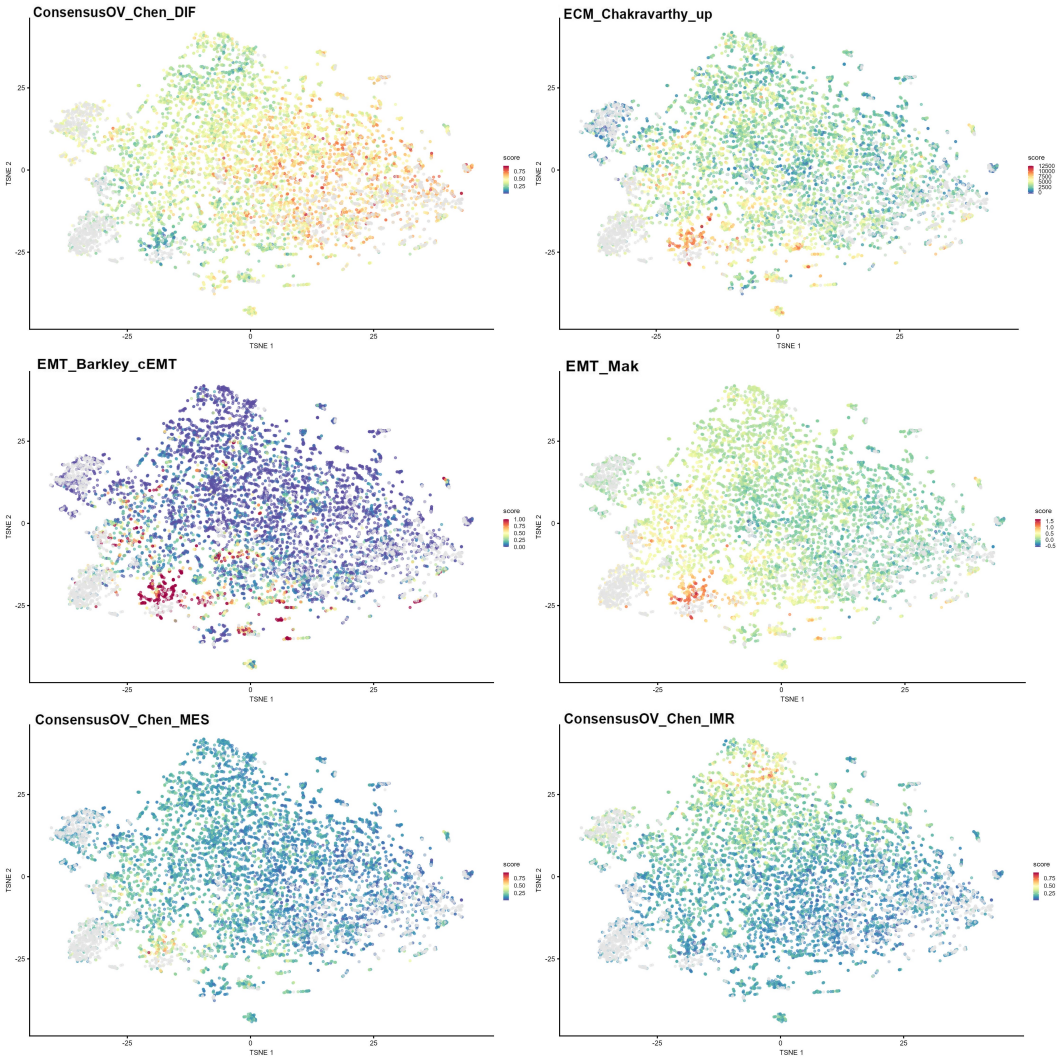

Figure 4.9: tSNE with cells coloured by signature. ConsensusOV_DIF is compared to EMC_Chakravarthy_up, EMT_Barkley_cEMT, EMT_Mak, ConsensusOV_Chen_MES and ConsensusOV_Chen_IMR.

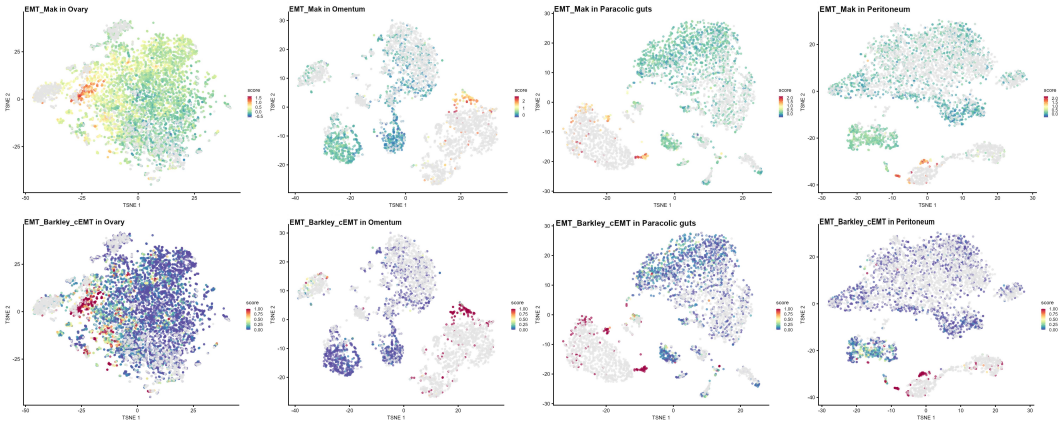

Figure 4.10: tSNE with cells coloured by signature. Comparison between EMT_Mak and EMT_Barkley_cEMT in all samples.

Continuing with EMT, it is also interesting to compare some signatures based on single-cell RNA sequencing and bulk RNA sequencing.

In particular, by comparing the EMT signatures in the datasets, we can deduce that EMT_Barkley_cEMT clusters similarly to EMT_Mak, highlighting the complete EMT transitions (Fig. 4.10).

Indeed, comparing the genes expressed in the Mak and complete Barkley signatures using the "geneHeatmapSignPlot" function of *signifinder* (Fig. 4.11), we can see that the genes shared between the two signatures constitute most of the EMT_Barkley_cEMT gene set.

Nevertheless, the scRNA-seq-based signature seems more efficient, showing less homogeneous scores for each cell in the different clusters.

I also found it interesting to compare some cellular responses such as the interferon response, hypoxic condition, and angiogenesis. From figure 4.12, it is possible to observe a tendency for the expression of these signatures to overlap, particularly in the three post-therapeutic metastatic datasets. It is well known that the angiogenic response can be activated by hypoxic conditions, and the presence of the Interferon_Barkley signature expression might indicate the activation of mechanisms that promote the inhibition of angiogenesis.

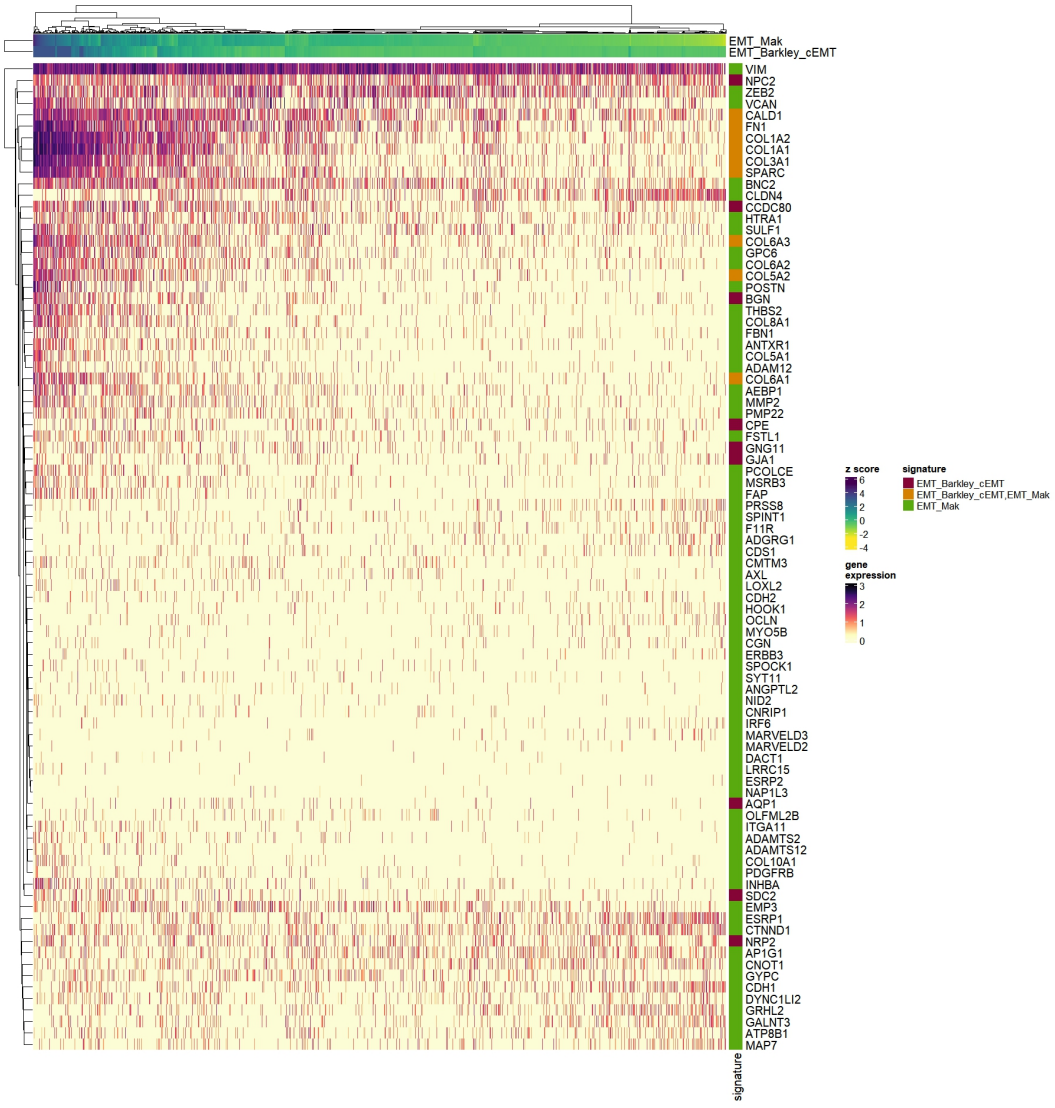

Figure 4.11: Heatmap of the expression values of EMT_Mak and EMT_Barkley_cEMT genes using geneHeatmapSignPlot function.

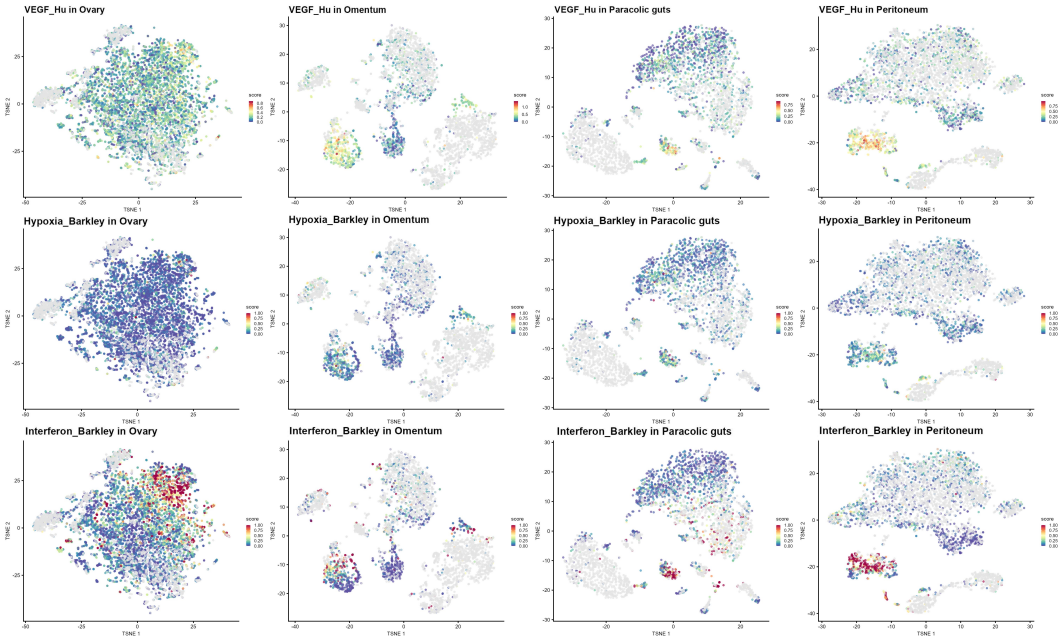

Figure 4.12: tSNE with cells coloured by signature. Comparison among angiogenic, hypoxic and interferon responses in all samples.

Finally, it is possible to visualise cellular expression through heatmaps. *signifinder* has a specific function called "heatmapSignPlot" that allows comparing the scores of the signatures being studied [3]. The function standardises the scores using a z-score transformation. It also allows dividing the heatmap into columns by cell type. This function enables the use of all the parameters of the "Heatmap" function from the R package *ComplexHeatmap*. Thanks to this, I was able to group each signature by topic to observe similar signatures more easily.

By comparing this type of plot for each dataset, I noticed that the expression of the signatures in the cells of the heatmap representing the ovary sample (Fig. 4.13) is distributed more uniformly compared to the other three samples (Fig. 4.14,4.15,4.16), making it less perceptible to identify cellular subgroups with specific characteristics conferred by the signatures. In the metastatic sample datasets, a clearer subdivision of cells expressing certain signatures can be observed.

All three heatmaps tend to have similar gene expression patterns.

Firstly, in fibroblasts, a low expression of signatures related to an immune response and a high profile for EMC_Chakravarthy signatures, linked to cancer-associated ECM genes and the epithelial-to-mesenchymal transition, can be observed. Additionally, some cells show a hypoxic state.

Conversely, in cytotoxic and NK cells, the expression profile highlights a high immune response and a low expression of ECM and EMT-related signatures.

In cancer cells, a more complex profile is evident, with certain groups of cells showing a low immune score for IPS_Charoentong, IPS_Charoentong_SC, and IPS_Charoentong_EC signatures, and a high level of scores related to ECM, EMT, certain cellular states like State_Barkley_Glandular, State_Barkley_Squamous, and State_Barkley_Basal, and cellular processes signatures like hypoxia, angiogenesis, oxidative phosphorylation, and interferon response. Other cellular groups, on the other hand, tend to have a higher immune response.

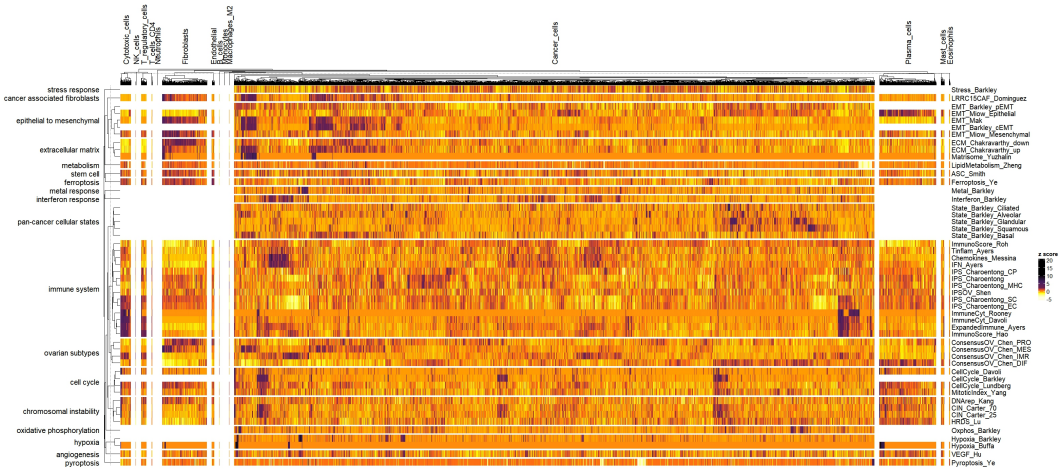

Figure 4.13: Signature heatmap for the ovary sample.

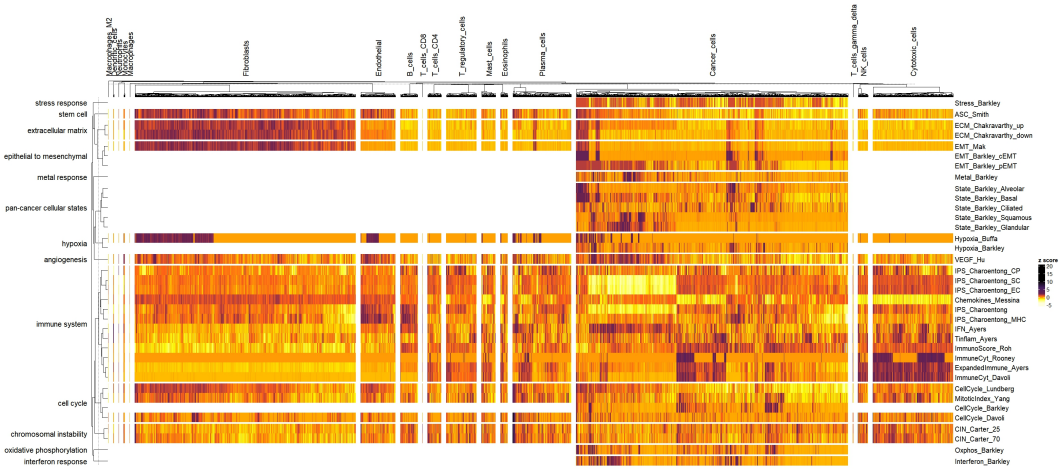

Figure 4.14: Signature heatmap for the omentum sample.

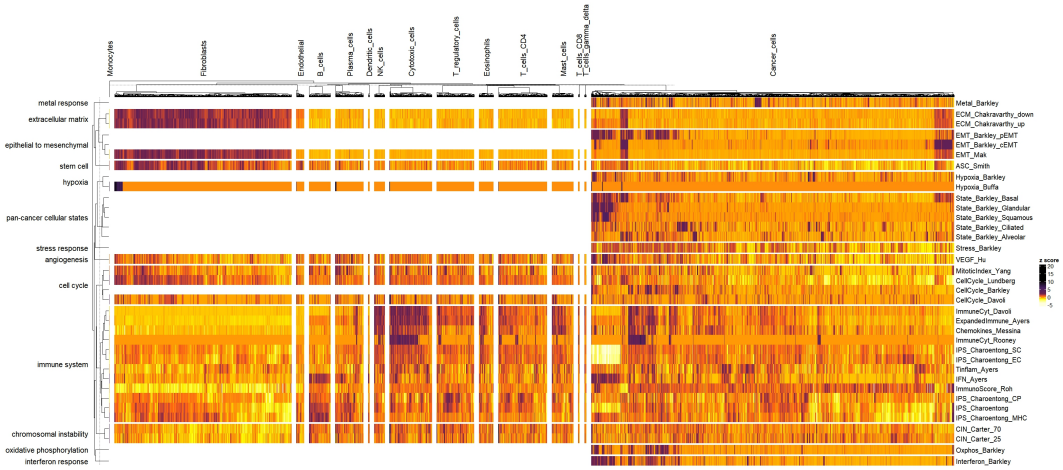

Figure 4.15: Signature heatmap for the paracolic guts sample.

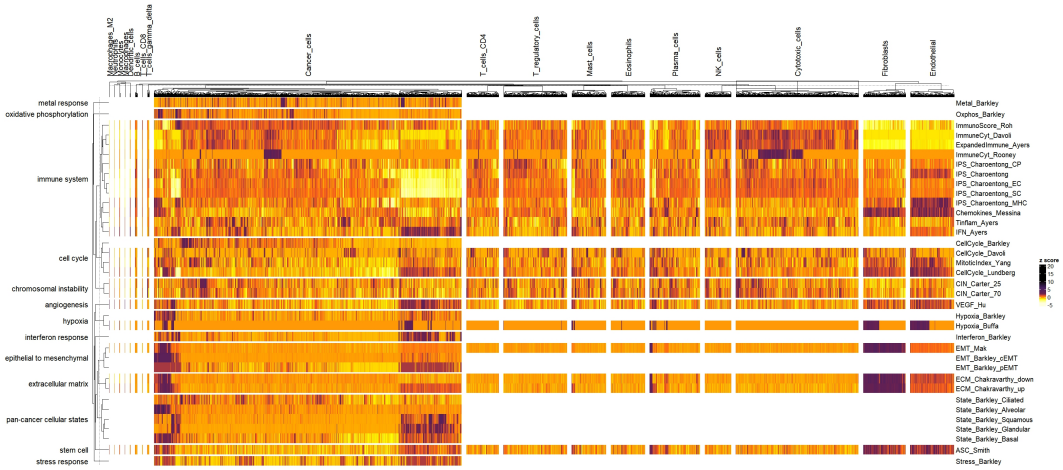

Figure 4.16: Signature heatmap for the peritoneum sample.

From all the plots shown so far, particularly from the correlation matrices and gene expression profiles of the heatmaps, two main aspects can be deduced. The first concerns a change in gene expression from the primary tumour in the ovary to the metastatic sites sampled after chemotherapy treatment. The tumour cells in the primary site dataset have more homogeneous and difficult-to-characterise expression levels. The second aspect concerns the cellular expression of the three tumour sites subjected to neoadjuvant chemotherapy: it is possible to identify cellular groups that have a low expression for all signatures, presenting only a slightly higher immune response. Other cellular groups, however, show a profile in which signatures might be linked to greater tumour complexity, due to the overlap of multiple cellular states; and to attempts at cell survival and proliferation, such as angiogenesis and a high mesenchymal potential.

Chapter 5

Conclusion

This thesis provides an insight into the impact that high-resolution technologies, such as single-cell RNA sequencing, could have on cancer research. In particular, it presents a study highlighting the importance of scRNA-seq-based signatures and the great need for a tool that can help users exploit them in their studies.

Thanks to the implementation of new public gene expression signatures created with high-resolution technologies in *signifinder*, it is now possible to have an improved version of the package, which ensures more detailed and precise analysis of users' tumour samples.

The collection contains various signatures related to processes and cellular states that can be detected in all tumours or in specific ones such as glioblastoma, metastatic melanoma, or breast cancer, thus allowing their use in a wide variety of samples.

The functions for computing the scores that I built and incorporated into *signifinder*, along with all the auxiliary functions of the package, have made it possible to create an example analysis of a case study.

The results of the case study on some HGSOV samples highlighted tumour development before and after chemotherapy, revealing cellular heterogeneity both spatially and temporally. Spatial heterogeneity is identified within the tumour samples where it is possible to observe cellular clusters showing profiles with resistance characteristics. Temporal heterogeneity is visible through a comparison between the sample in the primary tumour site and those in metastases; a less homogeneous distribution of expression profile scores can be observed in metastatic sites, likely due to tumour changes caused by neoadjuvant chemotherapy.

It was also possible to compare some signatures on the same topic created

with bulk sequencing and single-cell sequencing, This demonstrates that, although applied to scRNA-seq data, the bulk-derived signature might reflect an average profile across all the types of cells that could not fully capture the heterogeneity of individual cells.

Therefore, single cell RNA sequencing derived signatures are indispensable to fully characterise the tumour heterogeneity beyond the classification of cell types. However, to properly work the single cell signature should cover many and multiple cell biology aspects and should be representative and tested in a large number of patients and this is currently far from the state-of-art scenario.

Increasingly developing and integrating signatures from high-resolution transcriptome technologies into tools like *signifinder* will make the tumour characterization easier and the signatures more accessible and usable for everyone.

Future perspectives for the *signifinder* package are promising and multifaceted. Continued efforts will focus on expanding the collection of both bulk and scRNA-seq-based gene expression signatures to further enhance the package's utility across a wider range of cancer types and research applications. Additionally, integrating spatial transcriptomics signatures will be a key objective, as this technology allows for the preservation of spatial context within the tumour microenvironment, providing an added layer of precision and insight. By continually updating and enriching the database with these diverse signatures, *signifinder* will become an even more powerful and indispensable tool for researchers, enabling comprehensive analyses of tumour samples. This ongoing development would facilitate deeper understanding of tumour biology and could contribute to the advancement of personalised cancer therapies.

Appendix A

Supplementary material

Code A.1: Metastatic melanoma signature by Tirosh et al. scoring function.

```
stateSign <- function(
    dataset, nametype = "SYMBOL", author = "Tirosh", whichAssay
        = "norm_expr",
    isMalignant = NULL, hgReference = "hg38") {

  .consistencyCheck(nametype, "stateSign")
  .isMalignantCheck(isMalignant, dataset)

  if(nrow(dataset)<2500){stop(
      "dataset must have at least 2500 genes to compute the
         signature")}

  datasetm <- .getMatrix(dataset, whichAssay)
  dataset <- .dataTransformation(
    dataset, datasetm, "TPM", hgReference, nametype)
  datasetm_n <- as.matrix(assays(dataset)[["TPM"]])

  sign_df <- State_Tirosh
  sign_df$SYMBOL <- .geneIDtrans(nametype, sign_df$SYMBOL)

  .percentageOfGenesUsed(
    "stateSign", datasetm_n,
    sign_df$SYMBOL[sign_df$class == "MITF"], "MITF")
  .percentageOfGenesUsed(
    "stateSign", datasetm_n,
    sign_df$SYMBOL[sign_df$class == "AXL"], "AXL")

  sign_df <- sign_df[sign_df$SYMBOL %in% rownames(datasetm_n),
     ]
  sign_list <- split(sign_df$SYMBOL, sign_df$class)
  names(sign_list) <- paste0("State_Tirosh_", names(sign_list))

  datasetm_n <- datasetm_n[,isMalignant]
```

```
  exp_lev <- log2(datasetm_n/10+1)
  rel_exp <- exp_lev - rowMeans(exp_lev,  na.rm = TRUE)

  agg_exp <- log2(rowMeans(datasetm_n, na.rm = TRUE)+1)
  ea_bin <- split(
    sort(agg_exp, na.last = TRUE), factor(
      sort(round(x = rank(agg_exp) %% 25, digits = 0))))
  ea_bin <- lapply(ea_bin, function(x){names(x)})

  scores <- as.data.frame(lapply(sign_list, function(x){
    Gcont <- unlist(lapply(x, function(y){
      u <- NULL
      for (i in seq_along(ea_bin)) {
        if (y %in% ea_bin[[i]]) {
          u <- i
          break}}
      sample(ea_bin[[u]][!(ea_bin[[u]] %in% x)], 100)}))
    score <- rep(NA, ncol(dataset))
    SC <- colMeans(
      rel_exp[x,], na.rm = TRUE)-colMeans(rel_exp[Gcont,], na.
          rm = TRUE)
    score[isMalignant] <- SC
    score
  }))

  return(.returnAsInput(
    userdata = dataset, result = t(scores), SignName = "",
        datasetm))

}
```

Table A.1: Table with all the signatures based on scRNA-seq and all the signatures based on bulk RNA sequencing used for the case study. The table also offers a description of the score for each signature.

scoreLabel	tumor	developedWith	description
EMT_Miow_Epithelial, EMT_Miow_Mesenchymal	ovarian cancer	microarray, rnaseq	Double score obtained with ssGSEA to establish the epithelial- and the mesenchymal-like status in ovarian cancer patients.
EMT_Mak	pan-cancer	microarray, rnaseq	Score of the level of epithelial or mesenchymal status in cancer. Positive score is correlated with mesenchymal while negative score with epithelial.
EMT_Barkley_cEMT, EMT_Barkley_pEMT	pan-cancer	sc	Two cancer module scores are detected and their expression defines recurrent cancer cell states related to epithelial-mesenchymal transition: a complete mesenchymal module (cEMT) and a partial mesenchymal module (pEMT) lacking canonical mesenchymal markers such as collagen genes. The two modules may represent two pathways converging on phenotypic properties conferred by mesenchymal differentiation including migration and drug resistance. The expression level of each module is scored in individual cells: if the score is higher than 0.5, the module is considered expressed.
Pyroptosis_Ye	ovarian cancer	rnaseq	Score is based on risk coefficients and expression data of selected pyroptosis genes selected for their association with survival in ovarian cancer patients. Higher the score higher the risk.
Ferroptosis_Ye	ovarian cancer	microarray, rnaseq	A ferroptosis-related prognostic gene signature for ovarian cancers. High scores mean high-risk, poor overall survival and low immune cells infiltration.
LipidMetabolism_Zheng	epithelial ovarian cancer	rnaseq	A prognostic signature based on lipid metabolism for ovarian cancer patients. Higher the scores higher the risk and poorer the overall survival of patients.
Hypoxia_Buffa	pan-cancer	microarray	A highly prognostic signature. The score increasement reflects hypoxia activity.
Hypoxia_Barkley	pan-cancer	sc	The module score expression defines a recurrent cancer cell state related to hypoxia activity. The expression level of each module is scored in individual cells: if the score is higher than 0.5, the module is considered expressed.
ImmunoScore_Hao	epithelial ovarian cancer	microarray, rnaseq	An immune related signature to investigate the in situ immune activity in ovarian cancer and the response to chemotherapy. High immune score displayed overall high expression of favorable prognostic genes.
ImmunoScore_Roh	pan-cancer	rnaseq	The score is based on expression of genes involved in cytolytic markers, HLA molecules , IFN-γ pathway genes, chemokines, and adhesion molecules. It is used to investigate immune activation in tumor microenvironment, higher the score higher the immune system activation on relation to tumor rejection.
ConsensusOV_Chen_IMR, ConsensusOV_Chen_DIF, ConsensusOV_Chen_PRO, ConsensusOV_Chen_MES	high-grade serous ovarian carcinoma	microarray, rnaseq	It implements a consensus classifier of the four major subtype classifiers for high-grade serous ovarian cancer as described by Helland et al. (PLoS One, 2011), Bentink et al. (PLoS One, 2012), Verhaak et al. (J Clin Invest, 2013), and Konecny et al. (J Natl Cancer Inst, 2014), thereby providing reliable stratification of patients with high-grade serous ovarian tumors of clearly defined subtype.
IPS_Charoentong IPS_Charoentong_MHC, IPS_Charoentong_CP, IPS_Charoentong_EC, IPS_Charoentong_SC	pan-cancer	rnaseq	Five immune related scores are returned: the overall immune score (IPS), the EC score for effector cells (activated CD8+/CD4+ T cells), the SC score for immuno suppressive cells, the MHC score for antigen processing molecules, the CP score for co-inhibitory and co-stimulatory molecules.
MitoticIndex_Yang	pan-cancer	rnaseq	The mitotic-index is constructed from genes that have been highly validated as being cell proliferation markers. The score reflects the fraction of dividing cells in a sample and can be used as a predictors of normal/-cancer status.
ImmuneCyt_Rooney	pan-cancer	microarray, rnaseq	The score is a quantitative measure of immune cytolytic activity based on transcript levels of two key cytolytic effectors, granzyme A and perforin. High scores are associated with counter-regulatory immune responses and improved prognosis.
IFN_Ayers	pan-cancer	rnaseq	IFN-γ Score based on genes related to IFN-γ predicts clinical response to PD-1 checkpoint blockade. Higher scores are found in responders. Signature derives from patients undergoing treatment with Pembrolizumab in clinical trials using multiple distinct tumor types. The score is higher in responders.
ExpandedImmune_Ayers	pan-cancer	rnaseq	The score predict clinical response to PD-1 checkpoint blockade based on genes associated with cytolytic activity, pro-inflammatory cytokines/chemokines, T cell markers, NK cell activity, antigen presentation and T cell checkpoints. The score is higher in responders.
Tinflam_Ayers	pan-cancer	rnaseq	The score is derived by the expression of T cell–inflamed representing genes and it predicts the response to Pembrolizumab across multiple solid tumors. A T cell–inflamed phenotype is necessary for the clinical activity of PD-1–/PD-L1–directed monoclonal antibodies. The score is higher in responders.
CIN_Carter_25, CIN_Carter_70	pan-cancer	microarray	The score characterizes aneuploidy in tumor samples based on coordinated aberrations in expression of genes localized to each chromosomal region. Higher the score higher the total level of chromosomal aberration. Net overexpression of this signature was predictive of poor clinical outcome in six cancer types.
CellCycle_Lundberg	pan-cancer	rnaseq	It is a representative of general cell-cycle activity and could be applied to any tissue sample. Higher scores represent a worse prognosis.
CellCycle_Davoli	pan-cancer	microarray, rnaseq	Cell cycle signature score represents the expression of a set of genes considered molecular markers of proliferation. Higher the score higher the proliferation. High level is associated with high level of SCNA (somatic copy number alterations).
CellCycle_Barkley	pan-cancer	sc	The module score expression defines the subset of cancer cells that is cycling at the time of the sampling. The expression level of each module is scored in individual cells: if the score is higher than 0.5, the module is considered expressed.

ASC_Smith	pan-cancer	microarray, rnaseq	The adult stem cell (ASC) signature explores the relationship between human stem cell and cancer transcriptional programs. Higher the scores, aggressive the cancer, poorer the clinical outcome. The ASC signature is associated with the presence of specific genomic and alterations and methylation profiles.
ImmuneCyt_Davoli	pan-cancer	microarray, rnaseq	Cytotoxic immune signature score represents the expression of a set of genes considered molecular markers of cytotoxic CD8+ T cells and NK cells. Higher score higher immune response. The score is associated with arm and chromosome aneuploidy thus low level of SCNA (somatic copy number alterations).
Chemokines_Messina	pan-cancer	microarray	Chemokine score to predict host immune reaction and the formation of unique ectopic lymph node-like structures associated with better overall survival. Higher the score the more lymph node-like structures are present better the prognosis in melanoma patients.
ECM_Chakravarthy_up, ECM_Chakravarthy_down	pan-cancer	rnaseq	Two ECM ssGSEA scores are derived from cancer-associated extracellular matrix (ECM) genes and predict response to immune checkpoint blockade. Higher the ECM_up score and lower the ECM_down score worst the prognosis. Scores are inversely correlated with tumor purity and ECM_up directly correlated with CAFs presence and TGF-b activation.
HRDS_Lu	ovarian cancer, breast cancer	microarray, rnaseq	A score based on homologous recombination deficiency (HRD), higher the score better the platinum response, the patient outcome, and higher presence of BRCA mutations or inactivation.
VEGF_Hu	pan-cancer	microarray	VEGF profile is a prognostic score that correlates with glycolytic enzymes, hypoxia and vessel formation in distant metastasis, higher the score worst the overall and relapse-free survival.
DNArep_Kang	serous ovarian cystadenocarcinoma	microarray	A DNA Repair based score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. The higher the score the better the prognosis.
IPSOV_Shen	ovarian cancer	microarray	Single sample gene set enrichment (ssGSEA) analysis was used for the immune genes from ImmPort database to develop an immune-based prognostic score for OV (IPSOV). IPSOV is a prognostic signature which stratifyies patients into low- and high-immune risk score and could be used to predict overall survival outcome in patients with ovarian cancer. The patients with low IPSOV scores have longer survival time.
State_Barkley_Alveolar, State_Barkley_Basal, State_Barkley_Squamous, State_Barkley_Glandular, State_Barkley_Ciliated, State_Barkley_AC, State_Barkley_OPC, State_Barkley_NPC	pan-cancer	sc	Eight pan-cancer cellular states scores related to cell identity are returned. AC, OPC and NPC are neurological cancer-specific modules. The expression level of each module is scored in individual cells: if the score is higher than 0.5, the module is considered expressed.
LRRC15CAF_Dominguez	pancreatic adenocarcinoma, breast cancer, lung cancer, ovarian cancer, colon cancer, renal cancer, esophageal cancer, stomach adenocarcinoma, bladder cancer, head and neck squamous cell carcinoma	rnaseq	Signature characterizing a population of cancer associated fibroblasts (CAFs) programmed by TGFB and expressing LRRC15 protein. This population has been initially identified in the stroma of pancreatic ductal adenocarcinoma, but it was also present in multiple cancer types from TCGA. An increased expression of the signature was observed in patients who fail to respond to ICB therapy specifically in immune-excluded tumors.
Stress_Barkley	pan-cancer	sc	The module score expression defines a recurrent cancer cell state related to stress respose. The expression level of each module is scored in individual cells: if the score is higher than 0.5, the module is considered expressed.
Interferon_Barkley	pan-cancer	sc	The module score expression defines a recurrent cancer cell state related to interferon. The expression level of each module is scored in individual cells: if the score is higher than 0.5, the module is considered expressed.
Oxphos_Barkley	pan-cancer	sc	The module score expression defines a recurrent cancer cell state related to oxidative phosphorylation. The expression level of each module is scored in individual cells: if the score is higher than 0.5, the module is considered expressed.
Metal_Barkley	pan-cancer	sc	The module score expression defines a recurrent cancer cell state related to metal response. The expression level of each module is scored in individual cells: if the score is higher than 0.5, the module is considered expressed.
State_Neftel_MES1, State_Neftel_MES2, State_Neftel_AC, State_Neftel_OPC, State_Neftel_NPC1, State_Neftel_NPC2	glioblastoma	sc	Six glioblastoma cellular states scores are returned. MES2-like state is associated with high expression of mesenchymal-related, hypoxia-response, stress and glycolytic genes. This is an evidence that in some tumors the mesenchymal state is linked to hypoxia conditions and increased glycolysis. MES1-like state is also associated with high expression of mesenchymal-related genes, but it is hypoxia independent. The AC-like state includes astrocytic markers. The OPC-like state includes oligodendroglial markers. The NPC1-like state contains neural progenitor markers; OPC-related genes are included and they reflect the potential of NPCs to differentiate towards OPCs. The NPC2-like state also contains neural progenitor markers; neuronal lineage genes are included and they reflect the potential of NPCs to differentiate towards neurons.
State_Tirosh_MITF, State_Tirosh_AXL	metastatic melanoma	sc	Two metastatic melamoma cellular states are returned. MITF program includes MITF and other MITF target genes. The second state is AXL program. It includes genes related to AXL and it is linked to resistance to various targeted therapies such as treatment with RAF and MEK inhibitors.
SCSubtype_Wu_Basal, SCSubtype_Wu_Her2E, SCSubtype_Wu_LumA, SCSubtype_Wu_LumB	breast cancer	sc	The scores are the result of a developed single-cell method of intrinsic subtype classification to reveal recurrent neoplastic heterogeneity. Four scores are returned that describe two luminal-like subtypes, LumA and LumB, a basal-like subtype (Basal), and and a human epidermal growth factor receptor enriched subtype (Her2E). Each score is calculated for each cell, the highest score value indicates the subtype to which the cell belongs to.

Bibliography

[1] Ibiayi Dagogo-Jack and Alice T. Shaw. "Tumour heterogeneity and resistance to cancer therapies". In: *Nature Reviews Clinical Oncology* 15.2 (Feb. 2018), pp. 81–94.

[2] Frederic Chibon. "Cancer gene expression signatures – The rise and fall?" In: *European Journal of Cancer* 49.8 (May 2013), pp. 2000–2009.

[3] Stefania Pirrotta et al. *signifinder enables the identification of tumor cell states and cancer expression signatures in bulk, single-cell and spatial transcriptomic data*. Mar. 2023.

[4] Rory Stark, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years". en. In: *Nature Reviews Genetics* 20.11 (Nov. 2019), pp. 631–656.

[5] Christoph Ziegenhain et al. "Comparative Analysis of Single-Cell RNA Sequencing Methods". In: *Molecular Cell* 65.4 (Feb. 2017), 631–643.e4.

[6] Simone Picelli et al. "Full-length RNA-seq from single cells using Smart-seq2". In: *Nature Protocols* 9.1 (Jan. 2014), pp. 171–181.

[7] Rashid Ahmed et al. "Single-Cell RNA Sequencing with Spatial Transcriptomics of Cancer Tissues". In: *International Journal of Molecular Sciences* 23.6 (Mar. 2022), p. 3042.

[8] Atefeh Lafzi et al. "Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies". In: *Nature Protocols* 13.12 (Dec. 2018), pp. 2742–2757.

[9] Martin Morgan et al. *SummarizedExperiment: SummarizedExperiment container*. R package version 1.34.0. Bioconductor. 2024.

[10] Robert Amezquita et al. "Orchestrating single-cell analysis with Bioconductor". In: *Nature Methods* 17 (2020), pp. 137–145.

[11] Dario Righelli et al. "SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor". In: *Bioinformatics* 38.11 (2022), pp. -3.

[12] Dalia Barkley et al. "Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment". In: *Nature Genetics* 54.8 (Aug. 2022). Publisher: Nature Publishing Group, pp. 1192–1201. ISSN: 1546-1718.

[13] Cyril Neftel et al. "An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma". In: *Cell* 178.4 (Aug. 2019), 835–849.e21.

[14] Itay Tirosh et al. "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". In: *Science* 352.6282 (Apr. 2016). Publisher: American Association for the Advancement of Science, pp. 189–196.

[15] Sunny Z. Wu et al. "A single-cell and spatially resolved atlas of human breast cancers". In: *Nature Genetics* 53.9 (Sept. 2021). Publisher: Nature Publishing Group, pp. 1334–1347. ISSN: 1546-1718.

[16] Federico M. Giorgi, Carmine Ceraolo, and Daniele Mercatelli. "The R Language: An Engine for Bioinformatics and Data Science". In: *Life* 12.5 (Apr. 2022), p. 648.

[17] Keith Engwall and Mitchell Roe. "Git and GitLab in Library Website Change Management Workflows". In: *The Code4Lib Journal* 48 (May 2020).

[18] Hadley Wickham et al. *roxygen2: In-Line Documentation for R*. R package version 7.3.2, https://github.com/r-lib/roxygen2. 2024.

[19] Ziyi Zhao et al. "The Effect of Oxidative Phosphorylation on Cancer Drug Resistance". In: *Cancers* 15.1 (Dec. 2022), p. 62.

[20] Anthony A. Mercadante and Anup Kasi. "Genetics, Cancer Cell Cycle Phases". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024.

[21] Ernest C. Borden. "Interferons α and β in cancer: therapeutic opportunities from new insights". In: *Nature Reviews Drug Discovery* 18.3 (Mar. 2019), pp. 219–234.

[22] Manfei Si and Jinghe Lang. "The roles of metallothioneins in carcinogenesis". In: *Journal of Hematology & Oncology* 11.1 (Dec. 2018), p. 107.

[23] Erasmo Orrantia-Borunda et al. "Subtypes of Breast Cancer". In: *Breast Cancer*. Ed. by Department of Medical Education, Dr. Kiran C. Patel College of Allopathic Medicine, Nova Southeastern University, FL, USA and Harvey N. Mayrovitz. Exon Publications, Aug. 2022, pp. 31–42. ISBN: 978-0-645-33203-2.

[24] Laura Masatti. "From bulk to single-cell RNA-sequencing data: the tumor heterogeneity evolution in multicellular ecosystems of ovarian cancer". Tesi di Dottorato. Università degli Studi di Padova, 2024.

[25] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.