Università degli Studi di Padova

Dipartimento di Biologia

Corso di Laurea Magistrale in Biotecnologie Industriali

Collection and implementation of gene expression signatures for cancer data interpretation

Relatore: Prof. Chiara Romualdi
Dipartimento di Biologia

Correlatore: Dr. Enrica Calura
Dipartimento di Biologia

Controrelatore: Prof. Valentina Gandin
Dipartimento di Scienze del Farmaco

Laureando: Fabiola Pedrini

Anno Accademico 2021/2022

Abstract

During the last decades, in literature, a variety of several types of gene expression signatures for the study of cancer biology have been described. These published signatures cover various aspects of tumor biology related to both cancer and normal cells present in the tumor microenvironment. Most of the proposed signatures lack a computational implementation that is fundamental for their usage and reproducibility. In the attempt of filling this gap of knowledge, during my thesis project, I worked on collecting existing gene expression signatures and providing a tool for their use in genomic data analysis. My contribution in this project has been collected in a new R package, called *signifinder*.

Signifinder offers a unique tool to allow the use and comparison of different signatures within and between samples, also providing utilities for the graphical exploration of results.

Contents

1 Introduction

1.1 The Omic sciences

The Human genome Project (HGP 1990-2003) represents the revolutionary point of break between classical genetics and the advent of the *omics* era with the introduction of genomics approaches. The explosion of *omics* sciences has been possible thanks to the development of both innovative technologies and computational approaches that are fundamental to preserve and analyze the copious amounts of genomic data generated worldwide. From genomic, the expansion of *omics* sciences, including the recent transcriptomics and proteomics, has emphasized that the more significant characteristic in differentiating organisms is the high degree of complexity. Indeed, an organism could be deemed as the result of the articulate and organized interaction between these *omics* [1]. The system biology was introduced as a field with the purpose to study how each *omic* interacts within and among others. Furthermore, system biology has also the purpose to create useful tools for investigating complex mechanisms in all different biological systems. The copious amounts of data derived from different technologies and different *omics* have boosted the implementation of more sophisticated computational methods to understand all biological systems.

Here below a detailed definition of the most common *omics* sciences:

- Genomics consists in the analysis of whole genomes of different organisms. Identification of mutations and variation in the genome allow the collection of valuable information about the cause or the development of numerous diseases [2].
- Transcriptomics is the *omic* science that focus on the study of the transcriptome, also known as the plethora of RNA transcripts, produced transcribing the genome. The expression of genes differs in time and space in response to specific circumstances, or under specific stimuli and in a specific cell type.

- Proteomics investigates the complete set of proteins produced in a cell, tissue, or biological sample. The proteins represent a complicated translation from nucleotides to amino acids. The complexity of the proteome is related to the splicing process and the presence of numerous post-translational modifications.

The *omics* described above need to be analyzed and interpreted through the support of bioinformatics and biostatistics approaches [3].

The focus of my internship and thesis has been on the development of bioinformatic tools to help in dissecting the transcriptomic data complexity.

1.2 The Transcriptomics

RNAs are macromolecules composed of linear chains of nucleotides produced during the transcription, which is the cellular process in which RNAs molecules are generated based on the genomic template [4]. Every cell within a specific organism contains the same genome and same genes. Whereas, different cells show different patterns of gene expressions, indicating that not all the genes are transcriptionally activated [5].

According to the Ensembl genome browser (March 2022), the genome of Homo sapiens is composed of 20465 protein-coding genes and 24849 non-coding genes [https://www.ensembl.org/index.html].

The human precursor mRNAs are processed and spliced into mature forms (mRNA) that are composed of coding regions used during translation, between the 5'-cap and 5'-UTR (untranslated regions), and 3'-UTR and poly-A tail [4]. Furthermore, numerous cellular processes are regulated by other RNAs, including transfer RNA, ribosomal RNA, small nuclear and nucleolar RNA, short interfering RNA, microRNA, and others. Thus, different RNAs perform a wide range of cellular activities.

The aims of transcriptomics are various, including the identification of every type of transcripts such as mRNAs, non-coding RNAs, miRNAs, but also the determination of structure of genes and quantification of levels of gene

expression under different conditions [6]. The identification of all activated genes and their quantification help in increasing knowledge of molecules which have implications in determining the peculiar characteristics of cells both in healthy and diseased tissues.

1.3 Transcriptomics Technologies

The measurement of gene expression is nowadays considered a routine and widespread research field. Indeed, rapid development of technological devices allows innovative approaches commonly used with an improved sensitivity [7].

There are two key techniques for transcriptome study:

- Microarrays, first published in 1995, quantify a set of predetermined sequences via their hybridization to an array of complementary probes [8].
- RNA sequencing (RNA-Seq), first published in 2006, uses high throughput sequencing to capture all sequences of transcript cDNAs [9].

1.3.1 Microarray

Microarrays consist of a solid substrate in which there are linked short nucleotide oligomers, called "probes". The technique includes RNA extraction from biological samples and its copying in cDNA, incorporating either fluorescent nucleotides or tag [10]. The subsequent step is the hybridization between transcripts and the probes that leads to the emission of fluorescence based on the abundances of transcripts linked. The transcript abundances are indicated by the fluorescence intensity at each probe location on the array (Figure 1). In order to generate the probes linked on the array, prior knowledge is required of the organism of interest. Indeed, information about the annotated genome sequence or a library of expressed sequence tags (ESTs) could be used as probes [11].

The process of analyzing the microarray images starts from the correct iden-

tification of the regular grid and quantifies the fluorescence intensity for each spot. The abundance of each sequence is proportional directly to the fluorescence intensity. The next step consists in the conversion of the images to sequence data, and it is typically managed by instrument software [11].

The measurements need to be normalized to make them comparable. With the assumption of the same quantity of RNA used, normalization is needed to adjust problems related, for example, with image acquisitions or the presence of artifacts.

It is important to consider that the microarray techniques present limitations, including the previous knowledge about genomic sequence or high background levels [12].

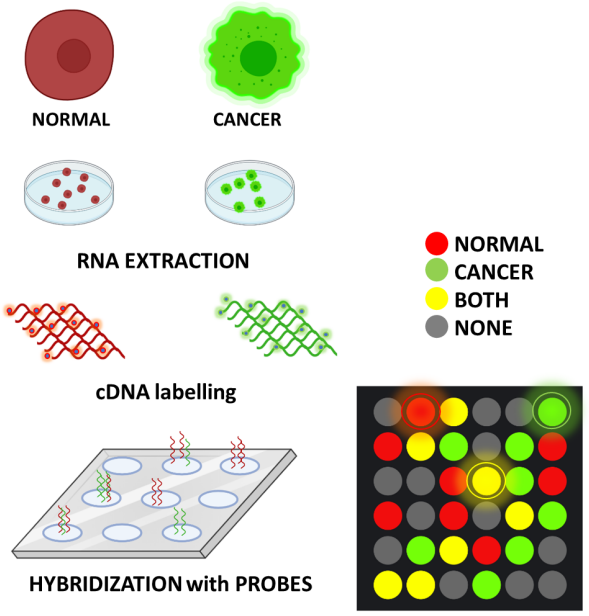

Figure 1: workflow the microarray experiment protocol.

1.3.2 RNA-Sequencing

A new method that consists both in mapping and quantifying transcriptomes is called RNA sequencing (RNA-Seq) [6]. RNA-Seq is an approach to quan-

tify transcripts in RNA extract using high-throughput sequencing methodology and computational methods. In RNA-Seq copious amounts of short transcripts are used to computational reconstruction of original RNA transcript by aligning reads to a reference genome. After the extraction of RNA, the transcripts are first enriched, then performed either conversion to a library of cDNA fragments or RNA fragments followed by amplification. The transcripts could be sequenced in just one direction (single-end) or in both directions (paired-end). The single-end approach is faster and cheaper than paired-end sequencing, however the second approach allows high fidelity alignments [11]. Paired-end sequencing is beneficial for gene annotation and the discovery of transcript isoforms.

In order to obtain sequences of different types of RNA, such as mRNA or rRNA, different methods are developed, especially using peculiar characteristics of every RNAs, like the presence of poly-A tails for mRNA [13].

cDNA transcripts may be amplified by PCR and the eventual use of unique molecular identifiers (UMIs) is widespread to individually tag sequences to create a set of unique tagged fragments. The use of UMIs, indeed, enable correction for amplification bias and allow accurate estimation [14].

The RNA-Seq analysis process is divided into four steps [11]:

- Quality control: the raw data are analyzed for the high-quality scores for base calls or guanine-cytosine content that matches the expected distribution.
- Alignment: is a step to link the expression of a gene to the sequence read abundance. The eukaryotic sequences require specialized handling of intron sequences, which are absent from mature mRNA.
- Quantification: may be performed at diverse levels: the gene, exon, or transcript. It requires probabilistic methods to estimate transcript isoform abundance from short read information.
- Differential expression: performed by normalizing, modeling, and sta-

tistically analyzing the expression data.

Illumina short-read sequencing is the most widespread RNA-Seq technology [13]. RNA-Seq does not require previous knowledge about genomic sequence. Thus, less RNA sample is required in comparison to microarrays. Indeed, the quantity of RNA necessary for RNA-Seq is around 1 ng while for microarray it is required around 1 ug. Moreover, this technique permits a large dynamic range of expression transcripts to be detected. However, RNA-Seq has some limitations including the different bias from types of fragmentation, the construction of libraries with presence of identical shorts reads or the risk to create artifacts during PCR amplification [6].

RNA-Seq allows capturing the gene expression levels from cell cultures or patient tissues characterized by different cell components, called bulk data. Beyond bulk RNA-Seq, the improvements of RNA-Seq allow the measurements of transcriptome at the single-cell level and the more recent spatial transcriptomic technology combines the gene expression measurements with the spatial localization of quasi-single-cell transcriptomes [13].

1.3.3 Data reproducibility and challenges

Bioinformatic is used to manage and analyze the large amount of gene expression data obtained from either microarray or RNA-Seq. The main goal of analyzing and using this type of data in biomedicine is the development of tools towards personalized medicine. In cancer, the data analysis offers the possibility to tailor the therapeutic treatments for patients or to predict the risk of therapy resistance. In the future prospective, indeed, the analysis of patients based on the use of biomarkers will help in the administration of the specific or combined treatment to provide increasing effect and to avoid the absence or adverse effect (Figure 2).

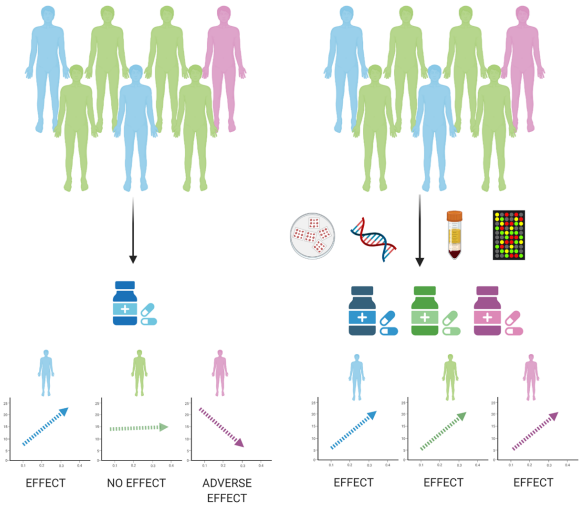

Figure 2: Personalized medicine: future vision.

In 2005, it was established The Cancer Genome Atlas (TCGA), the consortium that included solid and liquid tumor molecular characteristics with the aim to achieve useful patients' stratification toward personalized medicine. At the beginning, TCGA included only few tumors: brain, lung, and ovarian cancer. Subsequently, the expansion of TCGA to all types of human cancers has made It one of the most important databases for tumor *omics* data.

The transcription levels of genes in different samples are collected in a matrix with rows representing the genes and the columns representing the samples, called gene expression matrix. There are several types of comparison that could be done with gene expression data including comparison between genes (rows) and comparison between samples (columns). Generally, a matrix reporting demographic and clinical patient information is associated to each sample, while information such as the chromosome localization, functions and other details are associated to genes. The information related to rows and columns are called meta-data (Figure 3).

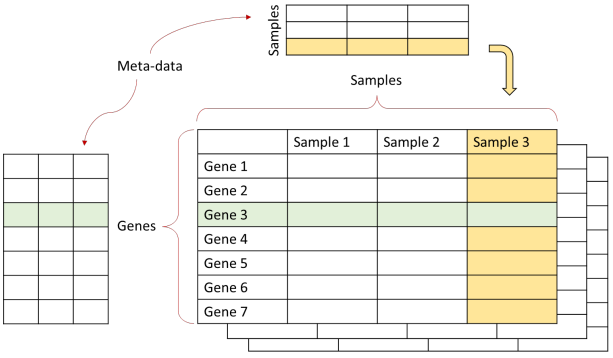

Figure 3: Gene expression matrix.

The processing of the data also provides crucial information, and it depends on the difference in set dimension, gene-expression platform, type of technologies used, annotation and referred to different external databases.

The management and analyses of expression data must follow specific rules defined by the Functional Genomics Data Society (FGED). FGED defines the Minimum Information About a Microarray Experiment (MIAME) or Minimum Information About a high-throughput SEQuening Experiment (MINSEQE) standards that should enhance reproducibility of data analyses.

The critical elements include raw data, processed data, sample annotation, annotation of molecular features and the data processing operation.

Given the above mentioned-characteristics of *omics* data, reproducibility is a crucial aspect also for gene expression data. The cause for result irreproducibility is due to selective reporting of data in most scientific articles, the pressure of publication and the low statistical power or poor analysis.

In data analysis, reproducibility is the condition when, starting from the same raw data and code, the same result is obtained from another analyst using the same analysis method and the same code. Whereas the concept of replicability implies the same conclusion starting from another data analyzed with the same methods [15].

The analysis of gene expression from the raw data to the result can be figu-

ratively represented by the so-called "garden of the forking paths" because of the numerous choices that could be made during the analysis. Indeed, filtering thresholds, the quality of the reads, the algorithm utilized and the method of quantifying the expression can be differentiated by the choice of the analyst [16].

Following reproducibility rules, the publications must provide not only the data but also the code and the meta-data.

1.3.4 The gene expression signatures

The opportunity to store and use the knowledge achieved from gene expression data allows researchers to compare data in several conditions and to clearly establish their properties.

Based on the increased insights, a "gene expression signature" can be defined as the fingerprint of a specific biological aspect.

Usually, a gene expression signature is composed by:

- The list of the genes that compose the signatures.
- A method that evaluates their expression and defines a summary score that represents the activity of the signature.

A gene expression signature is considered a multi-gene biomarker helpful to measure biological alterations in the samples and they can be used for multiple purposes when healthy (such as screening and early diagnosis) and when having diseases (such as prognosis and therapy response) [17].

The process to produce a new signature is composed of different steps: i) the identification of genes to be included in the signature, ii) the selection of the mathematical function that summarizes their expression, iii) estimation of the best parameters such as the cutoff points, which help in the application of the score to samples [18].

In order to obtain robust gene-expression signatures, the fundamental strat-

egy is represented by the division between the step of development and the step of validation [19]. In the development step, the patient selection criteria, the sample size, and the types of analysis should be correctly planned based on the intended use [17]. In the validation step, the new signature should be applied to independent validation samples. The application of the signature must present high specificity, precision, and accuracy. Indeed, five levels of evidence (LEO) were described by the ASCO Tumor Markers Guidelines Committee that should be used in evaluation of a new tumor marker such as a gene expression signature [17].

1.3.5 Gene expression signatures in cancer

Since the publication of the first signature in the late nineties, the study of gene expression signatures has been widely applied to the study of cancers. Indeed, the rise of high-throughput gene expression techniques revolutionized cancer genomics boosting the research towards the identification of a variety of new biomarkers.

This innovative technology is widely used to answer to the fundamental questions related to tumor biology, patient's risk at the time of diagnosis, ways to monitor cancer progression, and selecting the ideal treatment. Signatures respond to these questions for each cancer by providing a deep understanding of its specific characteristics, by predicting the outcome and by giving an insight of the performance of the treatments [19].

Gene expression signatures could be classified in distinct categories: prognostic, predictive and diagnostic signature. A prognostic signature is a clinical or biological characteristic that provides information about the outcome of cancer disease while a predictive signature consists in identifying patients who more benefit a therapy not necessarily related to prognosis [20]. The prognostic score divides, generally, the patients in different groups based on the risk of tumor recurrence, from low risk to moderate and high risk. Whereas the predictive score enables to divide patients that most likely benefit for a specific treatment and who less likely benefit [21]. Lastly, diagnostic

signature is composed by genes that help and simplify the increasing of ability to diagnose a specific clinical condition.

The successfully validated signatures are the ones that can be incorporated into clinical practice in the future. Therefore, the prospective usability in clinical studies ideally represents the aim of every new signature development [17].

Cancer cells have peculiar characteristics (Figure 4) including ability in chronic proliferation with losing control of growth signals that occurs in an acquired ability to support its own proliferation. Furthermore, cancer cells also prevent the negative regulation of cell proliferation, evading growth suppressors and they are resistant to cell death, like apoptosis. Indeed, apoptosis is well known to be the barrier to cancer development, so tumor cells use different strategies to limit or circumvent this and other forms of regulated death. Another crucial characteristic is the enabling of replicative immortality that consists in the ability of cancer cells to have unlimited replication to prevent a limitation in the number of cell divisions [22].

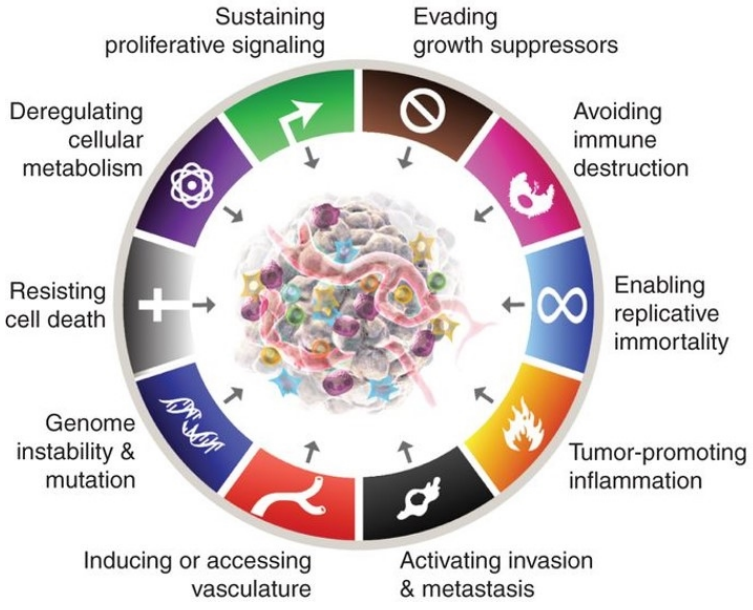

Figure 4: The Hallmarks of cancer [22]

Cancer cells also present a consistent deregulation of cellular metabolism due to support for chronic and uncontrolled cell proliferation [22].

Tumor tissues require sustainability nutrients and oxygen like normal tissues. In order to obtain these fundamental needs, cancer cells are able to induce production of growth factors to stimulate development of new vasculature. This process is called angiogenesis. However, these vasculatures are the result of defective regulation, so the structure of these new vessels presents the same irregularities presented in cancer tissue [23].

Apart from the tumor cells, tumor tissues are also composed by different types of normal cells, infiltrated and not infiltrated in the tumor mass, which compose the tumor microenvironment (TME) together with their extracellular counterpart [24]. It has been demonstrated that TME supports the tumor growth and promotes the immune response escape [25]. Thus, the study of the TME composition is a hot topic of cancer research.

Since the gene expression provided by microarray and RNA-Seq technology capture the cumulative gene expressions of many cells combined, they provide useful data for the study of both the tumor and its TME [26].

The altered relationship between cancer cells and its TME supports the tumor ability to invade other tissues. The process of metastasization implies alterations in cancer cells, such as epithelial to mesenchymal transition (EMT) process through which tumor cells acquire the ability to disseminate [27].

In the TME, a crucial role for cancer formation and progression is also related to the interaction between cancer cells and immune infiltrates. Indeed, immune evasion represents one of the most important cancer hallmarks [28].

In 2020, cancer was the cause of death of over ten million people according to the World Health Organization (WHO) and it was the second worldwide cause of death after cardiovascular diseases.

The most common forms of cancers in 2020 were breast with around 2.26 million of new cases, lung with 2.21 million of cases and colon and rectum cancer with around 1.93 million of new cases. The highest cancer mortality rates were for lung cancer with around 1.80 million deaths and colon and rectum cancer with less than one million deaths [https://www.who.int/].

However, there are types of cancer, like ovarian cancer, which see negligible improvements in therapy options and survival due to the difficulty in identifying the tumor in its initial stages and the lack of treatments due to therapy resistance acquisition [29].

Other cancers, like melanoma can be treated by innovative treatment based on immune checkpoint blockade, however only a subset of patients benefits from durable clinical response, while the others become resistant [30]. For bladder cancer there are no clinical markers to identify the carcinoma in situ lesions that are associated with the high decrease of overall survival [31], or glioblastoma that presents the highest mortality with a 5-year survival rate around 5% [32].

The use of prognostic and predictive signatures in clinical practice would have profound implications for the improvement of cancer treatments and patient quality of life. Although many cancer types still lack of effective personalized tools, there are some examples where the use of signatures are successfully applied to stratify patients. This is the case of breast cancer. This cancer is the most widespread around the world and there are signatures currently used in clinic for prognosis and therapy prediction, like MammaPrint® (Agendia, Amsterdam, The Netherlands) [33], which is able to guide adjuvant treatments in node negative breast cancer, or Oncotype DX, that is able to predict 10-year distant recurrence in estrogen receptor-positive patients and also the response to chemotherapy and endocrine therapy [34].

2 Aim

The high-throughput gene expression techniques, including microarray and RNA sequencing, allowed the identification and development of a set of biomarkers. These biomarkers, called "gene expression signatures", are used to understand cancer biological processes and prognostic features that can characterize the sample. The purpose of my work is to collect signatures from literature and provide their computational implementation.

My contributions helped to develop *signifinder*, an R package that serves as a tool for investigating and exploring cancer transcriptomes.

3 Methods and Results

The main results of my project are twice and correspond to the two main tasks that I performed during my stage: (i) I generated a compendium of gene signatures collecting and screening research papers from literature; and (ii) I contributed to providing the computational implementations of the selected signatures in a new R package named *signifinder*. These two tasks are detailed below.

3.1 Collection of gene expression signatures

Given the translational applicability of the package *signifinder*, we needed a reliable and solid collection of cancer gene expression signatures. Thus, we defined a series of stringent criteria for the signature inclusion.

- Signatures should rely on cancer topics.
- The signature, by definition, should provide a score that helps in characterizing a sample for a specific cancer feature. Thus, given this signature definition, all the included signatures are composed by a gene list and by a method to calculate an expression-based score.
- The selected signatures should be based exclusively on transcriptomic data; exceptions have been made in case of combination of transcriptomic data and survival or histopathological data.
- Signatures must be developed from bulk tumor samples. Thus, since the different biological purposes and data structures, even if extremely rare, we discarded all signatures developed for single-cell transcriptomic data.
- The material and method of the original work must be clear and complete; authors should clearly indicate the type of input data and the set of the considered genes.
- Genes should have official Gene Symbol (Hugo consortium) or an unequivocal translation versus this kind of annotation. Genes without

Gene symbol are removed if they accounted for less than the 5% of the gene signature. Signature with a total amount of untranslatable gene names greater than the 5% were discarded.

During the literature screening, I learnt the importance of providing detailed materials and methods while publishing scientific research. A lot of the screened papers and the discarded signatures have incomplete or un-reproducible methods. Specifically, the type of input data is crucial information often absent in the published work, this strongly affected the signature reproducibility, and it represented the exclusion criteria for most not-included signatures.

After the screening, we select 47 signatures for *signifinder*, 30 of which are tumor type specific (which cover a total of 14 different tumor types) and 17 that can be considered pan-cancer. Signatures are summarized in supplementary Table 1 and described along with their implementations in a dedicated section called "The Cancer Signatures compendium".

3.2 Computational implementations of the signatures

The following section is focused on the description of the signature development in the R language. The choice of the R programming language, as a tool to implement the signatures, is related to the fact that it represents one of the most widespread programming languages normally used for the analysis of gene expression data. It includes the ways and the procedures adopted for data collection and code implementation. Moreover, it includes a more general overview of the tools used for the *signifinder* development and a description of the package that I contributed to develop.

The following steps have been done for every signature collected.

Assessment on the type of input data required

Each signature has its own proper input (defined by the signature proponent and described in the original work). The proper input type has been collected and stored in the signature data. Data transformation procedures have been provided inside the signature implementation to match the type of user data with the type of data needed by the signature.

The possible types of input data retrieved from all the collected signatures are summarized by the below Table 1.

Type of experiment	**Type of data**	**Description**
Microarrays	Normalized gene expression matrix	A microarray normalized gene expression matrix is obtained to compare the level of gene expression between samples.
RNA-sequencing	normalized counts	RNA-Seq normalized gene expression matrix is obtained to compare the level of gene expression between samples.
	FPKM/RPKM	Fragments/Reads per kilobase of exon per million mapped fragments (FPKM) is a within-sample normalization method to compare gene expression levels rescaled to correct for both library size and gene length.
	TPM	Transcript per million (TPM) is a between-sample normalization. The sum of all TPMs in each sample are the same to make them comparable.

Table 1: Table of the types of input data required for the selected signatures

The statistic and its implementation in R

Following the description of the method, we implemented in R the mathematical model that combine the gene expressions profiles providing a final score (reproducing the same results obtained in the original work). Sometimes it happens that more than one signature share the same function, due to a similar formula or a similar cancer topic.

Data preparation for the R environment

When the proposed method for signature calculation is acquired, we need to create a data table in R language containing the collection of the list of genes used to obtain that signature score. Particular attention has been paid to the collection of genes with official gene symbols, avoiding genes that cannot be connected in any way to expression data.

The use of a GIT repository for the collection of signature data and code

Once the signature was implemented, I transferred the data and the relative R code to a shared repository, in order to be stored and made available to the other developers of *signifinder*.

To accomplish this task, we used a GIT repository shared across all the *signifinder* developers. GIT repositories allow the collection of files of different versions of a part or of an entire project, the files imported into this local server present continually updates and modifications. GIT repository presents different important advantages such as the possibility to keep track of every modification, store all different versions of the same files, allow simultaneous changes by different users, and create different versions of a project.

The GIT repository, created for *signifinder* code management, contained the structure of directories required for the R package development.

A brief description of the GIT workflow includes three steps: (I) the modification of a file in local, (II) the selection of the modifications to be part of a commit that can be added and (III) the commit of the files modified from the

local work area to store permanently on the GIT directory.

R package development

A package is the most common way to provide shareable code. A package is constructed with different elements including the functions, the input data, the documentation, and the tests.

In order to provide a tool for a broader public, the development of *signifinder* has been maintained as clear and simple as possible, compatible with the widely used tool for gene expression data analysis.

The input data for each function is the normalized gene expression matrix for both microarray and RNA-Seq data.

The *signifinder* package is composed of several functions that provide signatures' estimation, the R data used by the functions and the tests to control the correct implementation of the method. Moreover, *signifinder* includes accessory tools to help users browsing and visualizing the results.

More in detail, the "SignatureFunction.R" script contains the code of each signature that can be divided in two different parts. The first part is the documentation, which includes the "Description" that gives the fundamental information about the function, the "Arguments" that explain the parameters of the function and their default values, and the "Value" that describes the output of the function where the signature scores are stored. The second part is the effective structure of the function, which manages the input and computes the scores of the signature.

Additional functions are stored in another R script called "UtilityFunction.R" and are used inside the signature functions.

It includes "signatureTable", a table that provides a list of information for each signature: the function name in which the signature was implemented; the tumor tissue on which the signature was built; the macro-category of the biological process involved; the reference of the original work; and the type of data input.

The testing of the correct implementation and functionality of each function is constantly monitored by the test scripts.

3.3 The Cancer Signature compendium

The topics covered by signatures collected in *signifinder* are different and are divided in 16 cancer topics:

- Epithelial-to-mesenchymal transition signatures
- Pyroptosis signatures
- Ferroptosis signatures
- Hypoxia signature
- Immune System signatures
- Cancer stem cell signatures
- Chromosomal instability signatures
- Extracellular matrix signatures
- Carcinoma in situ signature
- Angiogenesis signature
- Altered metabolism signatures
- Mitotic index
- Autophagy signatures
- Cell cycle signatures
- Cancer molecular subtypes
- Platinum resistance signatures

Signatures details have been provided in supplementary Table 1 and are also described below.

3.3.1 Epithelial-to-mesenchymal transition signatures

The process of epithelial to mesenchymal transition (EMT) is the process in which cells lose epithelial characteristics and are transformed into motile mesenchymal cells [35]. EMT is strongly related with dissemination, invasion and drug resistance in various cancers [36].

The *EMTSign* is the function dedicated to the study of the EMT process and it allows the possibility to calculate three different signatures for ovarian cancer, breast cancer and a pan-cancer analysis. These signatures are based on Miow *et al.* [37], Cheng *et al.* [38] and Mak *et al.* [39] studies.

The EMT score proposed by Miow *et al.* is based on an enrichment score able to establish the epithelial and the mesenchymal-like status in ovarian cancer samples. The scores give an indication of which status is predominant in the sample.

In breast cancer, the EMT score proposed by Cheng *et al.* investigates the modulation of late recurrence. The breast samples presenting high EMTscore are related with high likelihood of developing a late recurrent disease and poor prognosis.

Lastly, the EMT signature proposed by Mak *et al.* is constructed as a pan-cancer score and it provides a quantification of the level of epithelial (EMTscore < 0) or mesenchymal (EMTscore > 0) status in the cancer sample.

3.3.2 Pyroptosis signatures

Pyroptosis is a form of programmed cell death that starts with the formation of numerous vesicles, followed by pores formed on the cell membrane, which results in contents flowing out [40]. These events involve activation and release of a variety of danger-associated signalling molecules and cytokines, accompanied by immune system activation and strong inflammatory response [41]. Pyroptosis' role is controversial in cancer tissue, in which it

seems to promote but also inhibit the tumor development. However, it is found to be related with different forms of cancer and it seems to contribute to delivering nutrients and accelerating cancer progression in late stages [42].

The *pyroptosisSign* is the function dedicated to the pyroptosis signatures. *pyroptosisSign* includes four signatures based on Ye et al. [43], Shao *et al.* [44], Lin *et al.* [45] and Li *et al.* [46] studies. It offers a tool for investigating the overall survival of patients with ovarian, gastric, lung and glioblastoma cancer, respectively. The signatures are constructed by using pyroptosis-related genes selected for their association with survival. Thus, the final goal of these signatures is to indicate the overall survival of samples dividing them based on the resulting pyroptosis scores.

3.3.3 Ferroptosis signatures

Ferroptosis is an iron-dependent and non-apoptotic form of death involved in various diseases including cancers [47]. During ferroptosis, the cells show peculiar characteristics including the intact cell membranes with a normal nucleus size and no chromatin condensation, a reduced mitochondrial volume and increased mitochondrial membrane density [48]. In order to overcome the problem of developing resistance, one possible treatment is to stimulate ferroptosis [49], however the role of this form of death is not completely elucidated in tumor suppression.

The *ferroptosisSign* is the function dedicated to the ferroptosis signatures. It includes four signatures based on Ye *et al.* [43], Liang *et al.* [50], Liu *et al.* [51] and Li *et al.* [52] studies. It offers a tool for investigating the overall survival of patients with ovarian, hepatocellular carcinoma, prostate, and oral squamous cell carcinoma cancer, respectively. The score signatures are estimated by analyzing the expression of ferroptosis genes associated with survival. The goal of these signatures is to provide a risk score able to predict the overall survival of patients dividing them into high and low risk of relapse.

3.3.4 Hypoxia signature

Low level of oxygen is generally found in every tumor tissue during the growth, and it represents the mechanism through which the cancer stimulates the growth of new vasculature needed for nutrient supply and dissemination [53][54][55].

The *hypoxiaSign* is the function dedicated to the study of hypoxia. It is based on the signature proposed by Buffa *et al.* [56] and offers an important tool to investigate the hypoxia levels in tumor tissue samples. The output score has been demonstrated to also have a prognostic significance. In particular, the patients with higher scores have the higher level of hypoxia and present a poor-prognosis tumors.

3.3.5 Immune System signatures

The macro category called Immune System includes different signatures that cover the various cancer aspects related to the involvement of immune system in cancer development.

- *immunoScoreSign*
- *chemokineSign*
- *immuneCytSign*
- *IFNSign*
- *expandedImmuneSign*
- *TinflamSign*
- *IPSSign*
- *PassONSign*
- *IPRESSign*
- *IPSOVSign*

- *TLSSign*

The *immunoScoreSign* is a function that includes two different immune scores proposed by Hao *et al.* [57] and Roh *et al.* [58] for the analysis of epithelial ovarian cancer and pan-cancer, respectively.

The Immune signature proposed by Roh *et al.* is based on 41 immune-related genes selected for melanoma. This score is composed of genes whose expression is associated to immune activation in the tumor microenvironment and can be used to evaluate the activation state of the immune system in cancer.

In order to elucidate the relationship between immune activity and cancer genotype in ovarian cancer, Hao *et al.* developed an immune score able to estimate the immune status of a sample. The final score is based on 76 favorable prognostic genes related to specific tumor-infiltrating immune cell types. Therefore, this score is highly reflective of pre-existing antitumor immunity and should be a strong prognostic signature in EOC (epithelial ovarian cancer).

The *chemokineSign* is the function based on the score proposed by Messina *et al.* [59]. This score is based on 12 chemokine genes to investigate melanoma samples, especially for the presence of lymphoid structures within the tumor. Patients with high scores show the presence of lymphoid cell infiltrates in melanoma metastasis which is also associated to a better outcome.

The *immuneCytSign* is based on two signatures proposed by Rooney *et al.* [60] and Davoli *et al.* [61] both for a pan-cancer analysis of the local immune cytolytic activity.

The cytolytic activity of the local immune infiltrate in solid tumors includes the activation of different types of cells with the ability to kill tumor cells and promote favorable outcomes. Starting from the analysis of key cytolytic effectors, Rooney *et al.* [60] developed a score that is able to quantify the level of this activity: the CYT score. CYT score can be used as a proxy of

survival, in fact high CYT score shows significant pan-cancer survival benefits. It has been demonstrated by authors that high levels of CYT bring the tumor to a condition of pressure in which subclones with resistant mutations expand due to the acquired ability to evade and suppress CYT.

Another immune system related score has been proposed by Davoli *et al.* [61] and it is based on the expression of a set of genes, which are considered molecular markers of cytotoxic $CD8^+$T cells and NK cells. Interestingly, samples that show high scores for this signature are also characterised by somatic copy number alterations.

The *IFNSign*, *expandedImmuneSign* and *TinflamSign* are based on the paper published by Ayern *et al.* [62] in which patients undergo treatment with Pembrolizumab in clinical trials across multiple cancer types. IFN-γ score is based on genes related to IFN-γ, while the expanded immune score includes cytolytic activity, pro-inflammatory cytokines/chemokines, T cell markers, NK cell activity, antigen presentation and T cell checkpoints. Both these scores are higher in responders than in non-responders to Pembrolizumab. The third signature is dedicated to the study of a T cell–inflamed phenotype and it is proven necessary for the clinical activity of PD-1–/PD-L1–directed monoclonal antibodies. Thus, the Tinflam score, which is derived by the expression of genes representing the inflamed T cell, also predicts the response to Pembrolizumab across multiple solid tumors.

Inhibition of tumor-mediated suppression of anticancer immune responses is the focus of recent treatments called immune-checkpoint blockade (ICB). The purpose of these treatments is to redirect the cytotoxicity of immune cells on tumor cells. ICBs are a class of treatments composed of numerous types of monoclonal antibodies to the receptor cytotoxic T-lymphocyte antigen-4 (CTLA-4) and programmed cell death protein 1 (PD-1), both expressed on T cells; or the PD-1 ligand (PD-L1), which is expressed by a variety of cell types, including some tumor cells [63] [64]. However, despite the promising results in preclinical studies, only few patients benefit from durable clinical responses from ICB therapies. The cancers that especially

benefit from ICB treatment include melanoma, small cell lung cancer [65] and colorectal cancer [66].

The *PassONSign* is based on the score proposed by Du *et al.* [67] and offers a tool to separate anti-PD1 responders and non-responders. The ability of this score is both in the prediction of a patient's clinical response to anti-PD1 therapies and in the identification of patients with better survival outcomes.

The *IPRESSign*, based on Hugo *et al.* [30] signature, allows the calculation of a predictive score able to distinguish between responders and non-responder's melanoma patients to anti-PD-1 treatment. PD-1 immune checkpoint blockade therapy induces a response, especially in melanoma. However, a high rate of innate resistance (60%–70%) in advanced metastatic tissues impacts the effective clinical use.

In ovarian cancer, specific alterations of the immune system are also prognostic [68]. The *IPSOVSign*, based on the signature proposed by Shen *et al.* [69] allows the investigation of the patient's prognosis in OV cancers based on immune system genes. IPSOV score helps in dividing patients in high and low risk patients based on their altered immune gene expressions.

The tertiary lymphoid structures (TLS) are the localization of B cells, a crucial component of the adaptive immune system that could increase the antigen presentation and the release of tumor-specific antibodies. The presence of TLS and B cells are associated with improved prognosis since they could recognize the tumor antigens and support the activation of $CD8^+$T cells against tumor cells. The TLSsign is a function based on the signature by Cabrita *et al.* [70] that studies the presence of TLS in melanoma patients. The genes used by this signature are considered TLS-hallmark genes and the signature is able to predict the overall survival of metastatic melanoma samples. In particular, high scores are related to better prognosis indicating the role of immune activation.

Finally, Charoentong *et al.* [71] proposed an approach based on the use of non-overlapping sets of genes representative for specific immune cell sub-

populations, by defining a set of pan-cancer metagenes for 28 immune cell subpopulations. The *IPSSign* is based on the immunophenoscore (IPS) proposed by the authors, it calculates four separate scores specific of different cell subpopulations: the effector cells (EC), represented by infiltration of activated CD8+/CD4+ T cells and Tem CD8+/CD4+ cells; the immunosuppressive cells (SC) composed by Tregs and MDSC; the MHC molecules (MHC) including MHC class I, class II and non-classical molecules; and the expression of certain co-inhibitory and co-stimulatory molecules called checkpoint/immunomodulators (CP). Then, it also provides an aggregate score, called IPS, comprehensive of the four categories described above. The IPS was found to be associated with the patient's survival in 12 solid cancers and it also has predictive power for identifying responders for treatment with CTLA-4 and PD-1 antibodies.

3.3.6 Cancer stem cell signatures

The stem cell phenotype is acquired by cancer cells during the progression toward a more aggressive phenotype. The stem cell phenotype acquisition and maintenance are strongly managed by microenvironmental cues and cellular crosstalk [72]. The dedifferentiation versus a stem cell phenotype is a widespread event in certain epithelial cancers especially when they become more aggressive [73] [74]. The *ASCSign* is based on the score proposed by Smith *et al.* [74] and it is composed of epithelial Adult Stem Cell (ASC) genes. The ASC score increases during cancer progression from the early to the advanced and metastatic disease. The score is also associated with overall survival of patients in different cancer types. Higher levels of this score are found in patients with the worst outcome and low overall survival compared to patients with low scores. Moreover, the score is associated with genomic alteration, including amplifications in oncogenes, such as TERT, and deletion of tumor suppressors. The ASC score is also found correlated with the gene expression of DNA methyltransferases especially in prostate and lung cancers.

In some cancers, such as colorectal cancer, the loss in regulation of the tissue homeostasis also involves the adult stem cells.

In intestinal crypts the presence of intestinal stem cells is essential for intestinal tissue regeneration [75]. Indeed, in the basis of crypts, intestinal stem cells (ISC), identified by the presence on the cell surface of the Lgr5 protein, proliferate continuously. The expression of another cell surface marker, the receptor tyrosine kinase EphB2, decreases from the crypt base toward the differentiated cell compartment [76]. The cells that derived from ISC are called transient amplifying (TA) cells, they undergo cell-cycle arrest and terminal differentiation close to the intestinal lumen [73]. The evidence suggests a role of ISC regulation especially in colorectal cancer recurrences. *ISCSign* function allows the study of ISC behaviour. Based on work by Merlos-Suárez *et al.* [73] the ISC scores are composed of four gene sets related to the expression of different intracellular and/or superficial markers. The four gene lists are representative of the presence inside the tumor of ISC cells (Ephb2 and Lgr5), late TA cells (lateTA) and proliferation cells (prolif). It has been shown that samples that present high levels of the EphB2- and Lgr5 scores have also high risk of relapse. While the proliferation signature is inversely associated with the risk of relapse and the Late TA shows no association.

Also in prostate cancer, the presence of stem cells has a role to identify cancers with a more aggressive phenotype. In order to initiate secondary tumor growth, the prostate cancer cells seem to acquire stem cell characteristics [77]. The *stemCellCD49fSign* is the function based on the score proposed by Smith *et al.* [78] related to prostate stem cells. Integrinα6, also known as CD49f, is among the proteins that have been identified in stem cell populations and it is used as a stem cell marker. The score is high especially in samples that present the most aggressive and metastatic cancer.

3.3.7 Chromosomal instability signatures

Chromosomal Instability (CIN) is known to be one of the most widespread characteristics of human tumors and it results from errors in chromosome segregation during mitosis. In cancer cells often co-occur the presence of CIN and aneuploidy events, which is the abnormal or non-diploid chromosome number [79]. It is widespread in 60%-80% of cancers and it can be induced by various events that include oncogenic signalling or defects in centrosome replication [80].

It is known that the presence of CI is proportional to tumor stage, recurrences and higher in metastatic cancer [81].

The *CINSign* function allows us to estimate the two CIN scores proposed by Carter *et al.* [82] for the quantification of the level of chromosomal instability in different cancer types. The 25-CIN is composed of the top 25 most important genes considered key regulators for the maintenance of a faithful replication and segregation of chromosomes, while the 70-CIN includes the top 70. Both scores are related to high instability and poor clinical outcomes in multiple cancer samples. Thus, the use of this score can help patients' stratification into two groups defined by high or low presence of chromosomal instability. Furthermore, CIN scores are higher in metastatic foci in different solid tumors, suggesting that they could also indicate the more aggressive phenotypes.

The genomic mutations are accelerated by the deficiency in DNA repair systems, especially for homologous recombination (HR) deficiency, in different cancer types including ovarian cancer [83]. Ovarian cancer cells carrying deficiencies in BRCA1 or BRCA2 genes have also altered ability in repairing DNA double-strand breaks (DSBs) during HR. Moreover, these altered cells are more susceptible to increased chromosomal damage under platinum-based chemotherapy [84]. The high levels of somatic mutations in ovarian cancer can be related to different levels of HR deficiency [85]. The *HDRSsign* is the function to calculate the HRD score proposed by Lu *et*

al. [86]. The score is based on the expression of genes correlated with high mutation rate. The patients are divided in two groups according to their HRD score: those with HRD score > 0 and those with HRD score < 0. The group with high scores (HRDS > 0) has a better outcome, presents longer progression free survival, and achieves complete response. The HRD score is also high in tumors with BRCA1/2 mutations, or epigenetically silenced or with deficiencies in genes related to HR. Valuation in BRCA1/2 mutation shows BRCA-deficient patients in the low-HRDS group were significantly lower than those of BRCA-deficient patients in the high-HRDS group.

Ovarian cancer samples following platinum-based chemotherapy also present alteration in genes involved in repair of these damages. The *DNArepSign* is the function based on the score proposed by Kang et al. [87]. The signature is constructed with genes related to DNA repair pathway, including nucleotide excision repair (NER), ataxia-telangiectasia mutated (ATM) and homologous recombination (HR). This signature offers a specific prognostic score to divide patients who have different outcomes based on differential expression of genes involved in repairing platinum-induced DNA damage. The patients with high scores present also a high complete response to platinum-based chemotherapy. Moreover, high scores are related to better outcomes and lower likelihood to have relapse and to die.

3.3.8 Extracellular matrix signatures

The extracellular matrix (ECM) is composed of cells, such as fibroblasts, endothelial cells, pericytes, macrophages, lymphocytes, and other immune cells, as well as an acellular part. Indeed, the variation of soluble factors in ECM is related to the type, stage, and location of the cancer. The tumor tissue is composed by interaction between stroma cells and cancer cells in a dynamic context [88]. In cancer, the remodelling of the extracellular matrix is mainly due to the activity of fibroblasts, especially cancer-associated fibroblasts (CAFs), and it gives the ability to disseminate, invade and colonize other tissues. The dysregulation of ECM is typically altered in cancer

where the presence of molecules produced by both cancer cells and CAFs increase the acquisition of more aggressive phenotypes [89].

The *ECMSign* is based on the score proposed by Chakravarthy *et al.* [90] and it gives two scores (ECM_up and ECM_down) both related to the outcome and to the response to immune checkpoint blockade at pan-tissue level. Higher the ECM_up score, lower the ECM_down score, poorer the patient prognosis and shorter the overall survival. The scores are also inversely correlated with tumor purity and ECM_up directly correlated with CAFs presence and TGF-β activation that is known to be capable of stimulating fibrosis, inducing EMT and driving metastasis [91].

The matrisome is constituted by the proteins present in the extracellular matrix of different cancer types. The *matrisomeSign* is based on the signature proposed by Yuzhalin *et al.* [92] and is composed of extracellular-matrix related genes. These genes are considerably overexpressed in different cancer types and their expression is associated with cancer progression. The score is found to be higher in patients that present lower overall survival and low disease-free survival. The score shows, also, an association with poor prognosis, like EMT, hypoxia and inflammation.

3.3.9 Carcinoma in situ signature

The *CISSign* is a function based on the score proposed by Robertson *et al.* [93]. The presence of carcinoma in situ (CIS) lesions in the urinary bladder is associated with a high risk of disease progression to a muscle invasive stage and CIS are also considered the precursors of invasive carcinomas [94]. Indeed, the score is applied on gene expression analyses of samples with urothelial bladder cancer. The luminal-papillary subtypes of bladder cancer present low CIS score and high overall survival, while a high score is found in basal-squamous subtypes.

3.3.10 Angiogenesis signature

Angiogenesis consists in the formation of new vessels from pre-existing ones; the reason underlying the activation of angiogenesis relates to low oxygen and nutrients levels during cancer growth. Cancer cells are able to induce production of growth factors to stimulate development of new vasculature that presents the same irregularities presented in cancer tissue [23]. Closely related with angiogenesis, there is the metastasization process. The metastases are the result from the spread of cells from the primary tumor through the blood or lymphatic system, and they represent the principal cause of cancer-treatment failure [95]. Cancer cells invade distant sites and form secondary tumors related to both cancer cells and signals from the microenvironment.

The *VEGFSign* is based on the score proposed by Hu *et al.* [96]. It is composed of cell-intrinsic and cell-extrinsic factors for breast cancer analysis. Indeed, a low expression of fibroblast/mesenchymal genes and a high expression of the VEGF profile are the distinct characteristics of breast distant metastasis. This score can be used to determine the relapse-free survival and overall survival in breast cancer metastatic patients. It also gives information about hypoxic conditions of the cancer tissues and correlates with the regulation of angiogenesis's factors (HIF1α). High levels of this signature suggest the ability in vessel-promotion, metastatic spread, living in under anaerobic conditions and loss of fibroblast dependence.

3.3.11 Altered metabolism signatures

One key aspect in cancer development is the reprogramming of cellular energy metabolism that is crucial to support continuous cell growth and proliferation [22]. Indeed, uncontrolled cell proliferation needs implementation of energy metabolism. Cancer cells disrupt their glucose metabolism by limiting their energy metabolism to glycolysis even in presence of oxygen [97] [98]. These alterations affect all aspects of energy metabolism leading to an increase of glucose [99] and they are associated with different types of

cancer including lung adenocarcinoma, breast, bladder, and renal cell carcinoma.

The function *glycolysisSign* is composed of two scores. The first score is proposed by Zhang *et al.* [100] as a prognostic score for patients with lung adenocarcinoma, while the second score for renal cell carcinoma patients is based on work by Xu *et al.* [101].

The score by Zhang is associated to the metastasis formation and overall survival in patients with lung adenocarcinoma (LUAD). The patients that have high-risk scores showed also higher mortality rates than those with low-risk scores. Indeed, the increase in glycolysis is also linked to the metastatic dissemination in LUAD [102].

In renal cell carcinoma, the patients with high score based on Xu *et al.* [101] have poor overall survival and a deep alteration in the immune microenvironment with higher level of T cells regulatory and lower level of macrophages M2 and dendritic cells. The mutation frequency of the genes by which the score is constructed is not high, suggesting an alteration in their post-transcriptional regulations or translation modifications.

The reprogramming and alteration of lipid metabolism are widespread marks of cancer. In particular, the increase of lipid uptake, it's storage and its genesis is observed in different cancers, contributing to a rapid growth [103] and making cancer cells more independent from external supplies. Moreover, lipid metabolism alteration has an important role in cancer cell dissemination and metastasis formation, indeed lipids also have a fundamental role as signalling molecules, crucial for mediate transformation and tumor growth [104].

The *lipidMetabolismSign* proposed by Zheng *et al.* [105] is based on lipid-metabolism related genes and is proposed to be a prognostic score of epithelial ovarian cancer (OV). The score could be used to stratify OV patients in two groups with different overall survivals: a high score is related to poor overall survival, especially for patients with cancer at FIGO stage III and IV.

In ovarian cancer cells, the lipid-metabolism alteration contributes to poor prognosis, cancer metastasis and stemness [106].

3.3.12 Mitotic index

Cancer cells are characterized by alteration of cell division rates, the mitotic index score obtained from *mitoticIndexSign* offers a tool to investigate the fraction of dividing cells in a tissue [107]. The mitotic index score is based on the work of Yang *et al.* [108] is strongly correlated to extrinsic factors which modulate the cumulative total number of divisions incurred per stem cell in each sample (TNSC) in cancer tissue. Indeed, the mitotic score could also help in the prediction of normal/cancer status.

3.3.13 Autophagy signatures

The cells under a condition of stress - such as nutrient deficiency - respond with activation of the autophagy process [109]. During autophagy, the cell destroys cellular organelles, such as ribosomes and mitochondria, in order to obtain catabolites used for biosynthesis and energy metabolism. The role of autophagy in cancer is not clear, some studies indicate that induction of autophagy can be a barrier to tumorigenesis that may operate independently or in concert with apoptosis [110]. Moreover, some stress situations can induce elevated levels of autophagy that can be cytoprotective for cancer cells [111]. Recent studies reveal that autophagy can prevent or delay tumor formation during the initial phase, but that autophagy can promote tumor progression and protect cancer cells once tumors are formed [111]. Autophagy has an important role in biological function in different human tumor types including clear cell renal cell carcinoma (ccRCC), ovarian cancer (OV) and glioblastoma (GBM).

The *autophagySign* is based on four different prognostic scores based on autophagy-related genes for prediction of overall survival (OS). The scores are based on works of Xu *et al.* [112], Chen M. *et al.* [113], Wang *et al.* [114], Chen H. *et al.* [115] that have proposed signatures for glioma, clear

cell renal carcinoma, glioblastoma, and cervical cancer respectively. Patients with high scores are characterised by the worst overall survival and prognosis.

The signature proposed by Chen for ccRCC provides two different scores, the first is specific for the overall survival (OS) while the second is specific for the disease-free survival (DFS). For both the ccRCC scores, the patients with high values present the worst OS. Further, the value of score from DFS provides additional information about the time of DFS: the patients with high score have a shorter DFS time than patients with low score.

3.3.14 Cell cycle signatures

A fundamental characteristic of cancer cells is their ability to sustain chronic proliferation [22]. Cell cycle is composed of four distinct phases: G1, S, G2 and M ruled by different proteins, such as cyclin family and their interactors the cyclin dependent kinases (CDK) [116]. During cell division, the accumulation and propagation of genetic errors must be prevented, and the cell cycle presents a tight regulation to avoid it. [117]. The cell cycle checkpoint could avoid progression of the cells in division and the death in presence of irreparable DNA damage [117].

Dysregulation of cell cycle is known to be one of the causes of cancer cell proliferation [118]. Indeed, recent works indicate that the ability to exit the cell cycle is the major compromised characteristic in cancer [117].

The *cellCycleSign* is the function dedicated to study the cell cycle, it includes two scores based on Lundberg *et al.* [119] and Davoli *et al.* [61]. Both these scores are constructed to be used in pan-cancer analysis. The score proposed by Lundberg *et al.* (the CCS score) is based on 463 cell-cycle related genes from three different databases (KEGG, HGNC, Cyclebase). The CCS score, in a pan-cancer analysis, showed a significant score association with the Progression Free Interval (PFI). Moreover, TP53 and PIK3CA, a well-known oncogenes, are found to be increasingly mutated with the increase

of CCS score.

The score proposed by Davoli *et al.* is composed of the most important hallmarks of cancer cell cycle. The genes component of this signature are markers of cell cycle regulation and cellular proliferation. This score is positively correlated with tumors with high levels of somatic copy number alterations (SCNAs). Moreover, high levels of SCNA levels in samples correlates with low overall survival of patients.

3.3.15 Tumor subtypes

The majority of ovarian carcinoma are high-grade serous histotypes (HGSOCs). They are frequently diagnosed as late-stage with poor survival and with few targeted therapies available.

The *consensusOVSign* is based on the consensusOV developed by Chen *et al.* [120]. The consensusOV is a classifier based on the consensus of multiple approaches that provide a standardized method for clinical application of HGSOCs classification. The classification methods are based on the work of Helland *et al.* (PLoS One, 2011), Verhaak *et al.* (J Clin Invest, 2013), and Konecny *et al.* (J Natl Cancer Inst, 2014). The consensuOV classification provides a consensus between the three methods and facilitates the ovarian tumors' categorization into well-defined subtypes.

The classifier divides samples into four groups: immunoreactive, differentiated, proliferative and mesenchymal.

Moreover, the subtyping classification divides patients into groups that have different overall survival with the immunoreactive subtype with high overall survival and mesenchymal subtype with low overall survival.

3.3.16 Platinum resistance signatures

A feature of cancer cells is the innate or acquired ability to evade the effects of chemotherapies and become resistant to treatments. The chemotherapeutic resistance is related to the host and the tumor factors and is most

related during the invasion and metastasization of cancer.

Ovarian cancer (OV) is characterized to have high mortality due to the development of resistance to current chemotherapy regimens. The treatment for patients with OV depends on the stage at the moment of diagnosis. Indeed, patients with stage IIIC and small metastasis have better survival if treated with primary debulking surgery [121], while in advanced disease better survival is achieved with chemotherapy with platinum agent in combination with a taxane. However, the majority of patients (75%) become resistant, although initially they respond [122]. Indeed, the 5-year survival is below 45% in OV patients with advanced disease due to drug resistance and the lack of alternatives for the treatment [123]. The chemotherapies used for OV treatment consist of different platinum drugs, including carboplatin or cisplatin and paclitaxel.

In literature, there are several works showing gene expression alterations in platinum resistant ovarian cancer. More in detail, a work by Sherman-Baust *et al.* identifies the differential expressed genes (DEGs) involved in drug resistance especially for cisplatin and paclitaxel [124], a work by Cheng *et al.* [125] identified the DEGs involved in carboplatin resistant samples and a work by Patch *et al.* [126] published in "Nature" provides the list of DEGs between the resistant and sensitive tumors to platinum treatments. Based on these different gene-lists, the *chemoresSign* offers a tool for investigating the presence of resistance genes. The function is based on the work by Winterhoff B.J. *et al.* [127] that describes a method to use DEG-resistant genes to obtain scores.

3.4 Signifinder

Signifinder is the R package that collects the compendia of signatures and their implementations. It enables the users to obtain a single-sample score from every signature with only the submission of a gene expression data matrix from microarray or RNA-sequencing. Moreover, the possibility to use different IDs of the gene-set of the signature is essential to allow and simplify

the use of the package starting from different sets of gene identifiers.

In this perspective, the users can simply submit their gene expression matrix with only the indication of the type of data (either RNA-Seq or microarray) and the gene ID used, without any other details. Users can select a single signature or a combination of the implemented signature.

The output of a signature function is generally a single and/or multiple scores, for which the importance and a brief interpretation is provided in the help function.

Although *signifinder* provides additional functions to better interpret and visualize the results. The analysis of *signifinder* allows not only to explore signatures independently, but also to compare signature relationships in terms of correlations and patients' stratification.

3.5 A case study on Ovarian Cancer using *signifinder*

As an example, I downloaded the ovarian cancer (OV) gene expression data from TCGA. I collected 304 patients, and I used *signifinder* to provide a signature results for this dataset.

In the following part, I reported the different types of plots produced by *signifinder*. The plots were produced using TCGA ovarian cancer data and demonstrate the potential of additional functions in *signifinder*. The focus of these plots is, therefore, purely illustrative, the interpretation of the results goes beyond the work of this thesis.

The first plot (Figure 5) available in the package could be obtained with the function "oneSignPlot".

It allows the user to see the distribution of the scores obtained from a single signature: a scatterplot on the left shows the ordered score values while on the right, we have a histogram and density distribution of the scores.

In both scores are reported the red dashed lines indicating the quantile values.

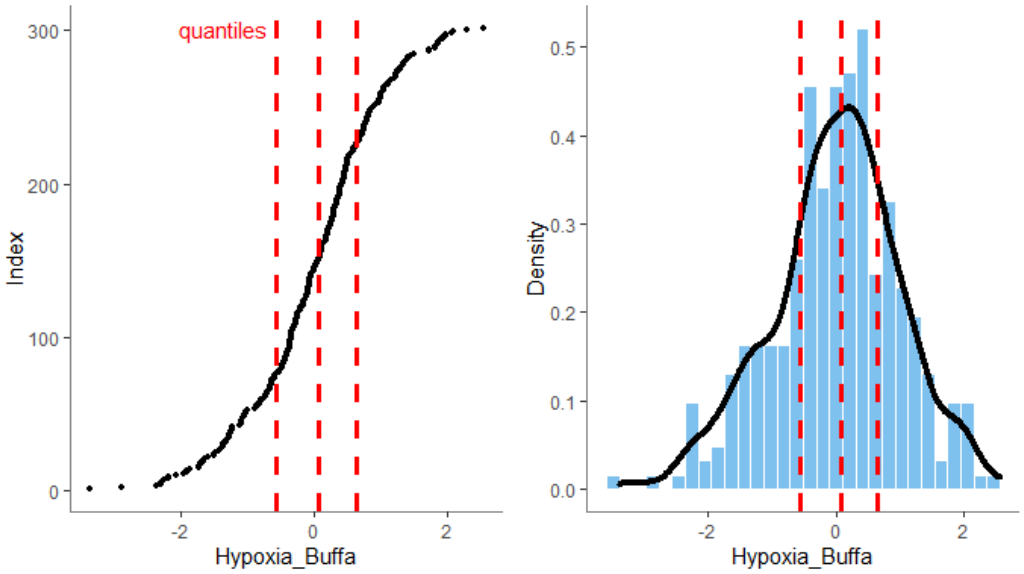

Figure 5: Plot from "oneSignPlot" function. The plot reported is an example produced with *HypoxiaSign*.

Through the function "survivalSignPlot", *signifinder* also offers the possibility to analyze the patient's survival according to the signature values, establishing if the scores can be considered prognostic.

Considering the information on the follow-up of patients it provides a typical plot for survival analysis (Kaplan-Mayer curves Figure 6).

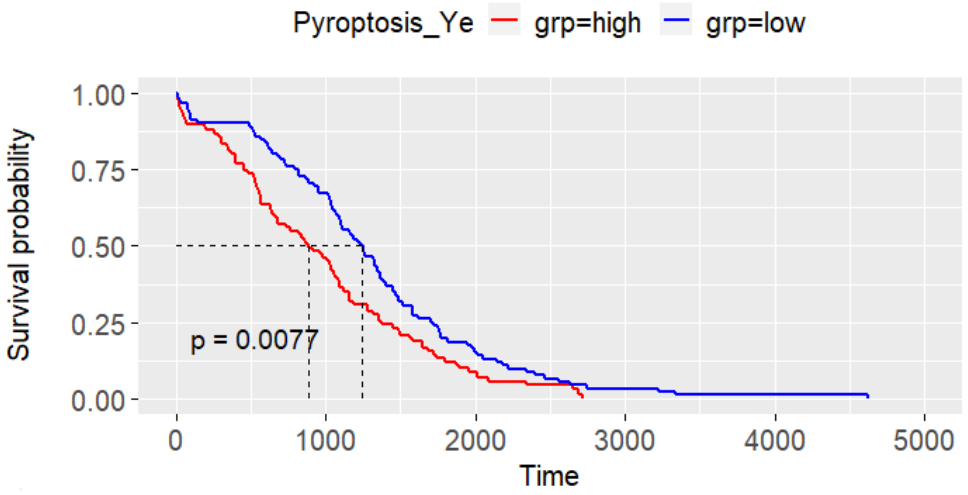

Figure 6: Plot from "survivalSignPlot" function. The Kaplan-Meyer reported is an example of survival analysis of *PyroptosisSign*.

It is also possible to directly compare the distribution of the scores between different signatures with the use of "ridgelineSignPlot" function. The ridge plot allows the analysis between samples divided based on different categories: a real example has been provided using *consensusOVSign* (four scores with different colors are plotted), and it is possible to see the distribution of single and/or multiple signatures (present in the y-axis) (Figure 7).

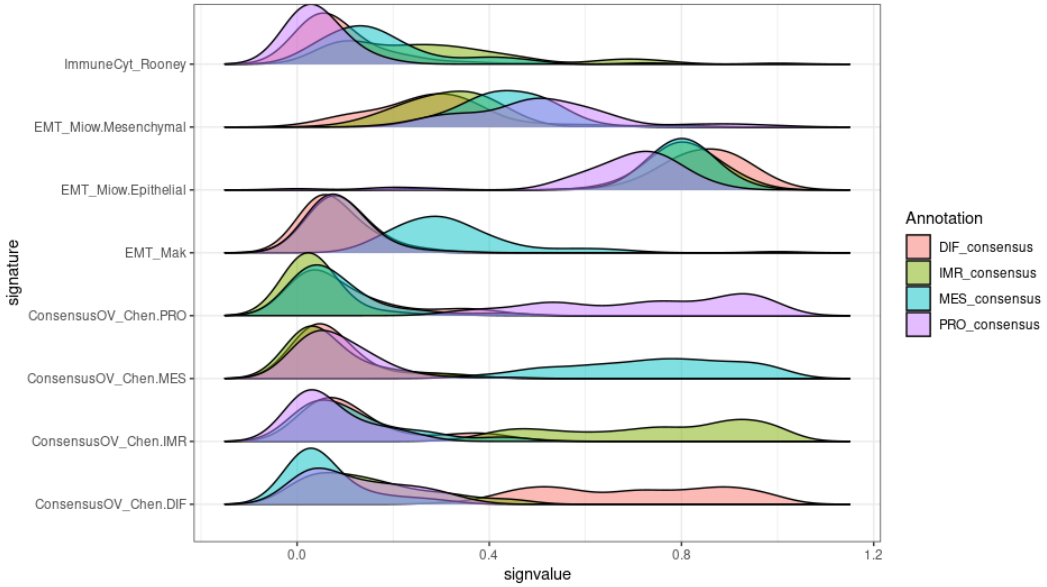

Figure 7: Plot from "ridgelineSignPlot" function. The plot reported is an example compared different signature scores distribution in the classification groups from *consensusOVSign*.

The "correlationSignPlot" allows users to analyze the relationship between different signature functions. The plot in Figure 8, indicates the positive (red) or negative (green) correlation between the different signatures.

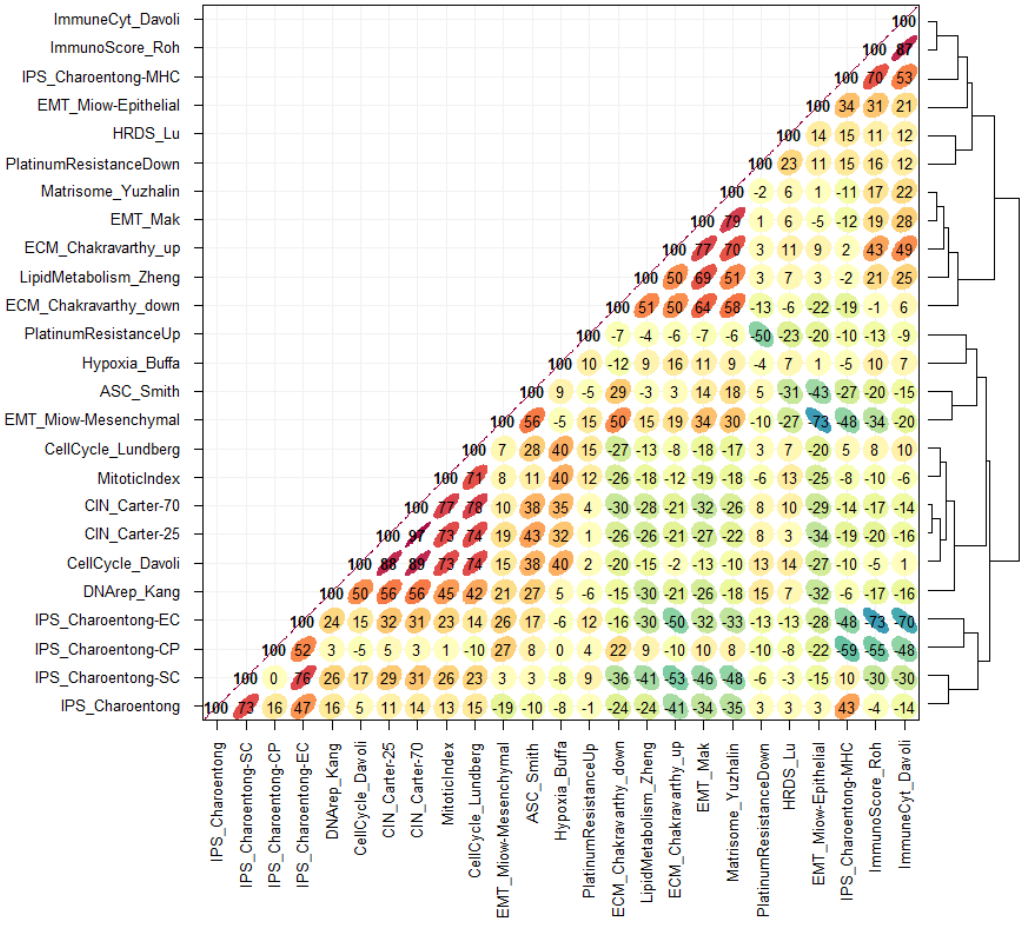

Figure 8: Plot from "correlationSignPlot" function. The plot reported is an example produced with different signatures.

This plot can give a wide perspective of the altered biological functions in a sample cohort, giving an idea of signatures' coherence.

The "geneHeatmapSignPlot" plot the heatmap of the genes used to estimate the signature. The user can see how the expression of the signature genes varies across samples, identifying those genes that affect the most the score and those, on contrary, that influence the less (Figure 9).

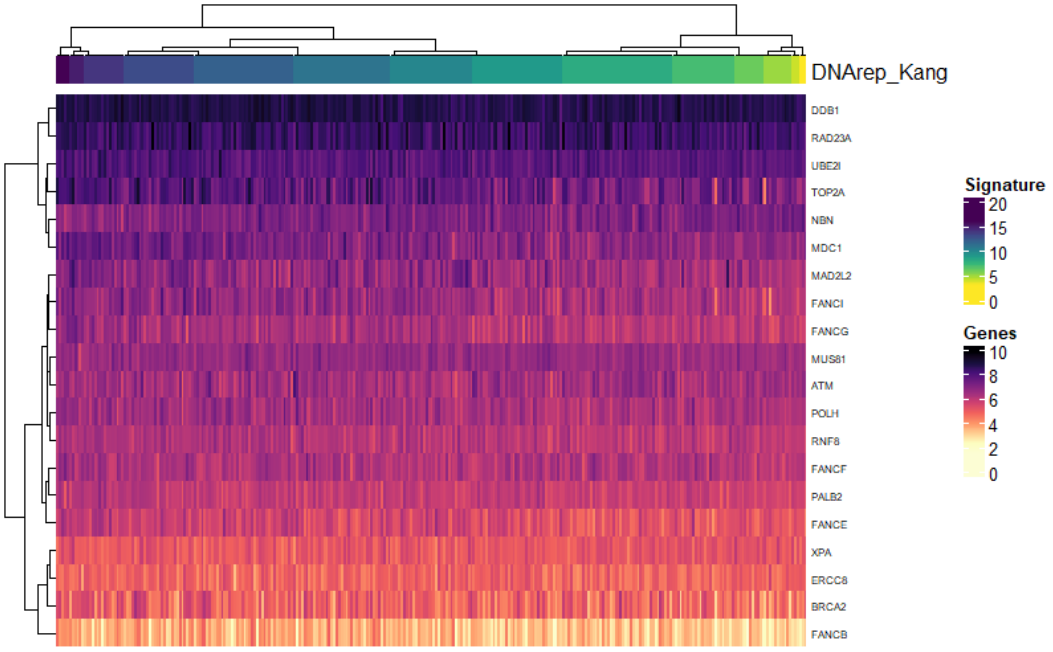

Figure 9: Plot from "geneHeatmapSignPlot" function. The plot reported is an example produced with *DNArepSign*. In horizontal there are the list of genes used by *DNArepSign* while the samples are reported on vertical.

Lastly, the "heatmapSignPlot" function plots a heatmap of the signature scores (Figure 10). The scores are scaled from 0 to 1 in order to be comparable between signatures.

The users can see in rows the signatures while on the columns the samples clustered on the score values.

The function offers the possibility to select the signatures to be plotted and to indicate one or more signatures that guide the clustering.

In the example reported, we select the four scores from *ConsensusOVSign* to guide the heatmap and different signatures to compare.

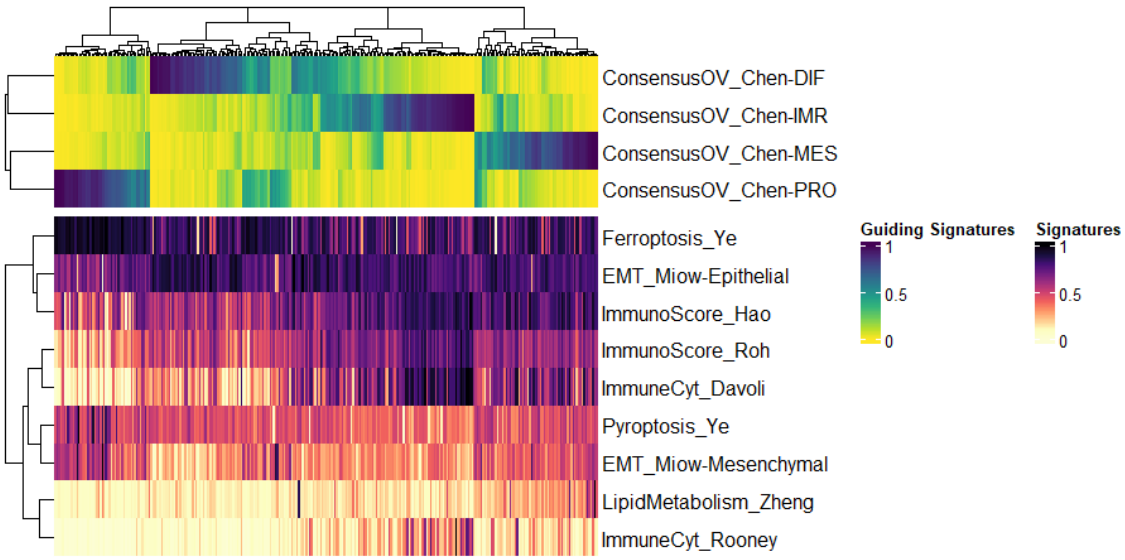

Figure 10: Plot from "heatmapSignPlot" function.The plot reported is an example produced using the four scores from *ConsensusOVSign* as guiding signatures.

4 Discussion

Cancer is a complex and dynamic entity that constantly evolves to survive and adapt in cross-connection with the host body. Different tumors share specific characteristics, called hallmarks, that include, among others, the resistance to cell death, the continuous proliferation and dysregulation in metabolisms. Additionally, tumor phenotypes are the result of the interactions between tumor cells and the tumor microenvironment (TME) that comprises normal cells and structures - lymphocytes, fibroblasts, endothelial cells, and the extracellular matrix (ECM) among others - modified to support the tumor growth [128].

In clinical practice, biomarkers can help in deciding treatments and predicting the prognosis. Their identification is an urgent need and numerous gene expression signatures have been developed to accomplish this task. In the last decades, gene expression signatures have been widely used to investigate specific tumor properties, such as the role of TME, and to provide predictions about tumor outcomes and evaluations of treatment efficacy [19]. Prime examples of these signatures can be found in breast cancer where MammaPrint® (Agendia, Amsterdam, The Netherlands) [33] is able to guide the use of adjuvant treatments in node-negative breast cancer, and Oncotype DX is able to predict 10-year distant recurrence in estrogen receptor-positive patients and also the response to chemotherapy and endocrine therapy [34].

In this scenario, the R package *signifinder*, that I contributed to develop, provides a new tool to improve the usability, reproducibility, and comparison across multiple gene expression signatures. In *signifinder*, the functions collected are implemented with R, the most common programming language for biological data analysis. *Signifinder* is able to produce scores starting from different types of gene expression input data.

Therefore, *signifinder* represents an innovative tool to make the analysis of the numerous signatures easy, fast, and reproducible.

The macro area covered by the signatures presented in *signifinder* cover the vast majority of cancer hallmarks. In order to be simple to use, *signifinder* is constructed with additional functions to provide easier visualization and interpretation of the results helping and simplifying the understanding of the role of different hallmarks within and between patient samples.

The future directions of the *signifinder* package consist in the continuous addition and implementation of signatures. Moreover, the development of recent technologies for the gene expression measurements, such as single-cell RNA-Seq (scRNA-Seq) and Spatial Transcriptomics RNA-Seq, open new perspectives in the application of the collected signatures at the single-cell level.

Acknowledgments

I would like to express my gratitude to my supervisors, Professor Chiara Romualdi and Dr. Enrica Calura, who guided me throughout this project. I would also like to thank Stefania and Laura for the time and to support me every day.

References

[1] Mete Civelek and Aldons J Lusis. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–48, 2014.

[2] FS Collins, ES Lander, J Rogers, RH Waterston, and IHGS Conso. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[3] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, 19(2):286–302, 2018.

[4] Kung-Hao Liang. *Bioinformatics for biomedical science and clinical applications*. Elsevier, 2013.

[5] J Adams. Transcriptome: connecting the genome to gene function. *Nat Educ*, 1(1):195, 2008.

[6] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[7] Paul A McGettigan. Transcriptomics in the rna-seq era. *Current opinion in chemical biology*, 17(1):4–11, 2013.

[8] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[9] Matthew N Bainbridge, René L Warren, Martin Hirst, Tammy Romanuik, Thomas Zeng, Anne Go, Allen Delaney, Malachi Griffith, Matthew Hickenbotham, Vincent Magrini, et al. Analysis of the prostate cancer cell line lncap transcriptome using a sequencing-by-synthesis approach. *BMC genomics*, 7 (1):1–11, 2006.

[10] Atul Butte. The use and analysis of microarray data. *Nature reviews drug discovery*, 1(12):951–960, 2002.

[11] Rohan Lowe, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee. Transcriptomics technologies. *PLoS computational biology*, 13(5): e1005457, 2017.

[12] Thomas E Royce, Joel S Rozowsky, and Mark B Gerstein. Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic acids research*, 35(15):e99, 2007.

[13] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[14] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[15] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 2016.

[16] Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348, 2013.

[17] Jyothi Subramanian and Richard Simon. What should physicians look for in evaluating prognostic gene-expression signatures? *Nature reviews Clinical oncology*, 7(6):327–334, 2010.

[18] R Simon. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British journal of cancer*, 89(9): 1599–1604, 2003.

[19] Frederic Chibon. Cancer gene expression signatures–the rise and fall? *European journal of cancer*, 49(8):2000–2009, 2013.

[20] Antoine Italiano. Prognostic or predictive? it's time to get back to definitions. *J Clin Oncol*, 29(35):4718, 2011.

[21] Richard Simon. The use of genomics in clinical trial design. *Clinical Cancer Research*, 14(19):5984–5993, 2008.

[22] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[23] Douglas Hanahan and Judah Folkman. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *cell*, 86(3):353–364, 1996.

[24] Johanna A Joyce and Douglas T Fearon. T cell exclusion, immune privilege, and the tumor microenvironment. *Science*, 348(6230):74–80, 2015.

[25] Rui Wei, Si Liu, Shutian Zhang, Li Min, and Shengtao Zhu. Cellular and extracellular components in tumor microenvironment and their application in early diagnosis of cancers. *Analytical Cellular Pathology*, 2020, 2020.

[26] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

[27] Kornelia Polyak and Robert A Weinberg. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nature Reviews Cancer*, 9(4):265–273, 2009.

[28] Douglas Hanahan. Hallmarks of cancer: New dimensions. *Cancer Discovery*, 12(1):31–46, 2022.

[29] David D Bowtell, Steffen Böhm, Ahmed A Ahmed, Paul-Joseph Aspuria, Robert C Bast, Valerie Beral, Jonathan S Berek, Michael J Birrer, Sarah Blagden, Michael A Bookman, et al. Rethinking ovarian cancer ii: reducing mortality from high-grade serous ovarian cancer. *Nature reviews Cancer*, 15 (11):668–679, 2015.

[30] Willy Hugo, Jesse M Zaretsky, LU Sun, Chunying Song, Blanca Homet Moreno, Siwen Hu-Lieskovan, Beata Berent-Maoz, Jia Pang, Bartosz Chmielowski, Grace Cherry, et al. Genomic and transcriptomic features of

response to anti-pd-1 therapy in metastatic melanoma. *Cell*, 165(1):35–44, 2016.

[31] Lars Dyrskjøt, Mogens Kruhøffer, Thomas Thykjaer, Niels Marcussen, Jens L Jensen, Klaus Møller, and Torben F Ørntoft. Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer research*, 64(11):4040–4048, 2004.

[32] Olivia G Taylor, Joshua S Brzozowski, and Kathryn A Skelding. Glioblastoma multiforme: an overview of emerging therapeutic targets. *Frontiers in oncology*, 9:963, 2019.

[33] Marc J Van De Vijver, Yudong D He, Laura J Van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[34] Maureen Cronin, Mylan Pho, Debjani Dutta, James C Stephans, Steven Shak, Michael C Kiefer, Jose M Esteban, and Joffre B Baker. Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay. *The American journal of pathology*, 164(1):35–42, 2004.

[35] Jean Paul Thiery, Hervé Acloque, Ruby YJ Huang, and M Angela Nieto. Epithelial-mesenchymal transitions in development and disease. *cell*, 139 (5):871–890, 2009.

[36] Samy Lamouille, Jian Xu, and Rik Derynck. Molecular mechanisms of epithelial–mesenchymal transition. *Nature reviews Molecular cell biology*, 15 (3):178–196, 2014.

[37] QH Miow, TZ Tan, J Ye, JA Lau, T Yokomizo, JP Thiery, and S Mori. Epithelial–mesenchymal status renders differential responses to cisplatin in ovarian cancer. *Oncogene*, 34(15):1899–1907, 2015.

[38] Qing Cheng, Jeffrey T Chang, William R Gwin, Jun Zhu, Stefan Ambs, Joseph Geradts, and H Kim Lyerly. A signature of epithelial-mesenchymal plasticity and stromal activation in primary tumor modulates late recurrence in breast cancer independent of disease subtype. *Breast Cancer Research*, 16(4):1–13, 2014.

[39] Milena P Mak, Pan Tong, Lixia Diao, Robert J Cardnell, Don L Gibbons, William N William, Ferdinandos Skoulidis, Edwin R Parra, Jaime Rodriguez-Canales, Ignacio I Wistuba, et al. A patient-derived, pan-cancer emt sig-

nature identifies global molecular alterations and immune target enrichment following epithelial-to-mesenchymal transition. *Clinical Cancer Research*, 22 (3):609–620, 2016.

[40] Edward A Miao, Jayant V Rajan, and Alan Aderem. Caspase-1-induced pyroptotic cell death. *Immunological reviews*, 243(1):206–214, 2011.

[41] Rong Tang, Jin Xu, Bo Zhang, Jiang Liu, Chen Liang, Jie Hua, Qingcai Meng, Xianjun Yu, and Si Shi. Ferroptosis, necroptosis, and pyroptosis in anticancer immunity. *Journal of Hematology & Oncology*, 13(1):1–18, 2020.

[42] Xiaojing Xia, Xin Wang, Zhe Cheng, Wanhai Qin, Liancheng Lei, Jinqing Jiang, and Jianhe Hu. The role of pyroptosis in cancer: pro-cancer or pro-"host"? *Cell death & disease*, 10(9):1–13, 2019.

[43] Ying Ye, Qinjin Dai, and Hongbo Qi. A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer. *Cell Death Discovery*, 7(1):1–11, 2021.

[44] Wei Shao, Zongcheng Yang, Yue Fu, Lixin Zheng, Fen Liu, Li Chai, and Jihui Jia. The pyroptosis-related signature predicts prognosis and indicates immune microenvironment infiltration in gastric cancer. *Frontiers in cell and developmental biology*, 9:1512, 2021.

[45] Wanli Lin, Ying Chen, Bomeng Wu, Zuwei Li, et al. Identification of the pyroptosis-related prognostic gene signature and the associated regulation axis in lung adenocarcinoma. *Cell death discovery*, 7(1):1–10, 2021.

[46] Xin-Yu Li, Lu-Yu Zhang, Xue-Yuan Li, Xi-Tao Yang, and Li-Xin Su. A pyroptosis-related gene signature for predicting survival in glioblastoma. *Frontiers in Oncology*, page 3168, 2021.

[47] Yanhua Mou, Jun Wang, Jinchun Wu, Dan He, Chunfang Zhang, Chaojun Duan, and Bin Li. Ferroptosis, a new form of cell death: opportunities and challenges in cancer. *Journal of hematology & oncology*, 12(1):1–16, 2019.

[48] Young-Sun Lee, Dae-Hee Lee, Haroon A Choudry, David L Bartlett, and Yong J Lee. Ferroptosis-induced endoplasmic reticulum stress: cross-talk between ferroptosis and apoptosis. *Molecular Cancer Research*, 16(7):1073–1076, 2018.

[49] Behrouz Hassannia, Peter Vandenabeele, and Tom Vanden Berghe. Targeting ferroptosis to iron out cancer. *Cancer cell*, 35(6):830–849, 2019.

[50] Jie-ying Liang, De-shen Wang, Hao-cheng Lin, Xiu-xing Chen, Hui Yang, Yun Zheng, and Yu-hong Li. A novel ferroptosis-related gene signature for overall

survival prediction in patients with hepatocellular carcinoma. *International journal of biological sciences*, 16(13):2430, 2020.

[51] Huan Liu, Lei Gao, Tiancheng Xie, Jie Li, Ting-shuai Zhai, and Yunfei Xu. Identification and validation of a prognostic signature for prostate cancer based on ferroptosis-related genes. *Frontiers in oncology*, 11, 2021.

[52] Hongyu Li, Xiliu Zhang, Chen Yi, Yi He, Xun Chen, Wei Zhao, and Dongsheng Yu. Ferroptosis-related gene signature predicts the prognosis in oral squamous cell carcinoma patients. *BMC cancer*, 21(1):1–16, 2021.

[53] Xinming Jing, Fengming Yang, Chuchu Shao, Ke Wei, Mengyan Xie, Hua Shen, and Yongqian Shu. Role of hypoxia in cancer therapy by regulating the tumor microenvironment. *Molecular cancer*, 18(1):1–15, 2019.

[54] Barbara Muz, Pilar de la Puente, Feda Azab, and Abdel Kareem Azab. The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia*, 3:83, 2015.

[55] J Martin Brown. Tumor hypoxia in cancer therapy. *Methods in enzymology*, 435:295–321, 2007.

[56] FM Buffa, AL Harris, CM West, and CJ Miller. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *British journal of cancer*, 102(2):428–435, 2010.

[57] Dapeng Hao, Jie Liu, Meng Chen, JingJing Li, Li Wang, Xiaobo Li, Qi Zhao, and Li-jun Di. Immunogenomic analyses of advanced serous ovarian cancer reveal immune score is a strong prognostic factor and an indicator of chemosensitivity. *Clinical Cancer Research*, 24(15):3560–3571, 2018.

[58] Whijae Roh, Pei-Ling Chen, Alexandre Reuben, Christine N Spencer, Peter A Prieto, John P Miller, Vancheswaran Gopalakrishnan, Feng Wang, Zachary A Cooper, Sangeetha M Reddy, et al. Integrated molecular analysis of tumor biopsies on sequential ctla-4 and pd-1 blockade reveals markers of response and resistance. *Science translational medicine*, 9(379):eaah3560, 2017.

[59] Jane L Messina, David A Fenstermacher, Steven Eschrich, Xiaotao Qu, Anders E Berglund, Mark C Lloyd, Michael J Schell, Vernon K Sondak, Jeffrey S Weber, and James J Mulé. 12-chemokine gene signature identifies lymph node-like structures in melanoma: potential for patient selection for immunotherapy? *Scientific reports*, 2(1):1–6, 2012.

[60] Michael S Rooney, Sachet A Shukla, Catherine J Wu, Gad Getz, and Nir Hacohen. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*, 160(1-2):48–61, 2015.

[61] Teresa Davoli, Hajime Uno, Eric C Wooten, and Stephen J Elledge. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, 355(6322):eaaf8399, 2017.

[62] Mark Ayers, Jared Lunceford, Michael Nebozhyn, Erin Murphy, Andrey Loboda, David R Kaufman, Andrew Albright, Jonathan D Cheng, S Peter Kang, Veena Shankaran, et al. Ifn-γ–related mrna profile predicts clinical response to pd-1 blockade. *The Journal of clinical investigation*, 127(8):2930–2940, 2017.

[63] Spencer C Wei, Colm R Duffy, and James P Allison. Fundamental mechanisms of immune checkpoint blockade therapy. *Cancer discovery*, 8(9):1069–1086, 2018.

[64] Drew M Pardoll. The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, 12(4):252–264, 2012.

[65] Scott J Antonia, José A López-Martin, Johanna Bendell, Patrick A Ott, Matthew Taylor, Joseph Paul Eder, Dirk Jäger, M Catherine Pietanza, Dung T Le, Filippo de Braud, et al. Nivolumab alone and nivolumab plus ipilimumab in recurrent small-cell lung cancer (checkmate 032): a multicentre, open-label, phase 1/2 trial. *The Lancet Oncology*, 17(7):883–895, 2016.

[66] Michael J Overman, Ray McDermott, Joseph L Leach, Sara Lonardi, Heinz-Josef Lenz, Michael A Morse, Jayesh Desai, Andrew Hill, Michael Axelson, Rebecca A Moss, et al. Nivolumab in patients with metastatic dna mismatch repair-deficient or microsatellite instability-high colorectal cancer (checkmate 142): an open-label, multicentre, phase 2 study. *The lancet oncology*, 18(9): 1182–1191, 2017.

[67] Kuang Du, Shiyou Wei, Zhi Wei, Dennie T Frederick, Benchun Miao, Tabea Moll, Tian Tian, Eric Sugarman, Dmitry I Gabrilovich, Ryan J Sullivan, et al. Pathway signatures derived from on-treatment tumor specimens predict response to anti-pd1 blockade in metastatic melanoma. *Nature communications*, 12(1):1–16, 2021.

[68] Kosuke Yoshihara, Tatsuhiko Tsunoda, Daichi Shigemizu, Hiroyuki Fujiwara, Masayuki Hatae, Hisaya Fujiwara, Hideaki Masuzaki, Hidetaka Katabuchi, Yosuke Kawakami, Aikou Okamoto, et al. High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway. *Clinical cancer research*, 18(5):1374–1385, 2012.

[69] Sipeng Shen, Guanrong Wang, Ruyang Zhang, Yang Zhao, Hao Yu, Yongyue Wei, and Feng Chen. Development and validation of an immune gene-set

based prognostic signature in ovarian cancer. *EBioMedicine*, 40:318–326, 2019.

[70] Rita Cabrita, Martin Lauss, Adriana Sanna, Marco Donia, Mathilde Skaarup Larsen, Shamik Mitra, Iva Johansson, Bengt Phung, Katja Harbst, Johan Vallon-Christersson, et al. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature*, 577(7791):561–565, 2020.

[71] Pornpimol Charoentong, Francesca Finotello, Mihaela Angelova, Clemens Mayer, Mirjana Efremova, Dietmar Rieder, Hubert Hackl, and Zlatko Trajanoski. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell reports*, 18(1):248–262, 2017.

[72] D Leanne Jones and Amy J Wagers. No place like home: anatomy and function of the stem cell niche. *Nature reviews Molecular cell biology*, 9(1): 11–21, 2008.

[73] Anna Merlos-Suárez, Francisco M Barriga, Peter Jung, Mar Iglesias, María Virtudes Céspedes, David Rossell, Marta Sevillano, Xavier Hernando-Momblona, Victoria da Silva-Diz, Purificación Muñoz, et al. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell stem cell*, 8(5):511–524, 2011.

[74] Bryan A Smith, Nikolas G Balanis, Avinash Nanjundiah, Katherine M Sheu, Brandon L Tsai, Qingfu Zhang, Jung Wook Park, Michael Thompson, Jiaoti Huang, Owen N Witte, et al. A human adult stem cell signature marks aggressive variants across epithelial cancers. *Cell reports*, 24(12):3353–3366, 2018.

[75] Nick Barker, Johan H Van Es, Jeroen Kuipers, Pekka Kujala, Maaike Van Den Born, Miranda Cozijnsen, Andrea Haegebarth, Jeroen Korving, Harry Begthel, Peter J Peters, et al. Identification of stem cells in small intestine and colon by marker gene lgr5. *Nature*, 449(7165):1003–1007, 2007.

[76] Eduard Batlle, Jeffrey T Henderson, Harry Beghtel, Maaike MW van den Born, Elena Sancho, Gerwin Huls, Jan Meeldijk, Jennifer Robertson, Marc van de Wetering, Tony Pawson, et al. β-catenin and tcf mediate cell positioning in the intestinal epithelium by controlling the expression of ephb/ephrinb. *Cell*, 111(2):251–263, 2002.

[77] Gennadi V Glinsky, Olga Berezovska, Anna B Glinskii, et al. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *The Journal of clinical investigation*, 115 (6):1503–1521, 2005.

[78] Bryan A Smith, Artem Sokolov, Vladislav Uzunangelov, Robert Baertsch, Yulia Newton, Kiley Graim, Colleen Mathis, Donghui Cheng, Joshua M Stuart, and Owen N Witte. A basal stem cell signature identifies aggressive prostate cancer phenotypes. *Proceedings of the National Academy of Sciences*, 112 (47):E6544–E6552, 2015.

[79] Christoph Lengauer, Kenneth W Kinzler, and Bert Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649, 1998.

[80] Samuel F Bakhoum and Lewis C Cantley. The multifaceted role of chromosomal instability in cancer and its microenvironment. *Cell*, 174(6):1347–1360, 2018.

[81] Samuel F Bakhoum, Bryan Ngo, Ashley M Laughney, Julie-Ann Cavallo, Charles J Murphy, Peter Ly, Pragya Shah, Roshan K Sriram, Thomas BK Watkins, Neil K Taunk, et al. Chromosomal instability drives metastasis through a cytosolic dna response. *Nature*, 553(7689):467–472, 2018.

[82] Scott L Carter, Aron C Eklund, Isaac S Kohane, Lyndsay N Harris, and Zoltan Szallasi. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature genetics*, 38(9):1043–1048, 2006.

[83] Penny A Jeggo and Markus Löbrich. How cancer cells hijack dna double-strand break repair pathways to gain genomic instability. *Biochemical Journal*, 471(1):1–11, 2015.

[84] Thomas Helleday, Justin Lo, Dik C van Gent, and Bevin P Engelward. Dna double-strand break repair: from mechanistic understanding to cancer treatment. *DNA repair*, 6(7):923–935, 2007.

[85] Kasey Rodgers and Mitch McVey. Error-prone repair of dna double-strand breaks. *Journal of cellular physiology*, 231(1):15–24, 2016.

[86] Jianping Lu, Di Wu, Chuanxing Li, Meng Zhou, and Dapeng Hao. Correlation between gene expression and mutator phenotype predicts homologous recombination deficiency and outcome in ovarian cancer. *Journal of molecular medicine*, 92(11):1159–1168, 2014.

[87] Josephine Kang, Alan D D'Andrea, and David Kozono. A dna repair pathway–focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *Journal of the National Cancer Institute*, 104 (9):670–681, 2012.

[88] Frances R Balkwill, Melania Capasso, and Thorsten Hagemann. The tumor microenvironment at a glance. *Journal of cell science*, 125(23):5591–5596, 2012.

[89] Paolo Cirri and Paola Chiarugi. Cancer associated fibroblasts: the dark side of the coin. *American journal of cancer research*, 1(4):482, 2011.

[90] Ankur Chakravarthy, Lubaba Khan, Nathan Peter Bensler, Pinaki Bose, and Daniel D De Carvalho. Tgf-β-associated extracellular matrix genes link cancer-associated fibroblasts to immune evasion and immunotherapy failure. *Nature communications*, 9(1):1–10, 2018.

[91] Laia Caja, Francesco Dituri, Serena Mancarella, Daniel Caballero-Diaz, Aristidis Moustakas, Gianluigi Giannelli, and Isabel Fabregat. Tgf-β and the tissue microenvironment: Relevance in fibrosis and cancer. *International journal of molecular sciences*, 19(5):1294, 2018.

[92] Arseniy E Yuzhalin, Tomas Urbonas, Michael A Silva, Ruth J Muschel, and Alex N Gordon-Weeks. A core matrisome gene signature predicts cancer outcome. *British journal of cancer*, 118(3):435–440, 2018.

[93] A Gordon Robertson, Jaegil Kim, Hikmat Al-Ahmadie, Joaquim Bellmunt, Guangwu Guo, Andrew D Cherniack, Toshinori Hinoue, Peter W Laird, Katherine A Hoadley, Rehan Akbani, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, 171(3):540–556, 2017.

[94] Charles H Spruck, Petra F Ohneseit, Mirella Gonzalez-Zulueta, David Esrig, Noriomi Miyao, Yvonne C Tsai, Seth P Lerner, Christoph Schmütte, Allen S Yang, Richard Cote, et al. Two molecular pathways to transitional cell carcinoma of the bladder. *Cancer research*, 54(3):784–788, 1994.

[95] Lance A Liotta and Elise C Kohn. Cancer's deadly signature. *Nature genetics*, 33(1):10–11, 2003.

[96] Zhiyuan Hu, Cheng Fan, Chad Livasy, Xiaping He, Daniel S Oh, Matthew G Ewend, Lisa A Carey, Subbaya Subramanian, Robert West, Francis Ikpatt, et al. A compact vegf signature associated with distant metastases and poor outcomes. *BMC medicine*, 7(1):1–14, 2009.

[97] Otto Warburg. On the origin of cancer cells. *Science*, 123(3191):309–314, 1956.

[98] Otto Warburg. On respiratory impairment in cancer cells. *Science*, 124 (3215):269–270, 1956.

[99] Ralph J DeBerardinis, Julian J Lum, Georgia Hatzivassiliou, and Craig B Thompson. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell metabolism*, 7(1):11–20, 2008.

[100] Lei Zhang, Zhe Zhang, and Zhenglun Yu. Identification of a novel glycolysis-

related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma. *Journal of translational medicine*, 17(1):1–13, 2019.

[101] Fangshi Xu, Yibing Guan, Li Xue, Shanlong Huang, Ke Gao, Zhen Yang, and Tie Chong. The effect of a novel glycolysis-related gene signature on progression, prognosis and immune microenvironment of renal cell carcinoma. *BMC cancer*, 20(1):1–19, 2020.

[102] Jianrong Lu. The warburg metabolism fuels tumor metastasis. *Cancer and Metastasis Reviews*, 38(1):157–164, 2019.

[103] Chunming Cheng, Feng Geng, Xiang Cheng, and Deliang Guo. Lipid metabolism reprogramming and its potential targets in cancer. *Cancer Communications*, 38(1):1–14, 2018.

[104] Marteinn Thor Snaebjornsson, Sudha Janaki-Raman, and Almut Schulze. Greasing the wheels of the cancer machine: the role of lipid metabolism in cancer. *Cell metabolism*, 31(1):62–76, 2020.

[105] Mingjun Zheng, Heather Mullikin, Anna Hester, Bastian Czogalla, Helene Heidegger, Theresa Vilsmaier, Aurelia Vattai, Anca Chelariu-Raicu, Udo Jeschke, Fabian Trillsch, et al. Development and validation of a novel 11-gene prognostic model for serous ovarian carcinomas based on lipid metabolism expression profile. *International journal of molecular sciences*, 21(23):9169, 2020.

[106] Guangyuan Zhao, Horacio Cardenas, and Daniela Matei. Ovarian cancer—why lipids matter. *Cancers*, 11(12):1870, 2019.

[107] Marija Dmitrijeva, Stephan Ossowski, Luis Serrano, and Martin H Schaefer. Tissue-specific dna methylation loss during ageing and carcinogenesis is linked to chromosome structure, replication timing and cell division rates. *Nucleic acids research*, 46(14):7022–7039, 2018.

[108] Zhen Yang, Andrew Wong, Diana Kuh, Dirk S Paul, Vardhman K Rakyan, R David Leslie, Shijie C Zheng, Martin Widschwendter, Stephan Beck, and Andrew E Teschendorff. Correlation of an epigenetic mitotic clock with cancer risk. *Genome biology*, 17(1):1–18, 2016.

[109] Beth Levine and Guido Kroemer. Autophagy in the pathogenesis of disease. *Cell*, 132(1):27–42, 2008.

[110] Eileen White and Robert S DiPaola. The double-edged sword of autophagy modulation in cancer. *Clinical cancer research*, 15(17):5308–5316, 2009.

[111] Robin Mathew, Vassiliki Karantza-Wadsworth, and Eileen White. Role of autophagy in cancer. *Nature Reviews Cancer*, 7(12):961–967, 2007.

[112] Yang Xu, Renpeng Li, Xiaoxia Li, Naijun Dong, Di Wu, Lin Hou, Kan Yin, and Chunhua Zhao. An autophagy-related gene signature associated with clinical prognosis and immune microenvironment in gliomas. *Frontiers in oncology*, page 2036, 2020.

[113] Mei Chen, Shufang Zhang, Zhenyu Nie, Xiaohong Wen, and Yuanhui Gao. Identification of an autophagy-related prognostic signature for clear cell renal cell carcinoma. *Frontiers in oncology*, 10:873, 2020.

[114] Zihao Wang, Lu Gao, Xiaopeng Guo, Chenzhe Feng, Wei Lian, Kan Deng, and Bing Xing. Development and validation of a nomogram with an autophagy-related gene signature for predicting survival in patients with glioblastoma. *Aging (Albany NY)*, 11(24):12246, 2019.

[115] Hengyu Chen, Qingchun Deng, Wenwen Wang, Huishan Tao, and Ying Gao. Identification of an autophagy-related gene signature for survival prediction in patients with cervical cancer. *Journal of ovarian research*, 13(1):1–10, 2020.

[116] Leland H Hartwell, Joseph Culotti, John R Pringle, and Brian J Reid. Genetic control of the cell division cycle in yeast: A model to account for the order of cell cycle events is deduced from the phenotypes of yeast mutants. *Science*, 183(4120):46–51, 1974.

[117] Helen K Matthews, Cosetta Bertoli, and Robertus AM de Bruin. Cell cycle control in cancer. *Nature Reviews Molecular Cell Biology*, 23(1):74–88, 2022.

[118] Marcos Malumbres and Mariano Barbacid. Cell cycle, cdks and cancer: a changing paradigm. *Nature reviews cancer*, 9(3):153–166, 2009.

[119] Arian Lundberg, Linda S Lindström, J Chuck Harrell, Claudette Falato, Joseph W Carlson, Paul K Wright, Theodoros Foukakis, Charles M Perou, Kamila Czene, Jonas Bergh, et al. Gene expression signatures and immunohistochemical subtypes add prognostic value to each other in breast cancer cohorts. *Clinical Cancer Research*, 23(24):7512–7520, 2017.

[120] Gregory M Chen, Lavanya Kannan, Ludwig Geistlinger, Victor Kofia, Zhaleh Safikhani, Deena MA Gendoo, Giovanni Parmigiani, Michael Birrer, Benjamin Haibe-Kains, and Levi Waldron. Consensus on molecular subtypes of high-grade serous ovarian carcinoma. *Clinical Cancer Research*, 24(20):5037–5047, 2018.

[121] Stephanie Lheureux, Charlie Gourley, Ignace Vergote, and Amit M Oza. Epithelial ovarian cancer. *The Lancet*, 393(10177):1240–1253, 2019.

[122] William P McGuire, William J Hoskins, Mark F Brady, Paul R Kucera, Edward E Partridge, Katherine Y Look, Daniel L Clarke-Pearson, and Martin Davidson. Cyclophosphamide and cisplatin compared with paclitaxel and

cisplatin in patients with stage iii and stage iv ovarian cancer. *New England Journal of Medicine*, 334(1):1–6, 1996.

[123] Roshan Agarwal and Stan B Kaye. Ovarian cancer: strategies for overcoming resistance to chemotherapy. *Nature Reviews Cancer*, 3(7):502–516, 2003.

[124] Cheryl A Sherman-Baust, Kevin G Becker, William H Wood III, Yongqing Zhang, and Patrice J Morin. Gene expression and pathway analysis of ovarian cancer cells selected for resistance to cisplatin, paclitaxel, or doxorubicin. *Journal of ovarian research*, 4(1):1–11, 2011.

[125] Lihua Cheng, Wei Lu, Bhushan Kulkarni, Tanja Pejovic, Xiaowei Yan, Jung-Hsien Chiang, Leroy Hood, Kunle Odunsi, and Biaoyang Lin. Analysis of chemotherapy response programs in ovarian cancers by the next-generation sequencing technologies. *Gynecologic oncology*, 117(2):159–169, 2010.

[126] Ann-Marie Patch, Elizabeth L Christie, Dariush Etemadmoghadam, Dale W Garsed, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, Peter J Bailey, et al. Whole–genome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553):489–494, 2015.

[127] Boris J Winterhoff, Makayla Maile, Amit Kumar Mitra, Attila Sebe, Martina Bazzaro, Melissa A Geller, Juan E Abrahante, Molly Klein, Raffaele Hellweg, Sally A Mullany, et al. Single cell sequencing reveals heterogeneity within ovarian cancer epithelium and cancer associated stromal cells. *Gynecologic oncology*, 144(3):598–606, 2017.

[128] Douglas Hanahan and Lisa M Coussens. Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer cell*, 21(3):309–322, 2012.

A Supplementary material

Function Name	Tumor Type	Category	Author Ref	Input
EMTSign	ovarian cancer	Epithelial-to-mesenchymal	Miow [37]	M, R
EMTSign	pan-cancer	Epithelial-to-mesenchymal	Mak [39]	M, R
EMTSign	breast cancer	Epithelial-to-mesenchymal	Cheng [38]	M, R
pyroptosisSign	ovarian cancer	Pyroptosis	Ye [43]	R
pyroptosisSign	gastric cancer	Pyroptosis	Shao [44]	M, R
pyroptosisSign	lung adenocarcinoma	Pyroptosis	Lin [45]	R
pyroptosisSign	glioblastoma multiforme	Pyroptosis	Li [46]	R
ferroptosisSign	ovarian cancer	Ferroptosis	Ye [43]	M, R
ferroptosisSign	hepatocellular carcinoma	Ferroptosis	Liang [50]	R
ferroptosisSign	prostate cancer	Ferroptosis	Liu [51]	M, R
ferroptosisSign	oral squamous cell carcinoma	Ferroptosis	Li [52]	R
lipidMetabolismSign	epithelial ovarian cancer	Altered metabolism	Zheng [105]	R
hypoxiaSign	pan-cancer	Hypoxia	Buffa [56]	M, R
platinumResSign	high grade serous ovarian cancer	Platinum resistance	Winterhoff [127]	M, R
immunoScoreSign	epithelial ovarian cancer	Immune System	Hao [57]	M, R
immunoScoreSign	pan-cancer	Immune System	Roh [58]	R
consensusOVSign	high-grade serous ovarian carcinoma	Tumor subtypes	Chen [120]	M, R
IPSSign	pan-cancer	Immune System	Charoentong [71]	R
matrisomeSign	ovarian cystadenocarcinoma, gastric adenocarcinoma, colorectal adenocarcinoma, lung adenocarcinoma	Extracellular matrix	Yuzhalin [92]	M, R
mitoticIndexSign	pan-cancer	Mitotic	Yang [108]	R
immuneCytSign	pan-cancer	Immune System	Rooney [60]	M, R
IFNSign	pan-cancer	Immune System	Ayers [62]	R
expandedImmuneSign	pan-cancer	Immune System	Ayers [62]	R
TinflamSign	pan-cancer	Immune System	Ayers [62]	R
TLSSign	melanoma	Immune System	Cabrita	M, R
stemCellCD49fSign	prostate cancer	Stem cell	Smith [78]	R
glycolysisSign	lung adenocarcinoma	Altered metabolism	Zhang [100]	R
glycolysisSign	renal cell carcinoma	Altered metabolism	Xu [101]	R
CINSign	pan-cancer	Chromosomal instability	Carter [82]	M, R
cellCycleSign	pan-cancer	Cell cycle	Lundberg [119]	M, R
cellCycleSign	pan-cancer	Cell cycle	Davoli [61]	M, R
autophagySign	glioma	Autophagy	Xu [112]	R
autophagySign	clear cell renal cell carcinoma	Autophagy	Chen M	R
autophagySign	glioblastoma	Autophagy	Wang [114]	R
autophagySign	cervical cancer	Autophagy	Chen H	M, R
ASCSign	pan-cancer	Cancer stem cell	Smith [74]	M, R
immuneCytSign	pan-cancer	Immune System	Davoli [61]	M, R
chemokineSign	pan-cancer	Immune System	Messina [59]	M, R
ISCSign	colorectal cancer	Cancer stem cell	Merlos-Suarez [73]	M, R
PassONSign	metastatic melanoma	Immune System	Du [67]	R
IPRESSign	mestatic melanoma	Immune System	Hugo [30]	R
ECMsign	pan-cancer	Extracellular matrix	Chakravarthy [90]	R
CISSign	bladder cancer	Carcinoma in situ	Robertson [93]	M, R
HRDSSign	ovarian cancer, breast cancer	Chromosomal instability	Lu [86]	M, R
VEGFSign	pan-cancer	Angiogenesis	Hu [96]	M, R
DNArepSign	serous ovarian cystadenocarcinoma	Chromosomal instability	Kang [87]	M, R
IPSOVSign	ovarian cancer	Immune System	Shen [69]	M, R

Table 1: M = microarray and R = RNASeq.