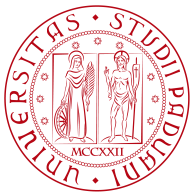


University of Padova

Department of Mathematics Tullio-Levi Civita

Master Thesis in Data Science

Exploring the effects of copy number variations on the tumor microenvironment in ovarian cancer at the single cell level

Supervisor

Professor Chiara Romualdi

University of Padova

Master Candidate

Elisa Ferrero

Academic Year

2023-2024

A tutte le persone che mi hanno cambiata.

Abstract

Cancer progression is driven by a complex interplay of the tumor cells with the surrounding environment, including the stroma and immune cells. Unveiling how the mutational processes involved in tumorigenesis affect this complexity is an important challenge in cancer research.

Among tumor types, high-grade serous ovarian cancer is characterized by great genomic instability, with large portions of the genome affected by copy number alterations. Mutational signatures are an important tool in decoding genomics instability, being recurrent patterns of mutations associated with specific biological processes and clinical variables.

In this thesis, through the analysis of a publicly available dataset, we explore the impact of copy number mutational signatures on the tumor microenvironment in high-grade serous ovarian cancer, evaluating how different patterns of copy number variation may lead to alterations in the abundance of different cell types, together with their gene expression programs and communication between cells.

Contents

Listing of figures

Listing of tables

Listing of acronyms

CNA . . . . . Copy Number Alterations

DEA/DEG . . . . . Differential Expression Analysis / Differentially Expressed Genes

FDR . . . . . False Discovery Rate

GOBP . . . . . Gene Ontology Biological Process

GSEA . . . . . Gene Set Enrichment Analysis

HGSOC . . . . . High Grade Serous Ovarian Cancer

LGSOC . . . . . Low Grade Serous Ovarian Cancer

ORA . . . . . Over Representation Analysis

scRNA-seq . . . . . Single Cell RNA Sequencing

SNP . . . . . Single Nucleotide Polymorphism

TME . . . . . Tumor Micro-Environment

WGS . . . . . Whole Genome Sequencing

1 Introduction

In this chapter we will briefly introduce the main concepts and background of this work. It is structured as follows:

- In section 1.1, we briefly introduce the main characteristics and issues associated with High-Grade Serous Ovarian Cancer (HGSOC);
- Then in section 1.2, we introduce the definition of copy number mutations and mutational signatures;
- In section 1.3 we then outline the tumor microenvironment and its role in cancer progression;
- In section 1.4, we present the experimental design and data structure;
- In section 1.5, we summarize the thesis work, giving an overview of the analyses performed.

1.1 High-Grade Serous Ovarian Carcinoma

Ovarian cancer is the seventh most common tumor and the eighth cause of cancer death in women worldwide, with an estimated 225500 cases and 140000 deaths each year, according to

the WHO [1]. This is partly due to the difficulties associated with diagnosis, as in over 70% of cases the disease has already progressed to FIGO stage III or IV by then [2].

Ovarian cancers can be classified into three main types: epithelial, germ cell and sex-cord-stromal, with epithelial cancers making up around 95% of cases [2].

Epithelial ovarian cancers can be further divided into serous, endometroid, mucinous, and clear-cell based on the tissue of origin, with serous carcinomas arising from the serous epithelial layer. The serous subtype is the most common, comprising arount 75% of all epithelial cancers [2].

Serous ovarian carcinomas are then classifies as either Low Grade or High Grade. LGSOC is characterized by low cell proliferation and point mutations in *KRAS* and *BRAF* genes. By contrast, HGSOC is by far the most common type of serous ovarian cancer (more than 90% of cases), and exhibits a very high frequency of *TP53* mutations [3]. It also the subtype of ovarian cancer that comprises the most deaths, between 70% and 80% [1].

It is further characterized by a high degree of genomic instability and copy number variation, with large parts of the genome containing amplified or deleted segments [1].

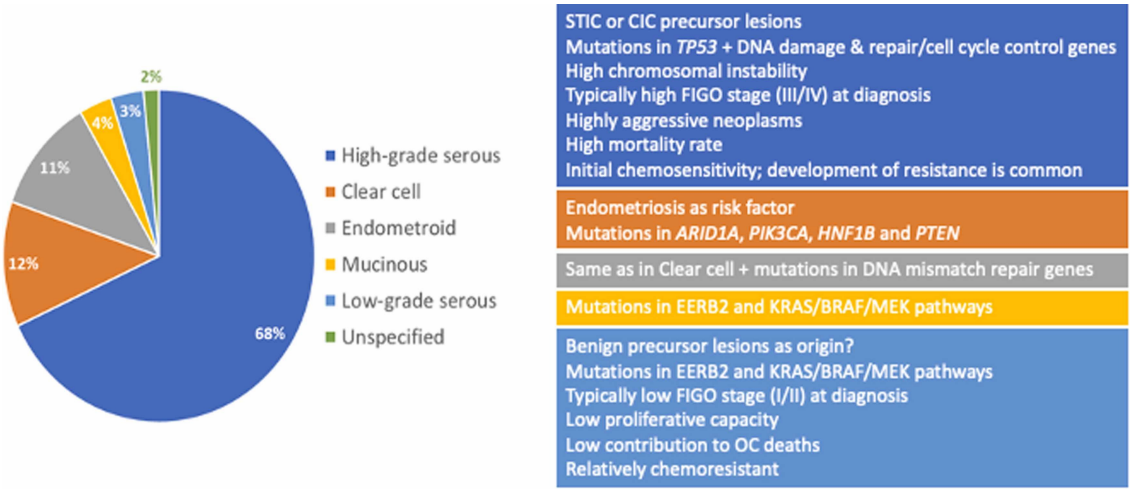


Figure 1.1: Histological classification of epithelial ovarian cancers and associated traits. As we can see, HGSOC is the most common and aggressive subtype. Figure from [4]

1.2 Copy Number Mutational Signatures

Somatic mutations are modifications in the sequence of DNA. They range between small mutations (SNPs, small insertions / deletions) and large ones who affect up to millions of base

pairs [5].

CN mutations are a particular type of large mutations, where a segment of the genome comprising many genes gains or loses copies. Examples of such events are whole genome duplication and loss of heterozigosity. CN mutations can have many underlying causes, among which impaired homologous recombination, replication stress and chromosome missegregation [6].

CN profiles can be inferred from SNP array data or WGS, allowing for the calculation of CN along the genome and its segmentation into regions with constant CN [5].

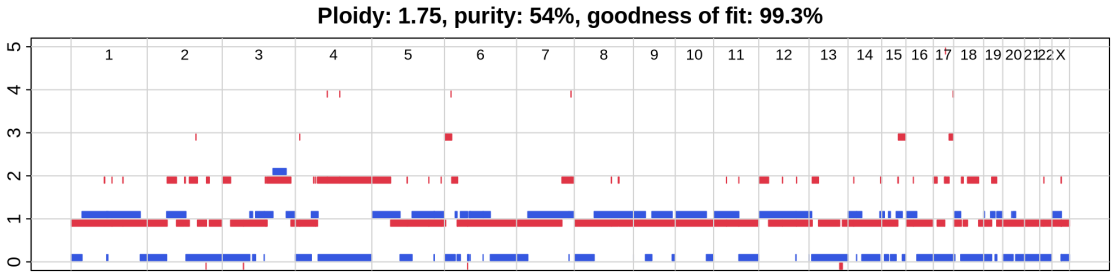


Figure 1.2: Example of CN profile reconstructed from WGS data using ASCAT. The x axis is the position in the genome of the segment, divided by chromosome. On the y axis is the estimated copy number for the segment, with the minor and major alleles in blue and red respectively.

Given a large number of CN profiles, mutational signatures can then be defined. They are recurrent patterns in mutational profiles, generally associated with a probable mutational mechanism as their cause [5]. We go more in depth on the extraction of mutational signatures in section 2.2.

CN mutational signatures are linked to specific mutational processes, and are often correlated with clinical variables and prognosis [6][7][8]. As such, there is a need to investigate how different CN mutational patterns affect the tumor microenvironment, in order to link mutational processes to the tumor-immune phenotype and improve our understanding of the underlying mechanisms.

1.3 scRNA-seq and the Tumor Micro-Environment

The tumor microenvironment (that is to say, the cellular environment in which the tumor develops) plays an important role in cancer development and progression [9]. It encompasses cells ranging from stromal cells, fibroblasts, and endothelial cells to adaptive (T and B lymphocytes) and innate (NK cells, macrophages) immune cells, together with noncellular components such as the extracellular matrix and signaling molecules [10].

During tumor development, cancer cells recruit and reprogram host cells in the TME to favor progression of the disease through the secretion of signaling molecules, such as cytokines and chemokynes [11]. For example, cancer cells can recruit components of the immune system to assume an immuno suppressive role [11], and convert fibroblasts into cancer-associated fibroblasts who promote tumor growth [9].

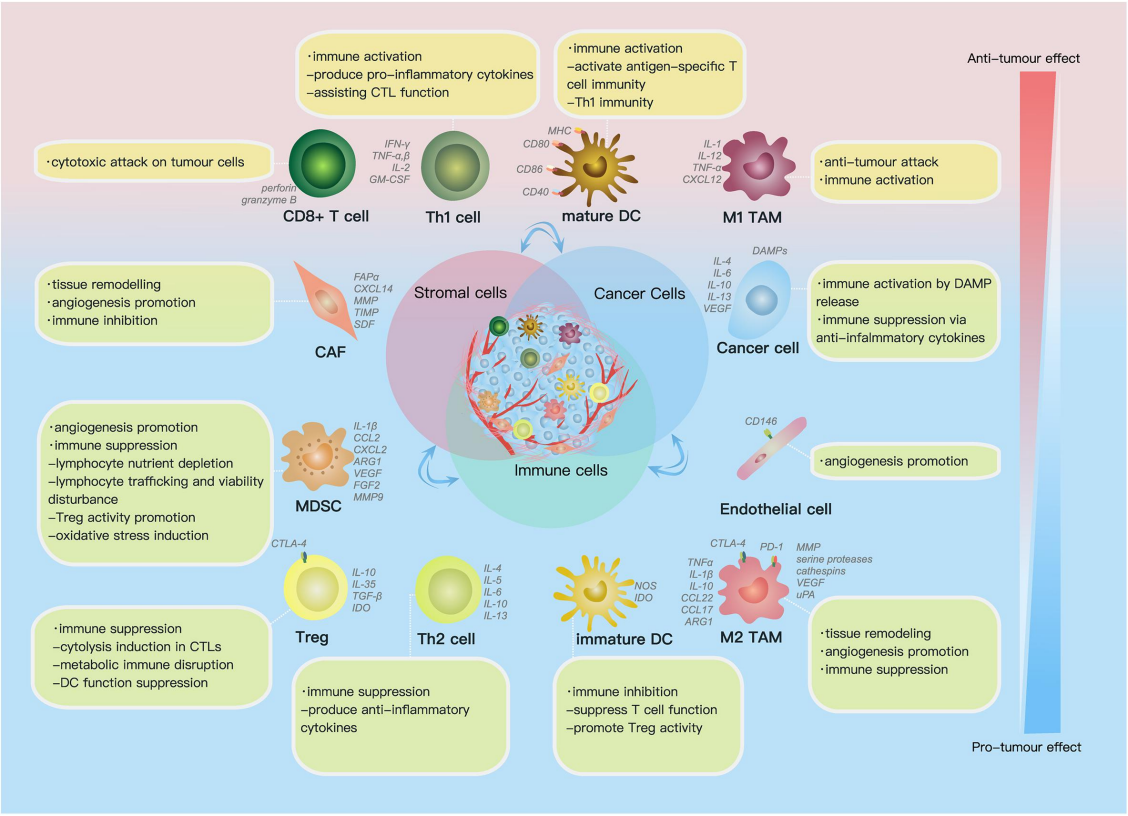


Figure 1.3: Summary of the main cell types in the ovarian cancer TME and their role. The TME comprises both the tumor itself and immune and stromal host cells, which can have both pro and anti tumor roles. Figure from [12]

In the case of ovarian cancer, the TME can range from solid (primary lesions, omentum) to liquid niches (ascites), with a high degree of heterogeneity both between patients and between different sites in the same patient [4].

Compared to bulk RNA-seq, single-cell RNA-seq allows for the transcriptomic profiling of single cells. This improved resolution makes it an important tool to investigate the complexity of the TME. In particular, it has allowed for the identification of novel subclusters of cells, both in cancer associated fibroblasts and immune populations [10].

1.4 The MSK-Spectrum Study

In this thesis we used the data from the MSK SPECTRUM study [13].

Tissue biopsies were collected from newly diagnosed and treatment-naive patients of HGSOC, for a total of 160 samples across 42 patients.

The collection took place at multiple anatomical sites, including the adnexa (primary lesions), omentum, peritoneum, bowel, ascites and other intraperitoneal sites.

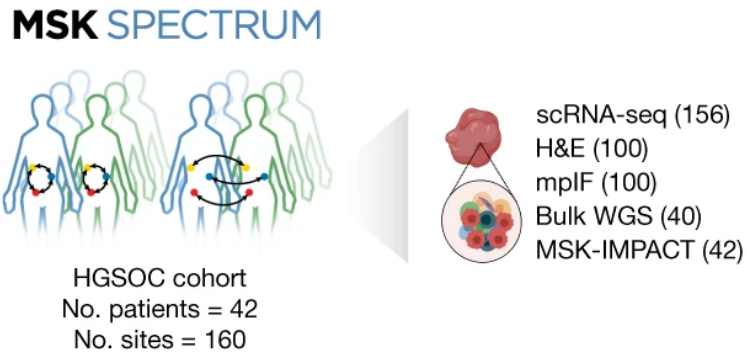


Figure 1.4: MSK SPECTRUM cohort overview. As shown, we have data from multiple anatomical sites for each patient, and from multiple patients for each site. Figure from [13]

These samples were first flow-sorted into CD45+ and CD45− fractions, and then profiled using scRNA-seq.

For 40 patients, WGS was also performed on matched normal and tumor samples from a representative site. However, one of these patients does not have any associated scRNA-seq samples, which leaves us with 39 suitable subjects for our analysis.

The authors then performed a thorough analysis of the immunophenotype of the collected samples, comparing them across sites and mutational subtype (distinguishing between homologous recombination depleted and foldback inversion-bearing tumors) to obtain a detailed report of the impact of mutational processes and anatomical site on the TME in HGSOC.

We explored a similar line of inquiry, focusing instead on copy number alteration as our mutational process of interest and quantifying its impact on the TME.

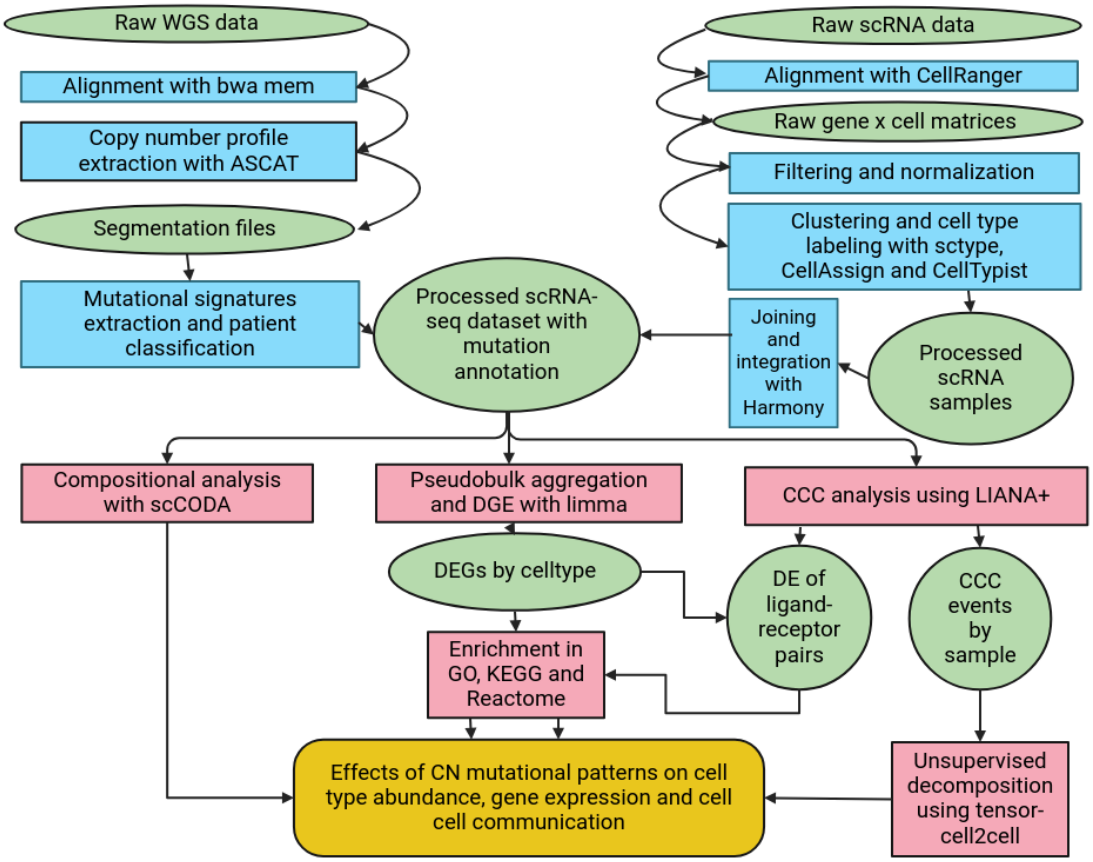


Figure 1.5: Summary of work. In the green circles we can find the datasets we worked on and the results we obtained from them; while in the light blue and pink rectangles we describe the processing steps the data went through, in preprocessing and downstream analyses respectively

1.5 Brief summary of thesis work

In this thesis, we investigate the impact of CN mutational signatures on the TME. Specifically, we stratify patients according to their mutational signature profiles inferred from three different algorithms to explore differences and similarities in terms of TME composition, gene expression and cell-cell communication

This thesis is structured as follows:

- In chapter 2, the computational methods for the inference of mutational signatures and patient classification are described;
- In chapter 3, we present the results of scRNA-seq data up to cell type annotation;

- In chapter 4, we show the results obtained from the comparative evaluation of patients characterised by different CNV signatures, in terms of TME cell type composition;
- In chapter 5, we report the result of differential expression analysis between groups of patients;
- In chapter 6, we explore the impact that different CNV signatures might have on the rewiring of cell-to-cell network communication;
- We then conclude in chapter 7, summarizing our results and proposing possible future directions

2

Mutational signatures from WGS data

As we explained in section 1.2, CN signatures are recurrent patterns in the CNA profiles of cancer patients. As they are linked to background mutational processes [6][7], the study of their impact on the tumor microenvironment is necessary to link mutational processes and resulting tumor-immune phenotypes in the microenvironment.

As such, our first goal is the extraction of these signatures from the matched tumor-normal WGS data and the stratification of patients, which are the subjects of this chapter.

In section 2.1, we explain how CN profiles can be inferred from WGS data using ASCAT. We then go over the extraction of mutational signatures and the four methods we used in section 2.2, to then conclude the chapter in section 2.3, in which we cluster the patients by their CN mutational signature activities.

2.1 Genome segmentation with ASCAT

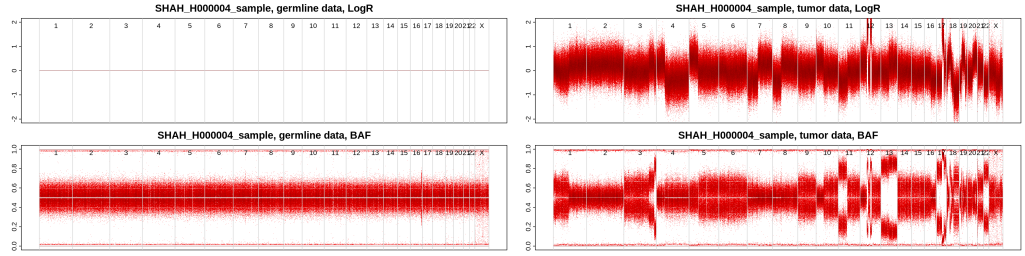
Our first step was the alignment of the raw WGS data. This was done using bwa-mem [14], with standard parameters and the GRCh38 human reference genome.

After mapping the reads and compressing the resulting SAM files into BAMs, we used these matched tumor and normal alignment files as input for the ASCAT pipeline, from which we obtained a segmentation table, describing start, end and CN for the major and minor alleles of the genome segments with consistent CN resulting from mutational events.

We used ASCAT (allele-specific copy number analysis of tumors) [15], versions 3.1.2. While originally intended for use on SNP array data, it can also extract CN profiles from WGS data.

To extract the genome segmentation from WGS data, ASCAT works in two steps:

1. From the .bam files for tumor and normal samples, logR and BAF tracks must be reconstructed. In SNP array data, the logR track is the internsity of the signal, while the BAF is the fraction of signal from the B (minor) allele. Since our data is from WGS, these tracks must be inferred;



(a) Normal sample **(b)** Tumor sample

Figure 2.1: logR and BAF tracks reconstructed from WGS data, for both germline and tumor samples

2. The genome is then segmented, fitting piecewise constant functions simultaneously to the logR and BAF tracks, constraining breaks to occur at the same positions in the genome in both functions. However, in order to impute the final CNs, since the purity of the sample is unknown, ASCAT estimates the purity and ploidy together, solving a minimization problem defined in the following way: given the smoothed data from the previous step, for each segment the major and minor copy numbers can be expressed as a function of the logR, BAF (known), purity and ploidy (to be estimated).

   Given a possible purity and ploidy pair, a loss function is defined as the sum, over each segment, of the squared distance between the calculated copy numbers and the closest nonnegative whole number. This calculation is performed for all possible purities and ploidies over a grid.

An example of CN profile extracted by ASCAT can be found at figure 1.2. For one patient, ASCAT was unable to solve the purity / ploidy optimization problem, leaving us with CN profiles for 38 patients.

After reconstructing the CN profile, our next step was the inference of copy number signature activities.

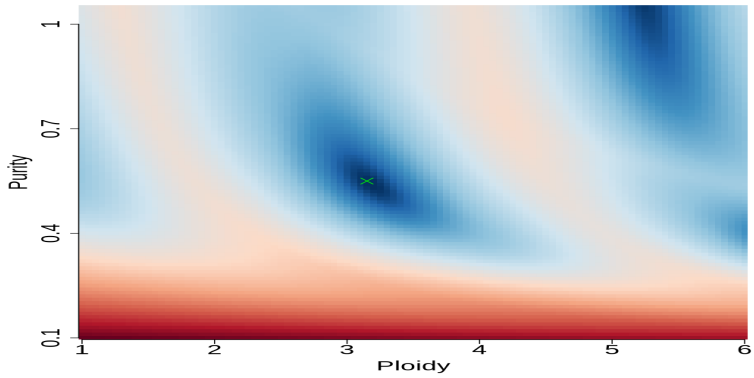


Figure 2.2: Loss function over the grid of possible purities and ploidies. All local minima are considered as possible interpretations, with the best one chosen using a goodness-of-fit measure

2.2 Extraction of mutational signatures

As said in section 1.2, CN mutational signatures are recurrent patterns of CNV. In particular, in this thesis we worked on definitions of mutational signatures from four different works [16][6][7][8].

In [16], the authors define a classification for mutational profiles of genomes in ovarian carcinoma based on two features extracted from the segmentation. We will cover this in more detail in section 2.2.1.

By contrast, the strategies used to define mutational signatures in [6], [7] and [8] are similar, based on Non-Negative Matrix Factorization. Given a $n \times m$ matrix V, NNMF finds the best approximation of V as $V \sim WH$, with W an $n \times p$ matrix and H an $p \times m$ one. When $p << n, m$, this can be interpreted as NNMF finding latent factors in the data, with matrix H describing these factors and W describing their activity in the original data.

To define mutational signatures, the authors start from a database of thousands of WGS samples and relative segmentations.

1. First, a number of features are extracted from each segment: these summarize the characteristics of the segment, and can range from the length [7] to the CN change [6];

2. Then, the features are used to define a classification for the segments, and the CN profile of the sample is summarized by vector of the number of segments in each class. The vectors representing each sample are then used to build a $sample \times component$ matrix is obtained;

3. NNMF is then used to decompose this matrix as the product of a $sample \times signature$ and a $signature \times component$ matrices. The first one is interpreted as the activity

of mutational signatures in the samples, while the second one defines the signatures as vectors of components;

4. Given a new CN profile, it can then be summarized by the vector of the number of segments in each class in the way defined above, and the activities of the previously defined mutational signatures can be quantified by decomposing this vector into a non-negative linear combination of the vectors describing the signatures.

For all three papers that use this method, we decomposed the CN profiles in our dataset into the signatures they defined: 17 signatures from [6], 25 from [7] and 14 from [8].

2.2.1 HU signature extraction

In [16], the authors classify somatic copy number alterations (SCNA) as arm-level (involving an entire chromosome arm), focal (affecting $< 25\%$ of the arm) or broad (between focal and arm-level).

They then classify the patients' mutational patterns using a decision tree that takes into account two variables:

- The **Copy Number Burden**, defined as the percentage of the genome affected by copy number changes;
- The **Somatic Copy Number Alteration Length**, based on the distribution of lengths of SCNAs, normalized by chromosome arm length. The original parameter was based on the distribution of SCNA lengths in shallow WGS. In our case, since our data has a much higher coverage, resulting in a more fragmented genome, we changed the SCNA length to be the $(1 - 2^{-4})$th quantile of the distribution.

Using these two variables, they define a classification for mutational patterns of genomes specifically designed for ovarian cancer:

- **Stable**(S) genomes are characterized by a generally stable genome, lacking relevant SCNAs;
- **Unstable** (U) genomes mainly show large arm-level SCNAs;
- **Highly Unstable** (HU) genomes are affected by many SCNAs, at all levels

The tree used for the classification is described in figure 2.3.

Applying this classification scheme on our dataset, we found 7 patients with an Unstable profile, and 30 with an Highly Unstable one. We also found a single patient with a Stable profile, that we discarded as an aberration.

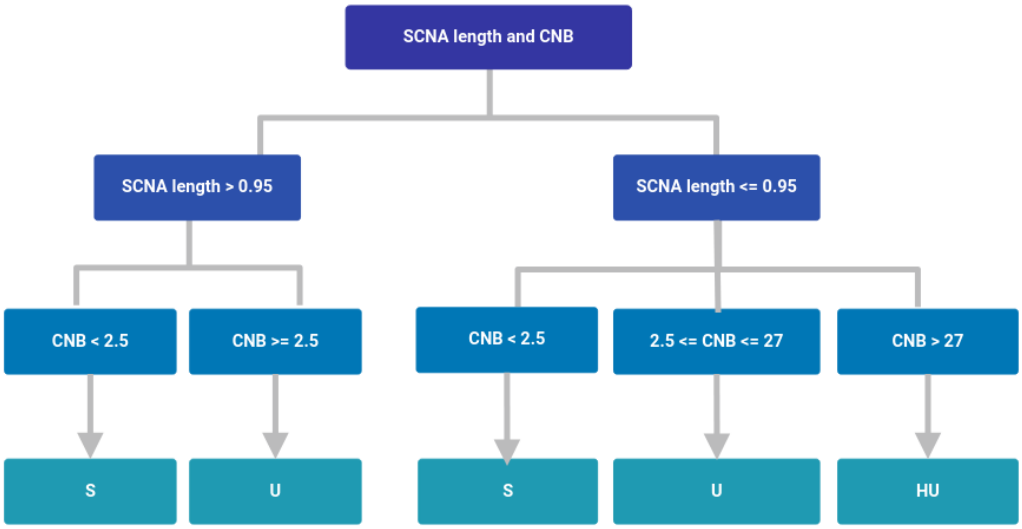


Figure 2.3: Decision tree for patient classification, based on SCNA length and CNB. Figure based on [16]

2.2.2 Drews et al. signatures extraction

We extracted the mutational signatures from [6] using the R package *CINSignatureQuantification*, version 1.2.0.

For each patient, after extracting the 5 features defined in [6] (copy number change between neighboring segments, segment length, breakpoints per 10 Mb and breakpoints per chromosome arm), we summarized the CN mutational profile by binning the segments into the 43 classes defined in [6] and imputed the activities of the 17 signatures. Adding together the activity for each signature in all patients, we find that the most active signatures in our dataset are CX1, CX2, CX3 and CX5, which is coherent with what is stated in [6]. The authors also propose aetiologies for each signature:

- CX1, which is related to whole-arm and whole-chromosome changes, is suggested to be caused by chromosome missegregation during mitosis, and negatively correlated with telomere length [6];
- CX2, CX3 and CX5 are all related to impaired homologous recombination, with an increased strength of the association from CX2 to CX5 and then CX3. In particular, CX3 and CX5 are also associated with replication stress.

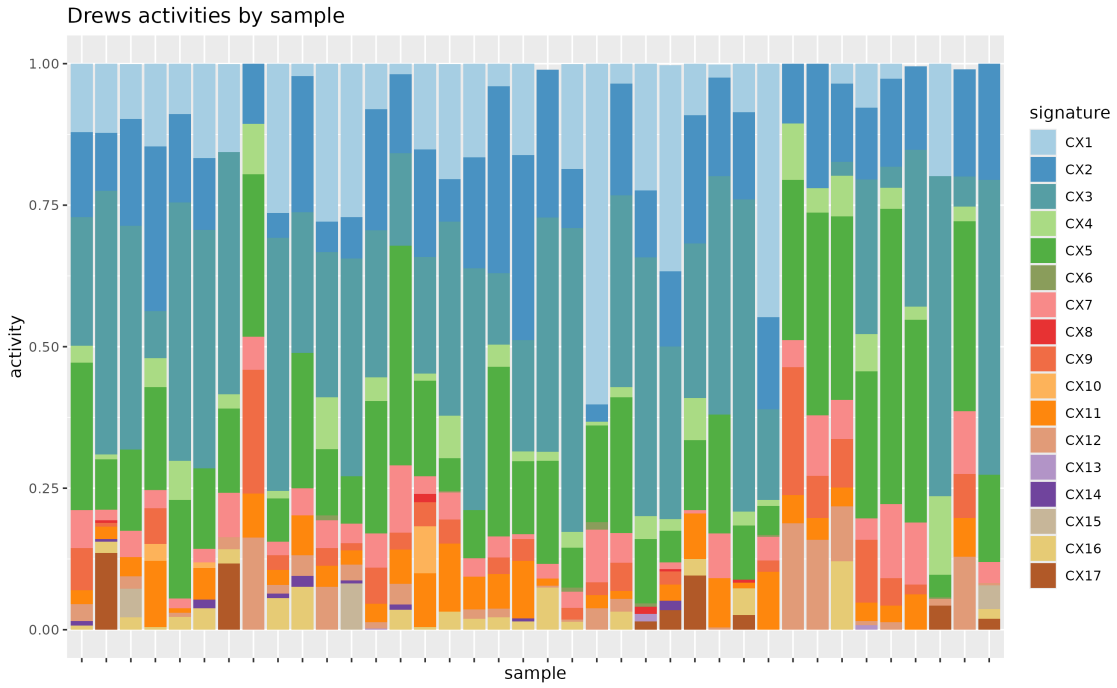


Figure 2.4: Activities of Drews signatures in our cohort. Each bar corresponds to a patient, and the colored segments represent the (normalized) activity of each signature.

2.2.3 Steele et al. signatures extraction

We also extracted the CN signatures described in [7], which form the COSMIC CN signature database [17].

We used *SigProfilerMatrixGenerator* (version 1.2.25) [18] to extract the components from our mutational profiles, which are based on three characteristics: length($0-100$ kb, 100 kb-1 Mb, $1-10$ Mb, $10-40$ Mb and >40 Mb), heterozygosity state (heterozygous, loss of heterozygosity and homozygous deletion) and absolute copy number (0, 1, 2, $3-4$, $5-8$ and >9 absolute CN), for a total of 48 bins.

After obtaining this $patient \times component$ matrix, we used *SigProfilerAssignment* (version 0.1.4) [19] to derive activities for the 25 signatures.

The most active signatures in our dataset are CN9, CN17 and CN20. In [7], the authors also propose aetiologies:

- CN9 is linked with chromosomal instability in a mostly diploid genome;
- CN17 is found to be associated with homologous recombination deficiency and tandem duplication;

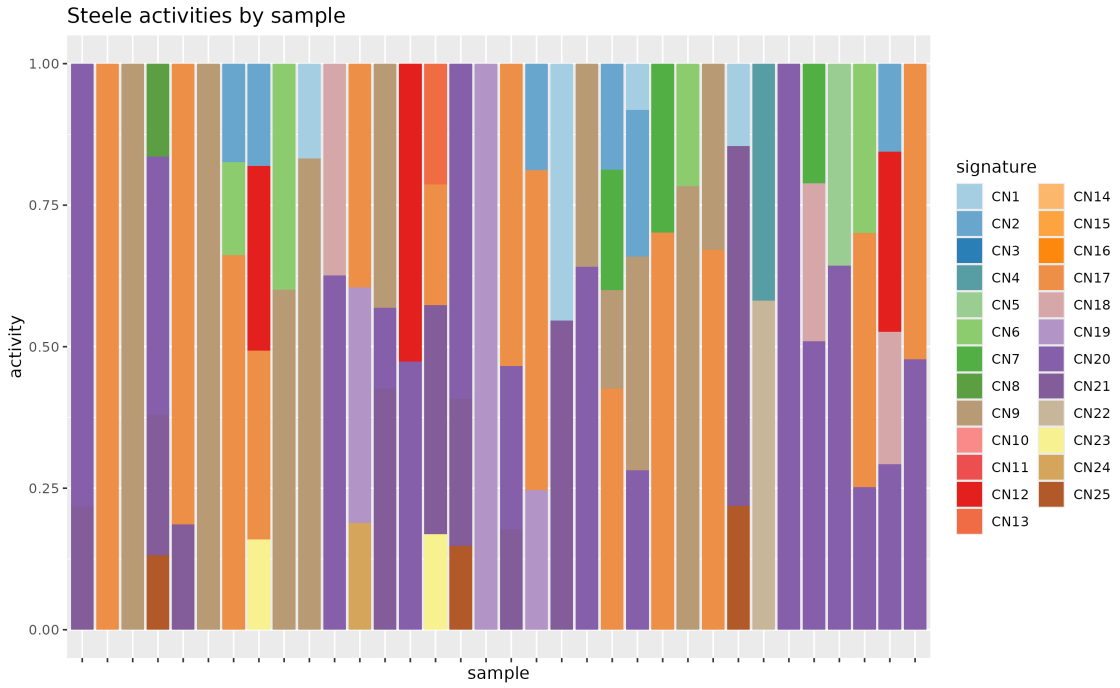


Figure 2.5: Activities of Steele signatures in our cohort. Each bar corresponds to a patient, and the colored segments represent the (normalized) activity of each signature.

- While CN20 has no proposed aetiology, it is associated with one-time genome duplication

2.2.4 Tao et al. signatures extraction

The last CN signature extraction method we considered is described in [8].

Their design of the components considers four features for the binning of the segments: the context (that is to say, the shape formed by the segment and its neighbors, for a total of 6 segment shapes), absolute CN (0, 1, 2, 3, 4, $5-8$ and ≥ 9), LOH status and size (<50 kb, $50-500$ kb, 500 kb-5 Mb, >5 Mb), for a total of 176 bins.

We used the R package *sigminer* (version 2.3.0) to calculate components and impute signature activities.

The most active signatures in our dataset are:

- CNS2, which has no known aetiology;

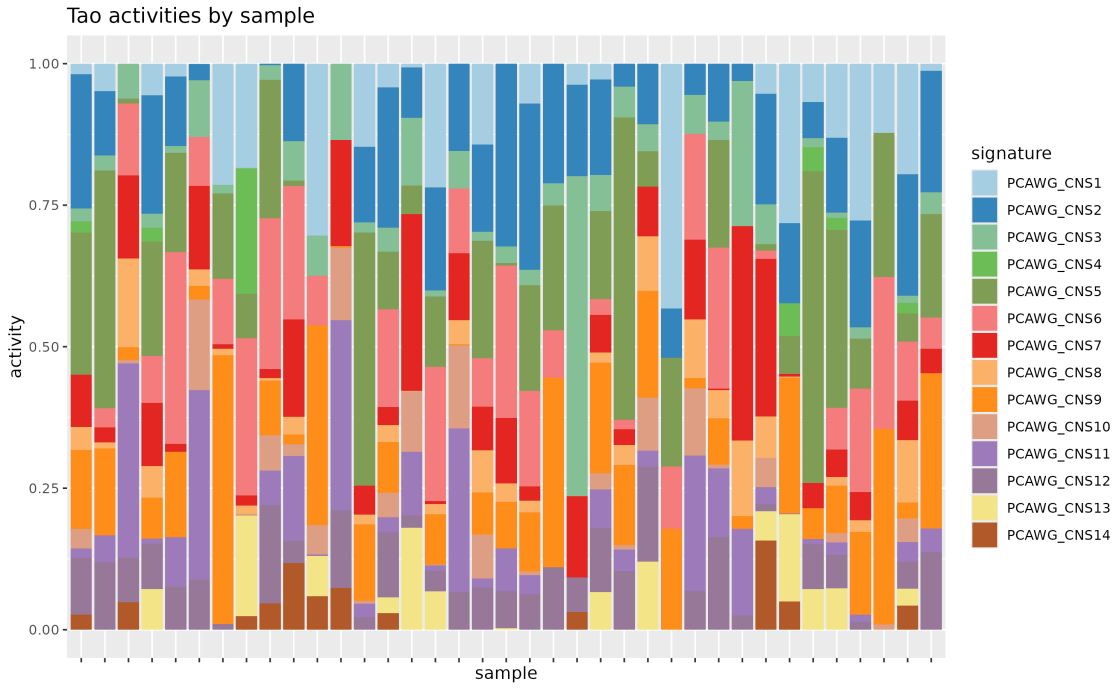


Figure 2.6: Activities of Tao signatures in our cohort

- CNS5, which has breakage-funsion-bridge as a probable cause;
- CNS6, correlated with loss of heterozigosity and amplification;
- CNS9, caused by whole genome duplication events.

While the three repertoires of CN signatures do not have clear correspondences between each other, there are recurrent mutational processes highlighted as aetiologies of highly active signatures: both the Drews and Steele signatures show homologous recombination depletion as an important mechanism in our mutational profiles, and all three methods estimate as highly present a signature correlated with whole genome duplication.

2.3 Patient stratification by mutational signatures

In order to quantify the effects of mutational signatures on the TME, we clustered patients based on the extracted signature activites, choosing the number of clusters between 2 and 8 corresponding to the maximum average silhouette score. We also looked at the difference in signature activities between Unstable and Highly Unstable patients.

2.3.1 Patterns of Drews signatures activities

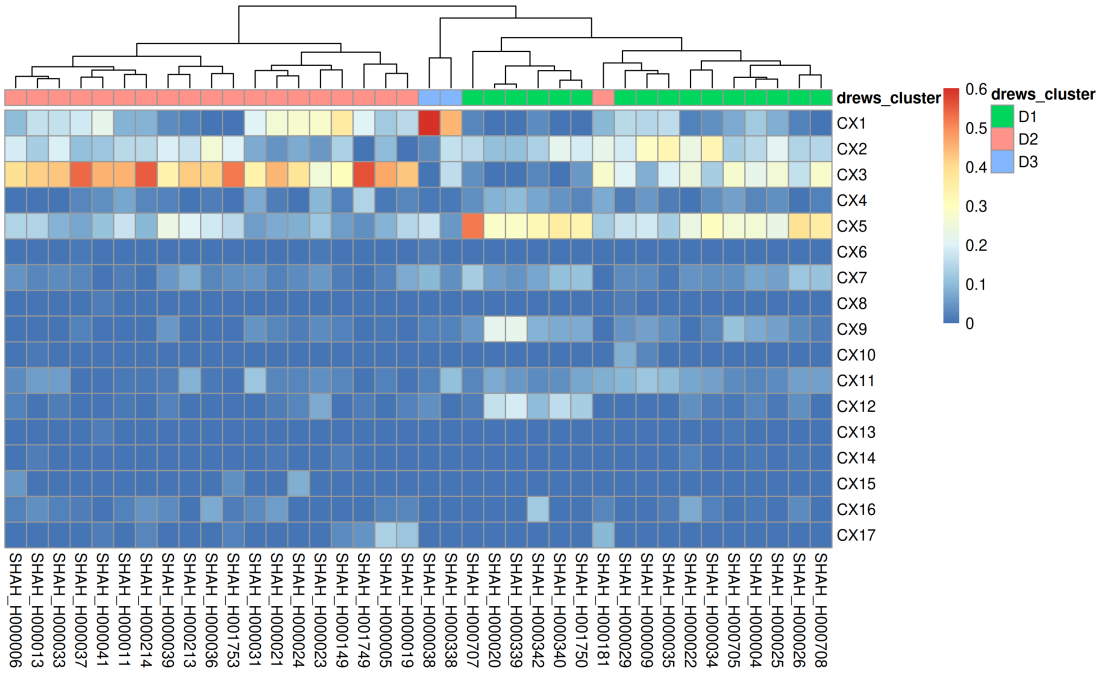


Figure 2.7: Clustering of patients based on Drews signatures activities.

Concerning the signatures we extracted in section 2.2.2, silhouette analysis identifies the optimal number of clusters at 3. These groups are cleary visible in figure 2.7. We identified the differences in activites across the clusters, which we have reported in the appendices.

Excluding the two-sample cluster of outliers in CX1 activity, we are left with two clusters: one of 16 patients, characterized by higher activity of signature CX5, which we termed D1, and the other comprising 20 patients, showing higher presence of signature CX3, and CX1 to a lesser extent, called D2. Considering their proposed underlying mutational mechanisms, we can hypotesize that patients in D1 show a higher degree of impaired homologous recombination, to the point of depletion [6].

We also found that HU and U patients clearly separate based on the activity of signature CX1, which is coherent with its characterization as a signature of whole-arm or chromosome changes [6]. Tables detailing the comparisons in Drews signatures' activities between groups of patients can be found in table A.2.

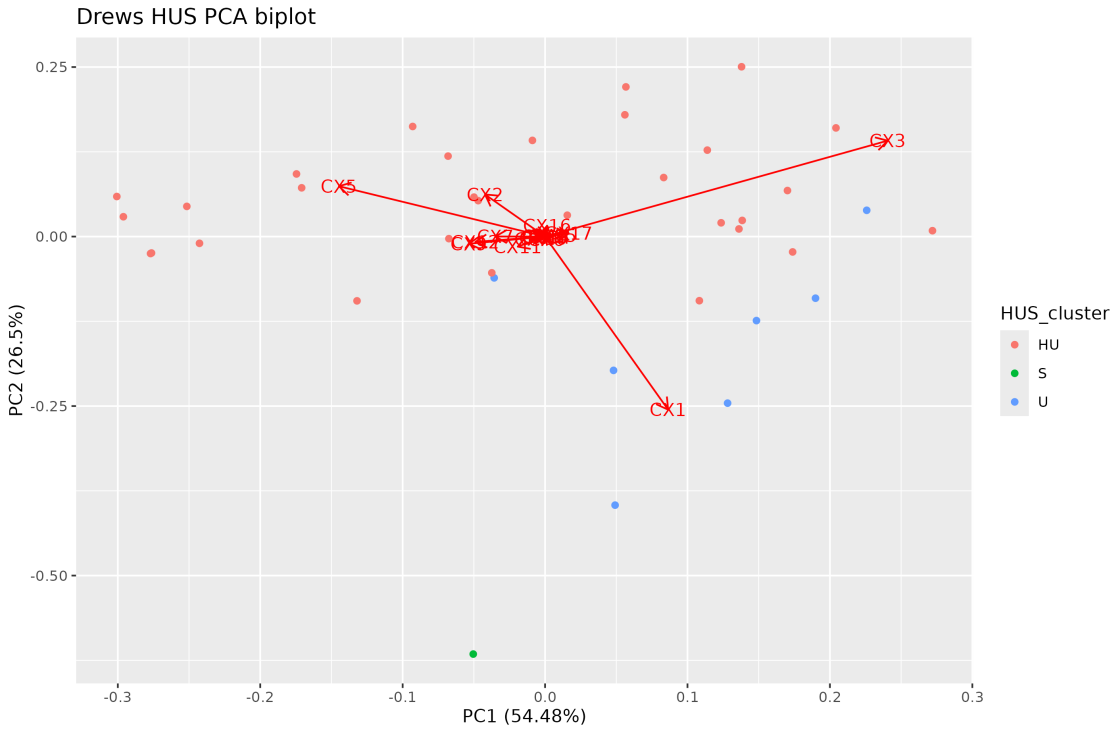


Figure 2.8: PCA biplot of Drews signatures activities, with patients colored by S/U/HU classification.

2.3.2 Patterns of Steele signatures activities

Concerning Steele signatures, the resulting $patient \times signature$ matrix of activities is very sparse. Furthermore, 3 patients were excluded due to the poor quality of the reconstruction of their mutational profile. Based on these factors, we decided to not investigate these signatures further.

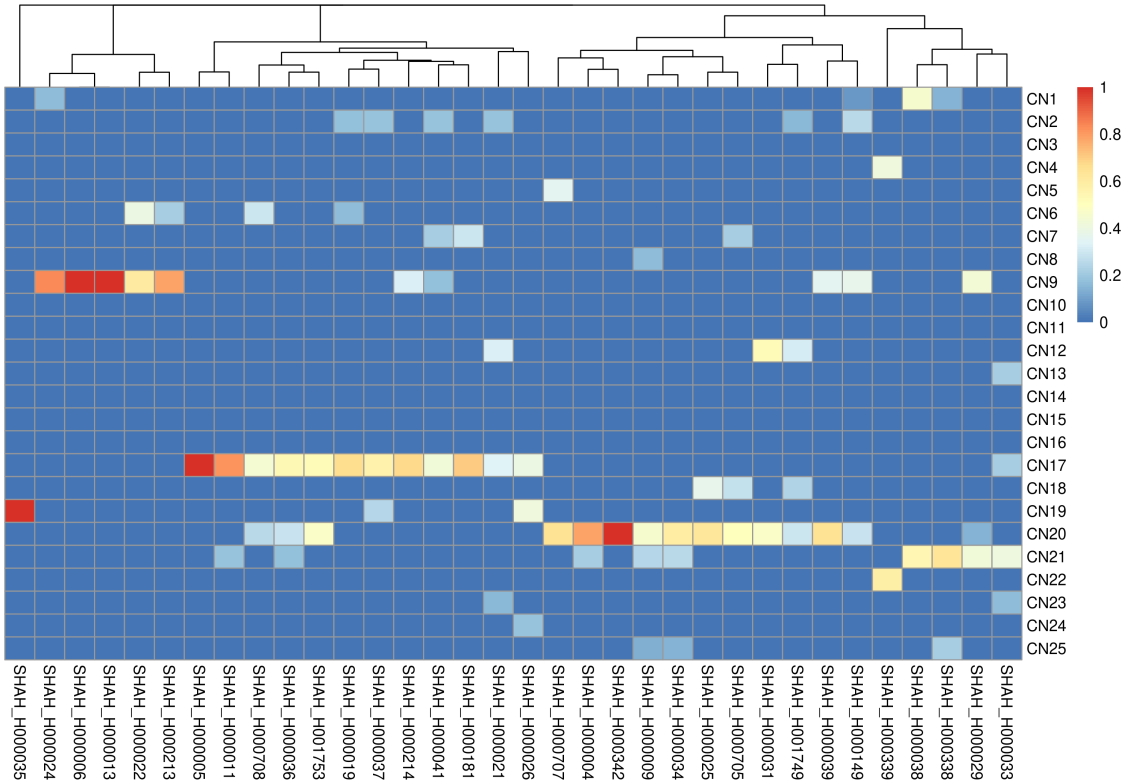


Figure 2.9: Clustering of patients based on Steele signatures activities

2.3.3 Patterns of Tao signatures activities

Silhouette analysis finds at 2 the optimal number of clusters. This leaves us with a cluster of 27 patients, termed T1, enriched in activities of signatures CNS5 and CNS9 and slightly in CNS1, with the remaining 10 patients grouped in T2 and enriched in CNS7, CNS11, and to a lesser extent in CNS3 and CNS10. Looking at proposed aetiologies [8], we can characterize T1 genomes as influenced by WGD and breakage-fusion bridges, and T2 by chromosome fragmentation and haploidity. Tables with the results of these comparisons can be found in table A.1.

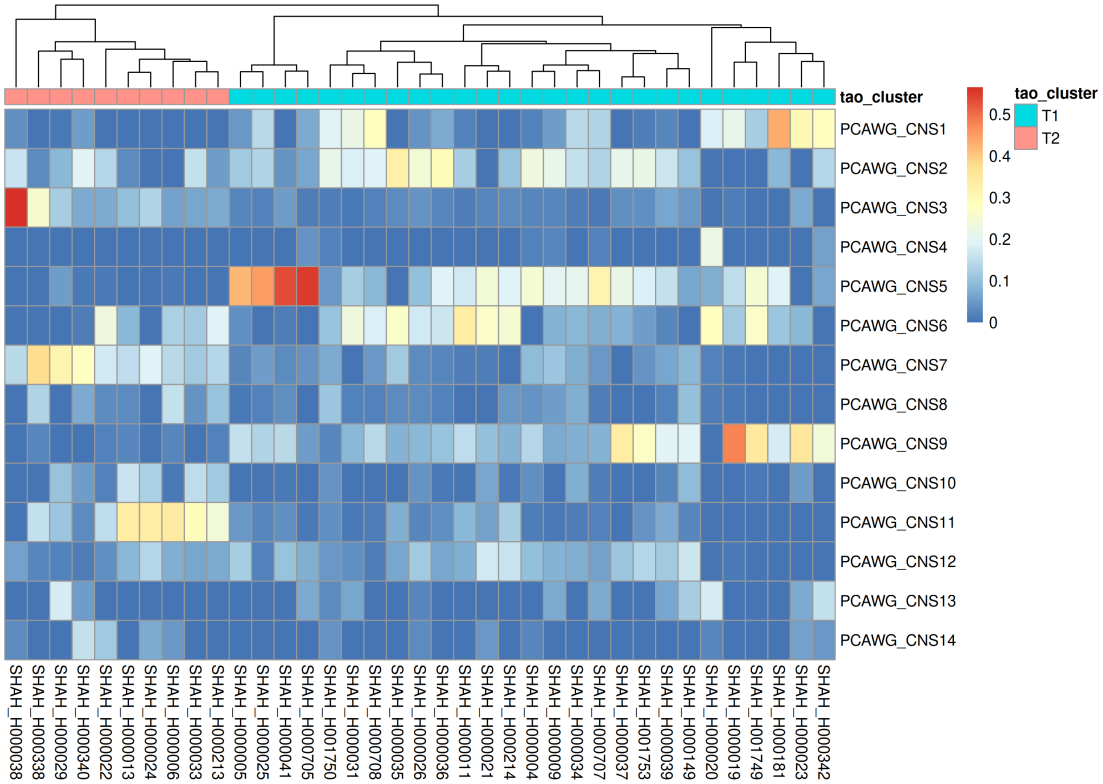


Figure 2.10: Clustering of patients based on Tao signatures activities

Having inferred CN mutational signatures from the WGS data and stratified the patients based on them, we now want to explore the differences between these groups in the TME through the scRNA-seq data. However, before that, we need to process the data in order to make it suitable for downstream analysis. We will explore these steps in the next chapter.

3

Processing of scRNA-seq data

Before scRNA-seq data is suitable for downstream analysis, it needs to go through various analyis steps, such as quality control, normalization, clustering and cell type annotation, which are the subjects of this chapter.

We processed our data following the recommendations of [20], a sistematic review of the steps necessary for the analysis of scRNA-seq data and the available tools.

The MSK dataset contains 283 samples from 40 patients, of which 134 were enriched in immune cells via flow-sorting, 143 in nonimmune ones, and 6 were left unsorted. As we introduced in 1.4, the samples are from different anatomical sites, ranging from primary lesions (adnexa) to ascites, peritoneum, omentum, upper quadrant and bowel samples.

We aligned our data using CellRanger (version 7.1.0) with standard parameters and the GRCh38 reference genome, obtaining $gene \times cell$ raw count matrices, which we then processed using the workflow described in this chapter.

3.1 Filtering, noise correction and normalization

scRNA-seq data analysis rests on the assumption that each droplet barcode corresponds to the RNA present in a single cell: however, this is often not the case, both due to the presence of ambient RNA and due to the possibility of a single droplet capturing more than one cell (this is often referred to as a "doublet") [20].

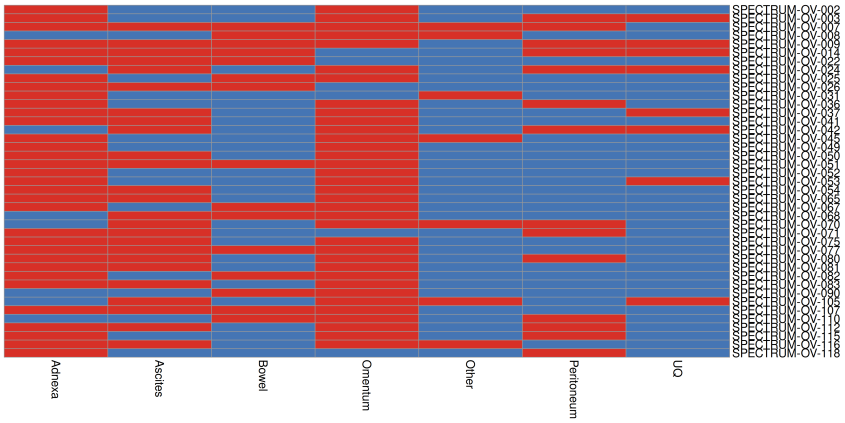


Figure 3.1: Anatomical sites of scRNA samples. For each patient (rows), cells in red are sites for which we have at least one sample

We used the R package *SoupX* (version 1.6.2) [21] to estimate the counts belonging to ambient RNA and remove them, followed by *scDblFinder* (version 1.18.0) to distinguish between doublets and proper single cells [22].

The filtered matrix was then loaded into a *Seurat* v5 object (version 5.1.0) in R. Next, we filtered out low-quality cells, which often represent dying cells [20]. Due to the high number of samples, we opted for automatic filtering using the perCellQCFliters function from the package *scuttle*, together with a filter for mitochondrial counts, only keeping cells where their fraction is $< 20\%$.

Once filtered, the data matrices were log-normalized, using size factors computed using the deconvolution method from the *scuttle* package.

We then merged these objects to create patient-level matrices and moved on to cell type labeling.

3.2 Dimensionality reduction, clustering and cell type labeling

We selected variable features, scaled and performed PCA and UMAP dimensionality reductions using the standard *Seurat* workflow.

Before clustering, we used *Harmony* (version 1.2.1) to correct the PCA embeddings across anatomical sites in order to account for technical variation between the samples.

After clustering, we performed cell type labeling using three automatic tools: *CellAssign*,

CellTypist and *scType*.

3.2.1 Tools for cell type labeling

The assignment of cell types to clusters or individual cells is a fundamental step in scRNA-seq analysis, necessary for the biological interpretation of the data.

In this thesis, after merging all samples from the same patient into individual *Seurat* objects, we annotated this patient-level objects separately, using two tools based on gene signatures (CellAssign [23] and scType [24]) and a pre-trained logistic classifier (CellTypist [25]).

Concerning the marker-based methods, they assign cell types based on a user-provided list of marker genes, which are genes whose expression is specific to one (or more) specific cell types. CellAssign was by far the better performing, while scType often left a significant number of cells unlabeled, which led us to ignore its labeling in further analyses.

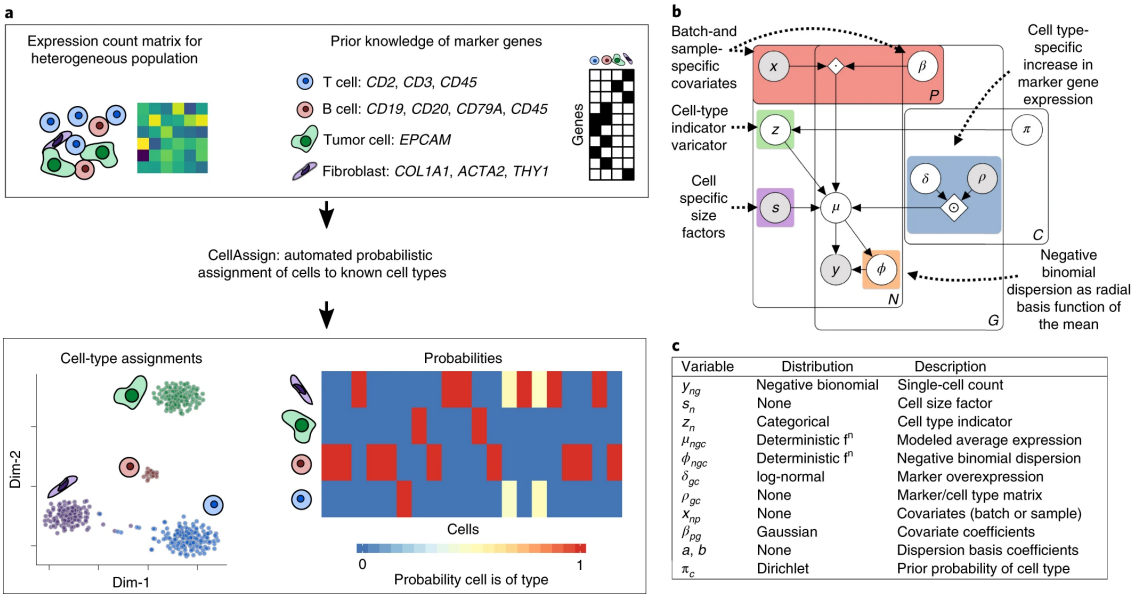


Figure 3.2: a: Starting from raw count data and prior knowledge on marker genes, cellassign returns a probabilistic cell type labeling matrix. b: Structure of the statistical model implemented c: Distributions of the parameters involved. Figure from [23]

Let Y be the $N \times G$ cell-by-gene matrix of raw counts, and $\mathbf{z}$ the vector encoding cell types ($z_n = c$ iff cell n is of type c). CellAssign performs statistical inference for the quantities $p\left(z_n = c|Y, \hat{\Theta}\right)$ using expectation maximization.

The model is

$$\mathbf{E}[\, y_{ng}|z_n = c] \; = \mu_{ngc} \tag{3.1}$$

where, given s_n vector of size factors for cells and X $P \times N$ matrix of covariates,

$$\log \mu_{ngc} = \log s_n + \delta_{gc}\rho_{gc} + \beta_{g0} + \sum_{p=1}^{P} \beta_{gp} x_{pn} \tag{3.2}$$

δ_{gc} represents the expected fold change for gene g in cells of type c.
The likelihood is then given by

$$y_{ng}|z_n = c \approx \mathcal{NB}\left(\mu_{ngc}, \tilde{\phi}_{ngc}\right) \tag{3.3}$$

with NB the negative binomial distribution. A more detailed representation of the model can be found in figure 3.2.

We used CellAssign on each patient-level object, with the list of marker genes provided by the authors of [13].

Concerning instead CellTypist, it is a collection of multinomial regression models pre-trained on high quality data. We used two approaches: a coarse-grained model to assign cell types to our clusters, and a more fine-grained one to analyze in more detail the T cells cluster. In both cases, we clustered the dataset at a high resolution in order to implement a majority voting procedure, where cells were assigned the most frequent label of cells in their resulting subcluster.

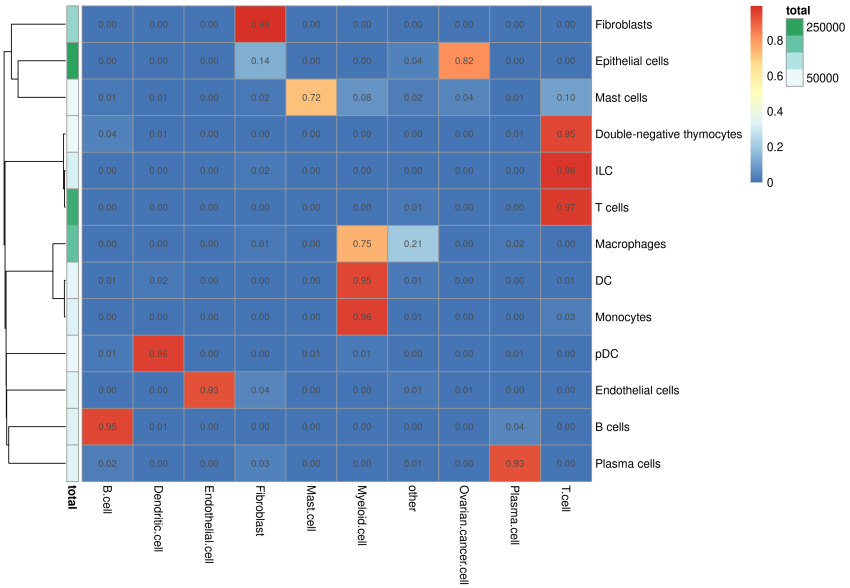


Figure 3.3: Agreement between CellTypist (rows) and CellAssign (columns). The heatmap represents the cross table between the two annotations, with rows divided by their total number of cells. The rows metadata is the original row sums.

As can be seen in figure 3.3, the two tools have a very high level of concordance. The main differences are that, since we used a model intended for the classification of immune cells, CellTypist can not distinguish between tumor and normal cells, and is instead more specific for the T cells and myeloid cells clusters.

We used the annotation provided by CellAssign as a starting point, refining it with the subclusters identified by CellTypist. The identified cell identities can be divided into 3 groups:

- The tumor cells;
- The immune fraction, with adaptive (T/B lymphocytes), and innate (a large cluster of myeloid cells, with clear subclusters of macrophages and dendritic cells, and innate lymphoid cells in the T cluster) fractions;
- The stromal component of the TME, consisting of endothelial cells and fibroblasts

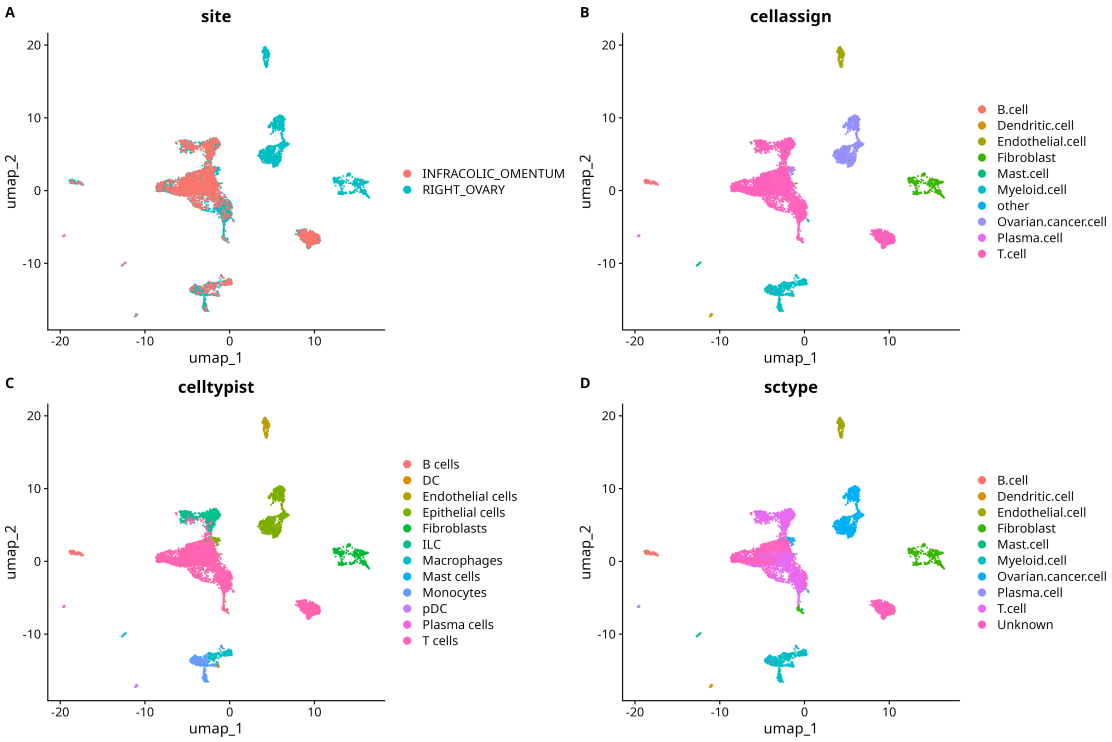


Figure 3.4: UMAP of patient level object, after cell type labeling. **A**: Cells annotated by anatomical source. **B**: Cells annotated by cellassign. **C**: Cells annotated by celltypist. As we can see, celltypist can not label tumor cells, but is more detailed on immune ones. **D**: Cells annotated by sctype

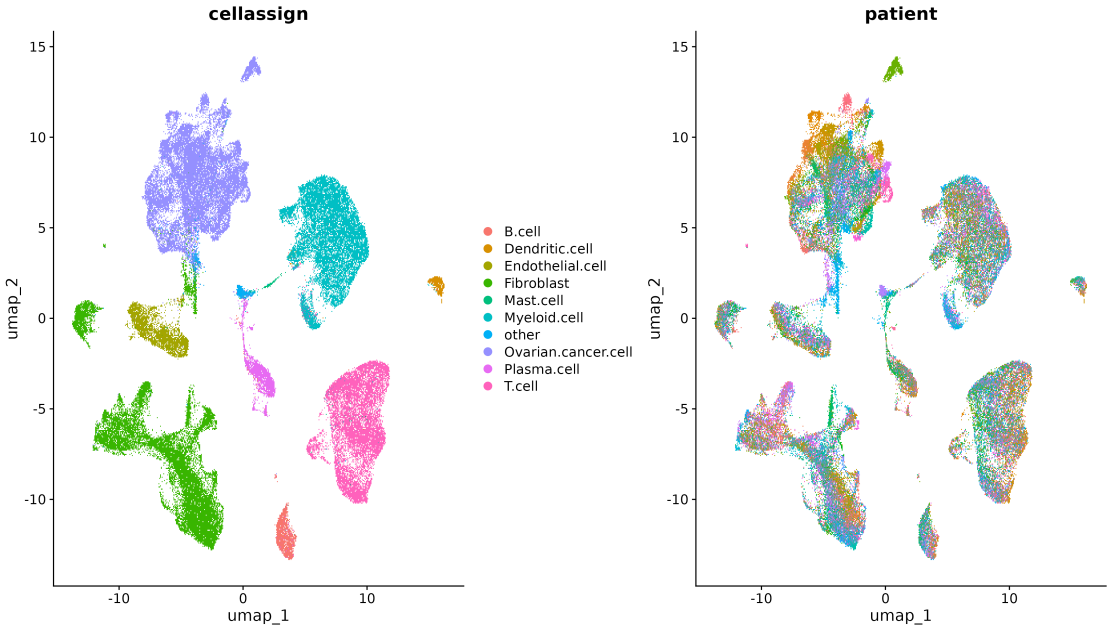
3.3 Merging into global object

We then merged all patient-level objects into a single Seurat dataset, using BPcells to transform the matrices to on-disk. In total, all our scRNA-seq data contains 1050665 cells by 28737 genes. A more detailed table with the number of cells for each patient and anatomical site can be found in table A.3.

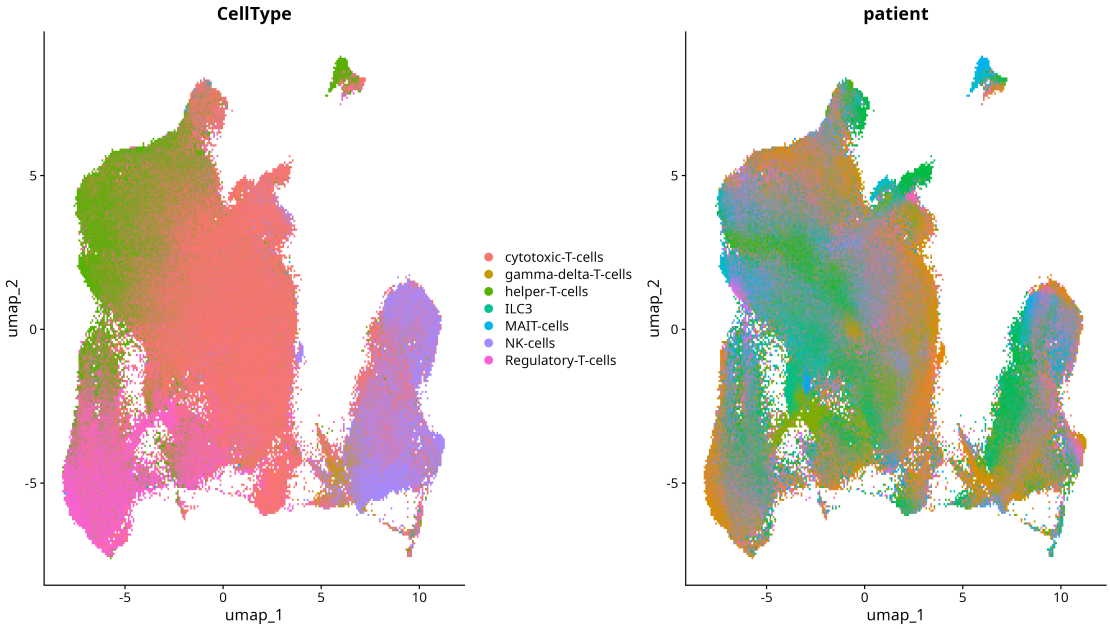
Inspired by [13], we also used CellTypist to re-assign labels to the T cell cluster, identifying cell identities which we aggregated into 7: 4 abundant classes corresponding to regulatory, cytotoxic, helper T cells and NK cells, together with 3 rarer populations of gamma-delta T cells, MAIT cells and Innate Lymphoid Cells.

We filtered out 3 low quality samples (identified by CellRanger metrics and fraction of cells left unlabeled) and those annotated with 'other' as anatomical site.

Finally, we integrated the stratifications of patients defined in chapter 2 with the scRNA-seq data, and prepared to evaluate the differences between these groups of patients in this processed data set. The first analysis we carried out, concerning differences in cell types abundances, is the subject of the next chapter.



(a) All clusters



(b) T cells subcluster

Figure 3.5: UMAP plots of the entire scRNA-seq dataset

4

Impact of mutational signatures on cell types abundances

After extracting mutational signatures from the WGS data and processing the scRNA-seq data, we are ready for downstream analyses.

In order to explore the effects of CNV on the tumor microenvironment, we will compare the groups defined in the previous chapter in terms of cell type composition of the TME, gene expression and cell-cell communication

The first analysis we carried out concerns changes in abundances of different cell types. Their identification requires ad hoc tools, since the limited number of cells captured by scRNA-seq forces us to consider these abundances as compositional data, with an inherent negative correlation between the different cell types.

In section 4.1, we introduce the scCODA tool for differential abundance analysis, its structure and assumptions, and the design of our model. We will then comment on our results in section 4.2.

4.1 Inferring differential abundances using scCODA

scCODA [26] is a Bayesian model, based on the multinomial Dirichlet distribution.

Let X be an $N \times M$ matrix of covariates, and Y a $N \times K$ matrix of cell counts for N

samples, M covariates and K cell types. The impacts of the covariates on cell counts $\gamma_{m,k}$ are then modeled individually using normally distributed effects $\gamma_{m,k}$, together with a logit-normal prior for feature selection, that is to say:

$$Y \sim DirMult(\phi, \bar{y}) \tag{4.1}$$

Using a log link function to model ϕ, we have

$$\log(\phi) = \alpha + X\beta \tag{4.2}$$

β is the matrix of effects we want to estimate. As we said, feature selection is implemented by a logit-normal prior:

$$\beta = \tau\tilde{\beta} \tag{4.3}$$

With τ logit-normal, with mean 0 and standard deviation 50.

$\tilde{\beta}_{m,k}$ is further defined as the product of a Half-Cauchy, covariate-specific variable (which acts as a scaling factor) and the normally distributed $\gamma_{m,k}$. A summary of the model can be found at figure 4.1

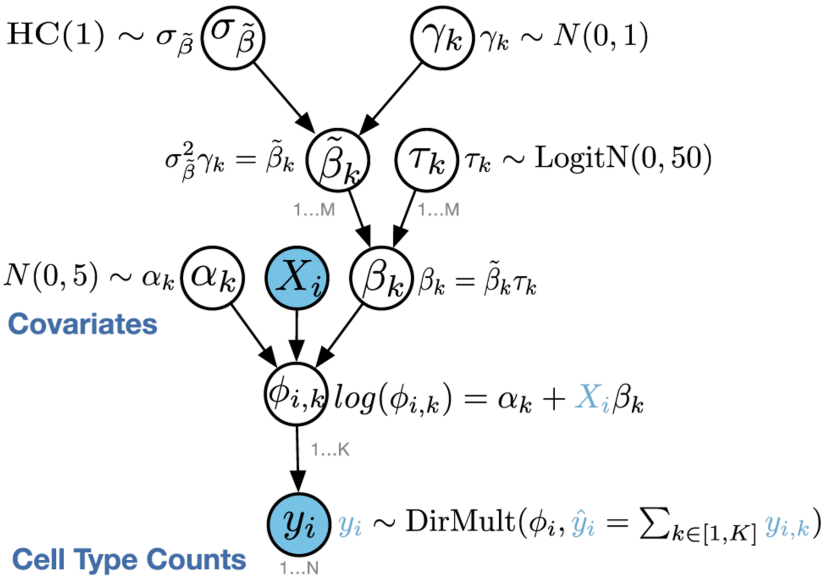


Figure 4.1: Structure of scCODA model. X_i and y_i, in blue, are the covariates and observed abundances respectively. The "roots" of the graph ($\sigma_{\tilde{\beta}}$, γ_k, τ_k and α_k) are estimated via HMC sampling. Figure from [26]

Parameters $\alpha_k, \gamma_{m,k}, \sigma_m, \tau_k$ are estimated using Hamiltonian Monte Carlo, with No U-

Turn Sampling over 11000 iterations, of which 1000 are discarded as burnout.

scCODA also performs false discovery rate (FDR) control, by using the fraction of iterations in which a certain parameter is $< 10^{-3}$ as an estimate of the probability of type 1 error.

In our case, after creating $samples \times celltype$ abundance matrices (one for the general clusters and one for the T / NK subgroups), we set up our analysis the following way: we used the extracted signature activities as covariates, only considering those that had activity > 0 in at least half of our patients. We included the age of the patient and the anatomical source of the sample as additional covariates to account for their impact on cell abundances. Then our matrix of covariates X can be described as $S_1|S_2|a$, with S_1 a block matrix $samples \times signatures$, where each sample is assigned the mutational signature activities extracted from the WGS data of the corresponding patient, S_2 a $samples \times site$ boolean matrix encoding the anatomical source of the sample, and a a block vector of ages for the patients corresponding to the samples, after Min-Max scaling to bring all covariates to the same scale and avoid biases due to the spike-and-slab prior.

Concerning the choice of reference cell type, it needs to be a sufficiently common type to be present at a sufficient level in most samples, with a relatively constant abundance. We used scCODA to create an abundance-dispersion plot, which we used to select dendritic cells as our reference for the general clusters and cytotoxic T cells for the T cells subcluster (see appendix A.2).

4.2 Differential abundances in the MSK dataset

When discussing our results, we must keep some complications in mind:

- The coefficients of the model are not directly interpretable. Still, they will allow us to at least identify credible positive or negative effects of mutational signatures on cell type abundances, even if we can not precisely quantify them;
- All our results are relative to a reference cell type (as we said, dendritic cells for the general clusters and cytotoxic T cells for the T/NK subclusters), which will be assumed to be invariant with respect to the covariates;
- The extracted activities of mutational signatures are also compositional, as they are normalized so that for each patient, the sum of the activities of all signatures is 1. This leads to an inherent collinearity, which we can mitigate by performing feature selection, both a priori (by only considering signatures present in at least half our patients), and with the spike-and-slab prior in-built in scCODA.

We used scCODA separately for the different types of signatures we discussed in chapter 3, and separately for the general clusters and T/NK subclusters. Since a FDR threshold of 0.05 proved too restrictive, we relaxed it to 0.1.

4.2.1 HU signatures

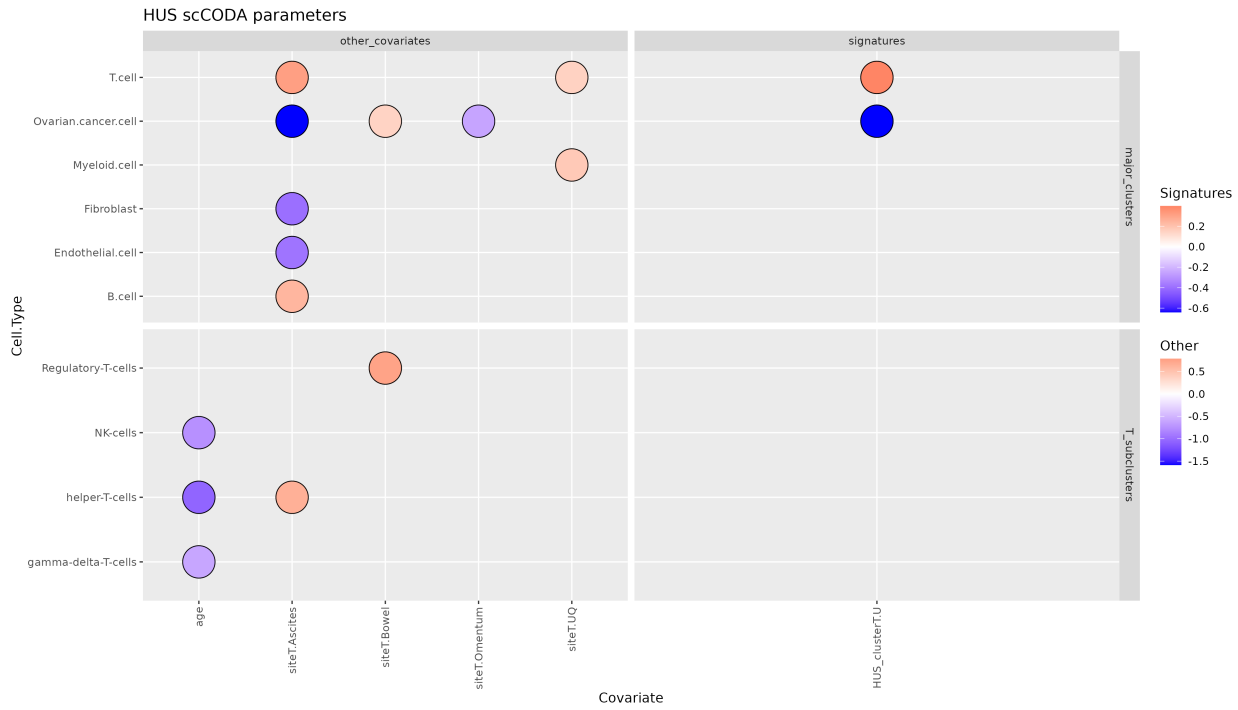


Figure 4.2: Effects identified by scCODA of HU signatures on cell type abundances. The top and bottom row of plots refer to the two separate models, fit to the general clusters and T/NK subclusters respectively. On the left, we have the coefficients of the other covariates of interest (age and anatomical site), while on the right are those relative to the mutational signatures. The color represents the value of the inferred parameter. Only the parameters identified as credible under a FDR threshold of 0.1 are shown.

scCODA identifies the T/NK cells cluster and cancer cells as differentially abundant between HU and U patients, with samples from U patients having a comparatively lower percentage of tumor cells and a higher one of T cells. This is coherent with our expectations, as HU patients have poorer outcomes [16]. However, there are no effects on the T/NK subclusters.

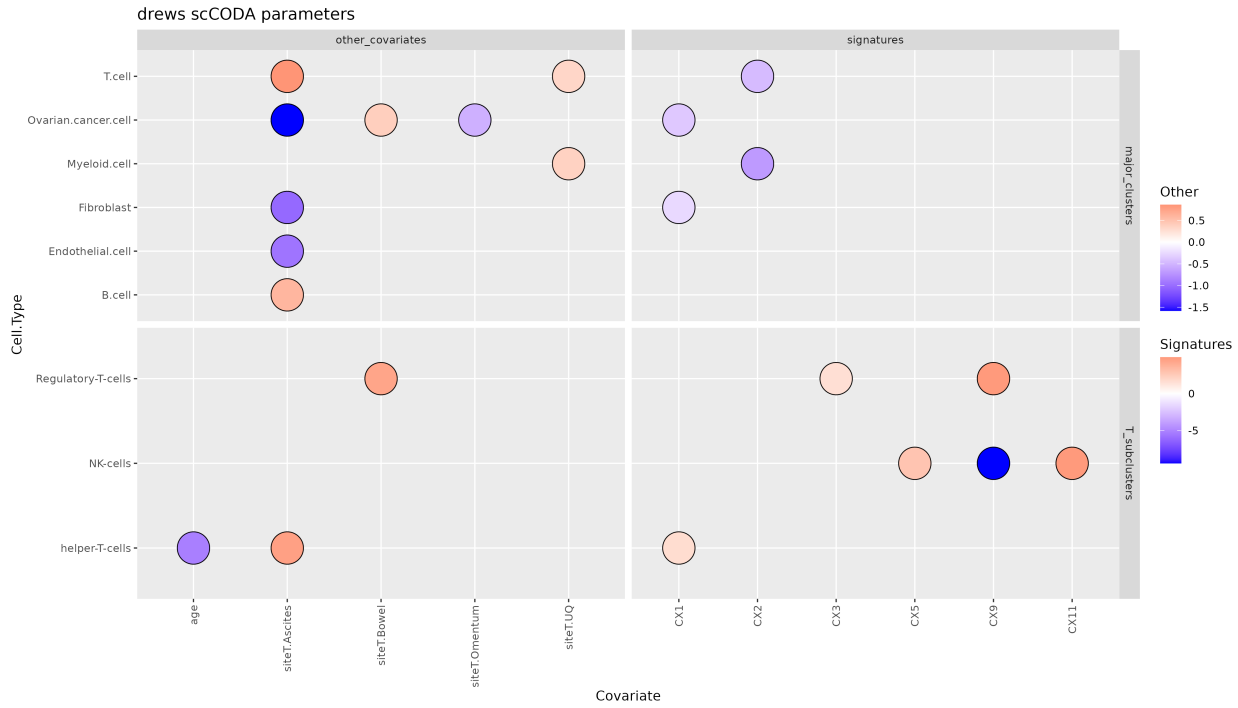


Figure 4.3: Effects identified by scCODA of Drews signatures on cell type abundances. The figure can be interpreted in the same way as figure 4.2.

4.2.2 Drews signatures

The first effects of immediate interest are those of CX1 (which, as we pointed out in chapter 2, is related to chromosome missegregation), which has a negative effect on tumor cells and fibroblasts, and a positive one on helper T cells. This further confirms our results concerning the HU mutational signature, since U patients were also characterized by higher levels of CX1.

Other notable effects identified by scCODA on the large clusters include a negative effect of CX2 on T/NK and myeloid clusters.

Looking at signatures with effects on the T cells subclusters, we find CX3 and CX5, two signatures with apparently opposite effects associated with IHR, with CX3 being associated with an increased level of regulatory T cells (which promote immune tolerance and immunosuppression) and CX5 with NK cells. Similarly, we can identify two signatures (CX9 and CX11) associated with replication stress and amplifications, where CX9 is associated positively with regulatory T cells and negatively with NK cells, which are instead enriched when CX11 has higher activity.

We also compared the patient clusters we identified in section 2.3.1, finding that cluster D2

patients, which we characterized by a higher degree of impaired homologous recombination, have a lower abundance of cancer cells and fibroblasts (see figure A.3).

4.2.3 Tao signatures

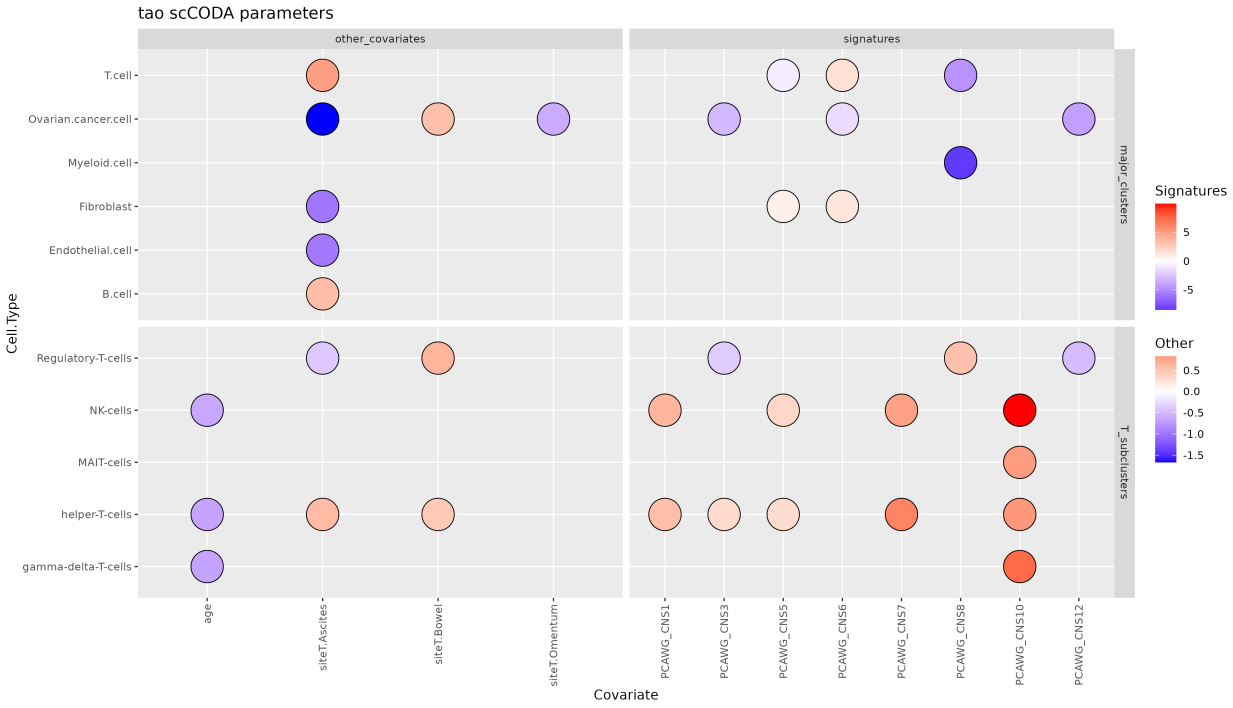


Figure 4.4: Effects identified by scCODA of Tao signatures on cell type abundances. The figure can be interpreted in the same way as figure 4.2.

Looking instead at Tao signatures, scCODA identifies many more effects. For simplicity, we will only consider the signatures that impact both the general clusters and the T/NK subclusters.

- CNS3 activity is correlated with higher fractions of helper T cells, and a reduced abundance of regulatory T cells and tumor cells. Looking at the proposed aetiologies in [8], CNS3 activity indicates a genome less affected by CNA, which might explain the possible reduced proliferation and immuno suppression;

- CNS5 (which has breakage-fusion bridges as a probable cause) is correlated with a higher fibroblast fraction, together with a lower T cell abundance and higher presence of NK cells and helper T lymphocytes;

- CNS8, which has no proposed mechanism, seems to lead to higher degree of immune suppression, with a lower fraction of lymphoid and myeloid immune cells, together with a higher presence of regulatory T cells in the T/NK subclusters;
- On the contrary, CNS12 (related to triploid chromosomes) is correlated with lower fractions of tumor and regulatory T cells, suggesting an opposite effect

Comparing instead the clusters we defined in section 2.3.3, T2 patients have a comparatively larger fraction of helper T and NK cells, but no credible effect at the level of the general clusters (see figure A.4).

Summarizing the results we just discussed, there are two apparent groups of signatures, with some leading to a higher presence of tumor and regulatory T cells, which might be indicative of a greater degree of immuno suppression and tumor proliferation(such as Drews signature CX2 and Tao signature CNS8), while others have the opposite effect, being correlated with higher fractions of immune cells and helper T cells. Interestingly, in multiple cases genomic patterns of whole arm/chromosome changes lead to reduced proliferation and immuno suppression, as seen in the comparison between U and HU patients, in the Drews CX1 signature and the Tao CNS12 one. However, there is no equivalent clear group for signatures with the opposite effect, as their proposed aetiologies range from impaired homologous recombination (CX2) to breakage fusion bridges (CNS5). Some signatures also have no proposed underlying mechanism, or do not fit clearly in one of the two groups.

4.2.4 Anatomical sites and age

Concerning the other covariates in our model, there are a number of high confidence effects of age and anatomical site, which are coherently identified by scCODA in all the models we discussed so far: first of all, age seems to be associated with a smaller fraction of helper T cells in the T/NK cells cluster. Concerning instead the anatomical sites, first we must keep in mind that all the effects found are relative to adnexal (primary) samples as a point of comparison. Ascites samples are then found to have a lower fraction of cancer and stromal cells, and a higher presence of lymphocytes, specifically helper T cells, which is expected, as in ascites cancer cells are suspended in fluid. In contrast, bowel samples are recurrently identified as having higher fractions of tumor cells and regulatory T cells. Omentum samples are also found to contain a lesser fraction of tumor cells.

While analyzing the abundances of cell types has allowed us to identify some probable effects of mutational signatures on the TME, the results presented in this chapter give us no indication of the underlying molecular mechanisms that result in the changes discussed above in the cell type composition. Therefore, we further analyzed our data in terms of gene expression and cell-cell communication.

In our further analyses, we choose to focus on tumor cells, T/NK immune cells, and fibroblasts: fibroblasts are known to play an important role in tumor progression, tumor propagation, and growth signaling [9], and analyzing the lymphoid cells mentioned above will allow us to further investigate the effects of CNV on the immune fraction of the microenvironment. We will not further consider gamma-delta T cells, MAIT and ILC3 cells, as they are too rare to form consistent subpopulations across samples. This leaves us with 6 cell populations of interest: cancer cells, fibroblasts, NK cells and helper/cytytoxic/regulatory T cells.

We will also no longer consider the activities of mutational signatures as continuous variables, limiting ourselves to the analysis of the clusters defined in Section 2.3, both for ease of interpretability of our results and to avoid issues of collinearity.

5

Mutational signatures and gene expression

While in the previous chapter we identified effects of mutational signatures on the cell type composition of the TME, the results we obtained give us no indication of the underlying processes that result in these changes. As such, we will now continue our analysis by investigating the changes in gene expression induced by copy number variation, in order to further explore the results we obtained in the previous chapter.

As we explained in section 4.2, we will focus on lymphoid cells, tumor cells and fibroblasts. We will also consider primary and metastatic samples separately, in order to account for the possibility of different effects of mutational signatures between primary lesions and metastases.

In total, we carried out 6 comparisons (primary or metastatic samples, for the three different groupings of patients based on HU, Drews or Tao signatures), each considering 6 cell types: tumor cells, fibroblasts, NK cells and cytotoxic/helper/regulatory T cells.

In section 5.1, we explain how we designed our differential expression analysis, and how we interpreted the results using gene set enrichment analysis. We then go over the main alterations in pathway and biological process activity we identified in section 5.2

5.1 LIMMA and mixed linear model fitting

In scRNA-seq data, treating each cell as an independent observation leads to a greatly overestimated number of degrees of freedom and FDR, due to the inherent correlation between transcriptional profiles of cells from the same sample. As such, our strategy was instead to build

pseudobulk profiles for each cell label and sample by aggregating the counts of each cell in the group, and apply tools for bulk RNA-seq analysis to the created profiles.

We aggregated RNA counts by cell type for each sample separately, considered only profiles consisting of at least 10 cells as relevant, and split our analysis between primary and metastatic samples, in order to account for the possibility of different effects in primary lesions and metastases.

For a statistically sound analysis, we need to account for a number of complicating factors:

- Samples from the same patient are inherently correlated, leading to a hierarchical structure in our data;
- When analyzing metastatic samples, different anatomical sites may have an effect on gene expression;
- Age is known to impact the expression of many genes.

In order to account for these factors, we used the R package *LIMMA* for our analysis. The pipeline we built for DGE involves the following steps:

1. First, we choose the comparison (primary or metastasis samples, HUS or Drews or Tao stratification) and cell type of interest (cancer cells, fibroblasts, cytotoxic/helper/regulatory T cells, NK cells), and build our matrix of pseudobulk gene expression. This means that in total we have 6 comparisons, each involving 6 cell types;
2. Lowly expressed genes are filtered using the *filterbyExpr* function from *edgeR*
3. For normalization of the library sizes, we use the *TMMwsp* method from R package *edgeR*, a modified version of trimmed mean of M values normalization better suited to zero-inflated data;
4. *voom* is then used to logCPM-transform the data, estimate the mean-variance relationship and compute precision weights for each gene, which are used to account for heteroscedasticity when fitting the linear model in the next steps;
5. A mixed linear model is then fit for each gene using *LIMMA*, with fixed effects for age, group stratification and anatomical site (if we are considering metastatic samples), and random effects for the variable encoding the patient. That is to say, the matrix for the fixed effects is $M_1|M_2|a$, with M_1 encoding the group stratification, M_2 the anatomical source of the samples, and a ages, similarly to our setup for scCODA, which we explained in section 4.1;

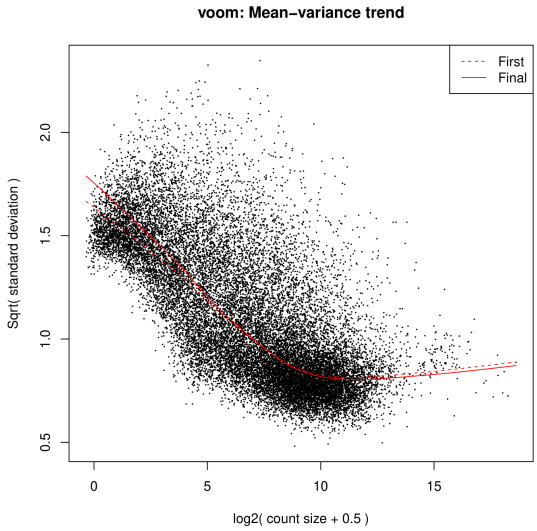


Figure 5.1: Mean-variance trend estimated by voom. The x axis is the logarithm of the total counts for that gene, while the y axis is its variance. As we can see, the variance sharply drops off as the average expression increases.

6. We also use the *duplicateCorrelation* function to account for the loss of residual degrees of freedom due to the inherent correlation between samples belonging to the same patient. Since we have many samples composed of varying amount of cells and including some outliers, we use the quality weights *voom* option to calculate, in addition to the observation weights, sample-level weights that are used to downweigh outliers in the linear model;

7. Finally, we identify effects of mutational signatures on gene expression by building contrasts between the coefficients relative to the two groups (for example, when considering the division of patients between U and HU groups, we build the contrast $U - HU$ between the coefficients corresponding to the two classes). We also apply Bayesian shrinking to the estimated parameters;

From this, we obtain a table of DGE results, including fold change, standard error, value of the test statistic, and raw and adjusted p values for each gene (FDR correction using the Benjamini-Hochberg procedure).

For cell types with at least 50 differentially expressed genes (adjusted p value < 0.1), we ranked our genes by the *LIMMA* test statistic, in order to obtain a ranked list of genes that we used as input for Gene Set Enrichment Analysis (GSEA). Given a ranked list of genes from DEA and a gene set, GSEA evaluates the enrichment of the gene set by going through the gene list and keeping a running score that increases whenever a gene belonging to the set is encoun-

gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
MRPL32	-1.466847	5.447096	-5.463510	1.895893e-06	0.03279684	4.9104128
PSMA2	-1.300438	7.056746	-5.333784	2.942872e-06	0.03279684	4.5100653
VAMP3	-1.389952	4.692118	-5.211415	4.448100e-06	0.03304790	4.1131679
ANKRD66	6.019519	-4.039852	5.008070	8.799914e-06	0.03471613	0.8064938
OR2A7	-2.339154	1.108194	-4.989942	9.349121e-06	0.03471613	3.0175283

Table 5.1: Top 5 differentially expressed genes for ovarian cancer cells, primary samples, comparison between U and HU patients. The columns represent the log fold change between groups, average expression of the gene, test statistic, raw and adjusted p values and the log-odds of differential expression for that gene.

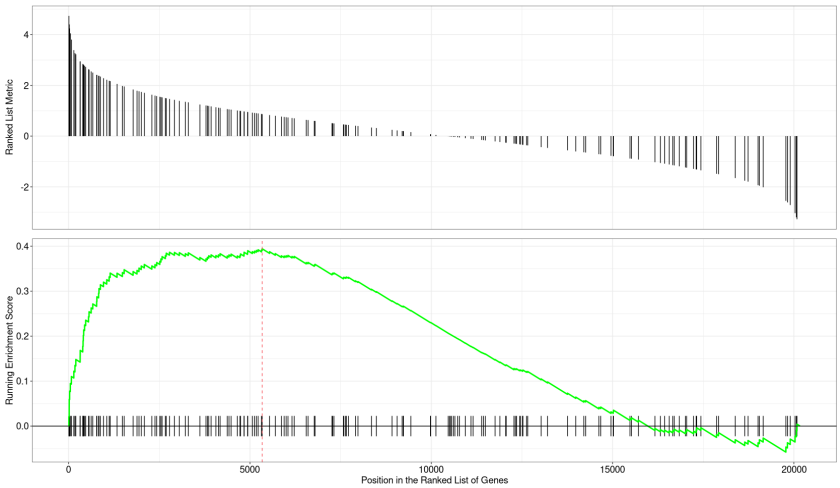


Figure 5.2: GSEA plots. The x axis represents the ranked list of genes. **top**: Distribution of the *LIMMA* test statistic over the gene set. **bottom**: Position of the gene set in the ranked list and running enrichment score (in green), with maximum value identified by the red line

tered, and decreases when not. The maximum (absolute) value is used as test statistic, with a null distribution generated by random permutation of the gene set in the ranked list. We collected gene sets from the MSigDb database. In particular, we used gene sets from its Hallmark collection, the Gene Ontology Biological Process collection and the Reactome pathway database. For each collection, we looked at the top 10 up and down regulated sets.

5.2 Enriched pathways in the MSK dataset

As we said in the previous section, we focus on comparisons that reveal at least 50 differentially expressed genes. This occurs for all cell types when comparing U and HU patients, and when comparing cancer cells in the D1 and D2 clusters. Comparing the T1 and T2 clusters, we found no cell type with significant changes in gene expression, which led us to exclude this comparison

from further analyses.

5.2.1 HU signature

We will start our discussion considering the differences between U and HU patients in primary samples. Plots with the top up and down regulated gene sets can be found at figure A.5

Across all cell types, HU patients show an increased expression of interferon-stimulated genes, which are known to play a role in anti-tumor immunity [27], together with a higher degree of inflammation, evidenced by the upregulated inflammation response program in multiple cell types and reduced production of anti-inflammatory cytokines.

In more detail:

- Cancer cells show upregulation in many processes and pathways related to cell growth and mitosis, such as mTORC1 signaling, targets of E2F and myc transcription factors, chromosome segregation during mitosis and DNA synthesis during the S phase. Other upregulated groups of genes include genes induced by the tumor necrosis factor a/ nuclear factor kB(TNFa/NF-kB) pathway and genes involved in protein secretion and cellular respiration. Concerning down-regulated processes, instead, we find down-regulated uptake of insulin-like growth factor (IGF) by IGF binding protein, together with multiple terms related to cell motility and cilium assembly, the organization of the extracellular matrix and oxidations;

- In fibroblasts, we find some similarities with cancer cells: they also show upregulation of terms related to cell growth and division and the TNFa/NF-kB pathway, together with downregulation of processes related to cell motility, the extracellular matrix and the transport of ions. Differently from tumors cells, they also show higher activity in many pathways related to cellular respiration, RNA translation and amino acid metabolism and regulation of apoptosis, together with increased activity of the KRAS signaling pathway, an important growth signaling pathway. Finally, they show upregulation of the stress response to unfolded proteins, rejection of allografts and expression of SLITs and ROBOs;

Considering instead the subclusters of T/NK cells, we find that:

- NK cells show upregulation of production of components of the complement system. On the other hand, we find downregulation of genes involved in the innate immune response, metabolism of amino acids and translation, and the cell cycle, which are possible signs of an impaired innate immune response;

- In cytotoxic T cells, we find higher activity of genes involved in RNA translation, cellular respiration, protein secretion together with production of components of the complement system, and cell cycle and growth. Considering less expressed genes, we find groups involved in Notch4 signaling, together with increased degradation of DVL and GLI1, which is indicative of downregulation of the Wnt and Hedgehog signaling pathways;
- Both helper and regulatory T cells how upregulation of pathways involved in RNA translation and SLIT/ROBO signaling. Considering helper lymphocytes, they also show increased signaling by TNFα, together with downregulation of multiple pathways involved in aerobic metabolism;
- Lastly, regulatory T cells show increased activity of pathways related to cell growth, mitosis and cell respiration. On the other hand, they show decreased IL2/STAT5 and IL6/STAT3 signaling, two pathways that play a central role in modulating CD4 T cell activity, together with decreased activities in multiple pathways related to interactions with the extracellular matrix.

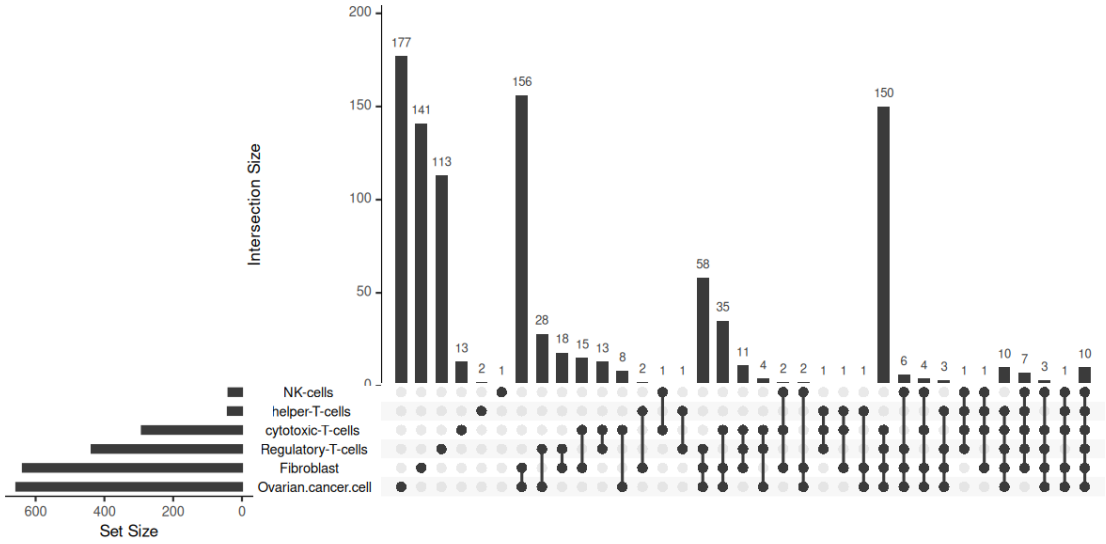


Figure 5.3: Upset plot of enriched terms in primary samples. As we can see, the groups of cells that share the most enriched terms are fibroblasts and tumor cells, and all cells but helper T and NK lymphocytes.

Among these results, some gene sets need to be discussed in more detail:

- First of all, we find repeated evidence of increased pro-inflammatory signaling, as evidenced by the upregulation of TNF-α signaling, together with further activation of the inflammation response in multiple cell types. Higher levels of TNF-α and other pro-inflammatory cytokines has been associated with poor prognosis in ovarian cancer [28];

- The increased level of interferon signaling we identified is also associated with a pro-inflammatory environment. However, their impact on tumor cells in not clearly defined [29];

- The upregulation in cancer cells of multiple pathways related to cell growth and mitosis, provides further evidence of the increased proliferation of tumor cells we identified in section 4.2. They also show reduced immune interactions with lymphoid cells, which might suggest a reduced capacity for attack by T/NK cells;

- Fibroblasts show higher metabolism, which is one of the hallmarks of cancer-associated fibroblasts [30], together with higher expression of genes related to the epithelial-mesenchymal transition, which drives tumor progression and metastasis and has been found to be driven in great part by cancer-associated fibroblasts in ovarian cancer [31];

- Multiple cell types show upregulation in production of elements of the complement system, which plays an important, although ambiguous, role in the ovarian cancer TME [32];

- We found repeated evidence of increased expression of Slit glycoproteins , with increased signaling by Roundabout (ROBO) receptors in all T cells subsets (but decreased in NK cells). Slit/ROBOs are guidance molecules involved in cell migration which have also been implicated in angiogenesis, cell proliferation and migration in multiple tumor types [33];

- All cell types show an increased stress response to amino acid starvation, together with widely altered activity of translation pathways and increased stress response to unfolded proteins in tumor cells, fibroblasts and regulatory lymphocytes, showing widely altered amino acid and protein metabolism

Comparing these results with those obtained from metastasis samples, we again find an increased level of interferon signaling and inflammation throughout the TME. However, results differ greatly, with an average Jaccard index between enriched terms in the same cell type between primary and metastasis samples of only 0.26.

- Tumor cells again show higher activity in pathways related to cell growth, mitosis and cellular respiration. We find many downregulated gene sets, involving interactions and modifications of the extracellular matrix (especially collagen formation), the RAS signaling pathway, ephrin receptors ,transforming growth factor beta (TGF-β) signaling, cell-cell adhesion and the epithelial-mesenchymal transition, hypoxia and the estrogen response;

- Similarly, we find again in fibroblasts upregulated pathways related to the cell cycle, mitosis and aerobiic metabolism, together with downregulated cell cell adhesion, interactions with the extracellular matrix, TNF-α and Notch4 signaling, hypoxia response, epithelial-mesenchymal transition, estrogen response and amino acid transport;
- NK cells show increased signaling by Notch receptors, production of complement proteins and reactions involved in mitosis, together with reduced response to amino acid starvation, immune interactions with non-lymphoid cells;
- Cytotoxic T cells show upregulation of processes involved in cell growth, division and respiration, but diminished activity of the Wnt, Ras and TNF-α signaling pathways, ion transport and the hypoxic response;
- In helper T cells, we find activated pathways related to leukocyte-mediated immunity and proliferation, together with increased response to starvation, signaling by Robo receptors and TNF-α and apoptosis;
- By contrast, regulatory T cells show greater activity of pathways related to the cell cycle and division, together with less activity of the TNF-α and Notch signaling pathways and genes involved in T cell differentiation and immune response

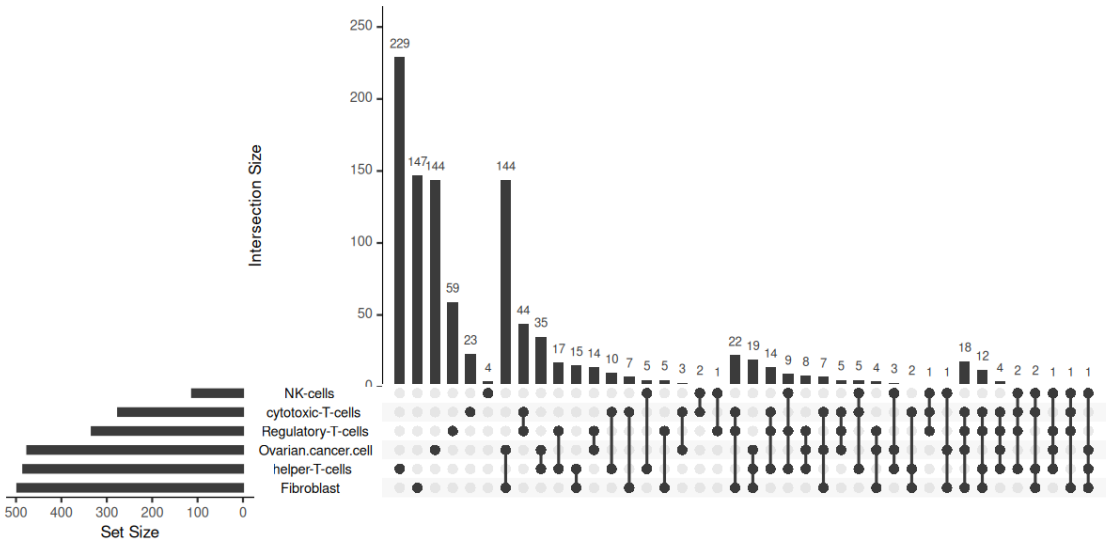


Figure 5.4: Upset plot of enriched terms in metastasis samples. While fibroblasts and tumor cells again share many terms, there are in general less shared enrichments across cell types.

While we again find increased inflammation and interferon signaling, we find widely downregulated signaling by TNF-α, which is the opposite of what we found in primary samples. We

also find repeated evidence of downregulated response to estrogen and hypoxia (which is coherent with the increased activity of aerobic metabolism in multiple cell types). Tumor cells and fibroblasts also both show decreased cell-cell adhesion and expression of gene programs related to the epithelial-mesenchymal transition.

5.2.2 Drews signatures

As we said, only tumor cells show a significant amount of differentially expressed genes when comparing the D1 and D2 clusters of patients (for plots of the top enriched/depleted gene sets, see figure A.7).

In primary samples, tumor cells from patients belonging to group D2 show increased antigen presentation and activation of the immune response (importantly, suppression of antigen presentation is one of the main mechanisms of immune escape by tumors [34]), together with increased interferon signaling, inflammation, response to estrogen and activity of the IL6/STAT3 and TNFa/NF-Kb signaling pathways and activity of the processes involved in programmed cell death. Furthermore, they show downregulation of the smoothened signaling pathway and processes involved in cell growth and division. All of this is coherent with our results from the previous chapter, which showed a lower abundance of tumor cells in D2 samples.

In metastasis samples, we again find downregulated processes associated with mitosis. Furthermore, we find upregulation in pathways related to amino acid metabolism, response to estrogen and androgen hormones, cell respiration and SLIT/ROBO signaling.

Although these results allow us to identify some broad effects of copy number instability on the tumor microenvironment, further investigation is needed to understand how CN mutational patterns induce these changes in the TME. Lastly, the changes we identified in metastases are much more heterogeneous across cell types and unclear, perhaps due to the diversity of metastasis sites.

Among the results in this chapter, we found multiple instances of altered signaling across the TME. We further investigated this aspect, which is the subject of the next chapter.

6

Mutational signatures and cell-cell communication

Cells constantly communicate with each other and with the environment around them, in three broad modalities:

- One cell may secrete a protein (ligand), which binds a transmembrane protein on the surface of another cell (receptor), triggering a downstream signaling cascade leading to changes in the target cell;
- Similarly, cells may communicate via contact, with both ligands and receptors being transmembrane proteins;
- Finally, adhesion molecules such as integrins may bind ECM proteins, informing the cell about the environment surrounding it.

In cancer, cell-cell communication is significantly altered, as cancer cells reshape it to favor disease progression by reprogramming the surrounding stroma to support them and creating a hostile environment for immune cells.

We assessed specific up and down-regulated interactions between our groups of patients using the Python package *LIANA* [35].

In section 6.1, we will briefly explain the two ways in which we identified disregulated interactions, and how we extracted biological insights from these. We will then go over our results

in section 6.2. Note that we only performed this analysis comparing U and HU patients, as this is the only comparison that showed significant changes in gene expression in multiple cell types.

6.1 Methods for comparative analysis of communication events

Cell-cell communication events are inferred by combining a resource of ligand-receptor pairs with a tool for scoring the co-expression of these pairs within a single or multiple datasets.

LIANA allows for a consistent identification of communication events by using a consensus resource and scoring method. Given a scRNA-seq dataset and cell type annotations, it scores interactions for each possible signal sender/receiver cell type pair using multiple methods and aggregates the results (see table 6.1). In particular, it aggregates the scores from CellChat, CellPhoneDB, Connectome, NATMI and SingleCellSignalR.

sample	source	target	ligand_complex	receptor_complex	scaled_weight	spec_weight	lrscore	magnitude_rank
SPECTRUM-OV-002_RIGHT_OVARY	Regulatory T cells	NK cells	B2M	KLRD1	1.495504	0.1480345	0.9546232	2.263237e-09
SPECTRUM-OV-002_RIGHT_OVARY	CD4 T cells	NK cells	B2M	KLRD1	1.388675	0.1382659	0.9531214	9.052851e-09
SPECTRUM-OV-002_RIGHT_OVARY	CD8 T cells	NK cells	B2M	KLRD1	1.380445	0.1375133	0.9529994	2.036870e-08
SPECTRUM-OV-002_RIGHT_OVARY	NK cells	NK cells	B2M	KLRD1	1.365501	0.1361467	0.9527752	3.621064e-08
SPECTRUM-OV-002_RIGHT_OVARY	Fibroblast	CD8 T cells	COL1A1	CD44	1.513442	0.2458949	0.9500007	5.657853e-08
SPECTRUM-OV-002_RIGHT_OVARY	Fibroblast	CD4 T cells	COL1A1	CD44	1.438992	0.2266677	0.9480312	1.108898e-07

Table 6.1: Example table of liana output. For each sample and sender/receiver cell type pair, LIANA aggregates scores for ligand-receptor pairs from different methods into a consensus rank.

As in the DGE analysis, we focused on interactions involving cancer cells, fibroblasts and T/NK lymphocytes.

We approached this analysis in two ways:

- To obtain a broad overview of the communication events present in our dataset and their changes, we first used *LIANA* to score interactions in each sample separately starting from the normalized matrix of gene expression and the cell type annotation for each sample, obtaining tables such as that in table 6.1, then used *Tensor-cell2cell* [36] to build a tensor with the results and decompose it into factors using Tensor Component Analysis, to find up and down regulated modules of cell-cell communication between U and HU patients. We will explain this in more detail in section 6.1.1;

- To analyze in more detail the changes in communication events involving tumor cells, we integrated the differential gene expression results for each cell type we obtained in chapter 5 with the database of ligand-receptor pairs provided by *LIANA* to identify disregulated interactions between cancer cells and other cell types. See section 6.1.2 for more details.

6.1.1 Unsupervised decomposition of cell-cell communication using Tensor-cell2cell

Tensor-cell2cell [36] is a Python package for the decomposition of cell-cell communication events across contexts, such as tissue, disease state and life stage.

After scoring communication events for each sample separately using *LIANA*, we built a tensor of the shared interactions using *Tensor-cell2cell*. The tensor is of order 4, with dimensions corresponding to samples, ligand-receptor pair, sender cell type and receiver cell type, so that a_{ijkl} is the score for interaction j, with ligand expressed by cell type k and receptor by cell type l, in sample i.

Using Tensor Component Analysis, an analogue of Principal Component Analysis for tensors, this tensor is then decomposed as a sum of components, with each component the external product of vectors of loadings:

$$A \sim \sum_{r=1}^{R} c_i^r \otimes p_j^r \otimes s_k^r \otimes t_l^r \tag{6.1}$$

The loadings are constrained to be non-negative to allow for biological interpretation. The best decomposition is chosen by minimization of the Frobenius distance between the communication tensor and its reconstruction, with the optimal number of factors chosen by elbow analysis.

We built and decomposed the tensors separately for primary and metastasis samples, and filtered out samples lacking 2 or more cell types of interest and interactions not appearing across all samples in order to reduce tensor sparsity.

We then used the decomposition to identify changes in cell-cell communication by comparing the loadings for each factor between U and HU samples, using a Wilcoxon test for primary samples and a linear model with anatomical site as an additional covariate for metastases.

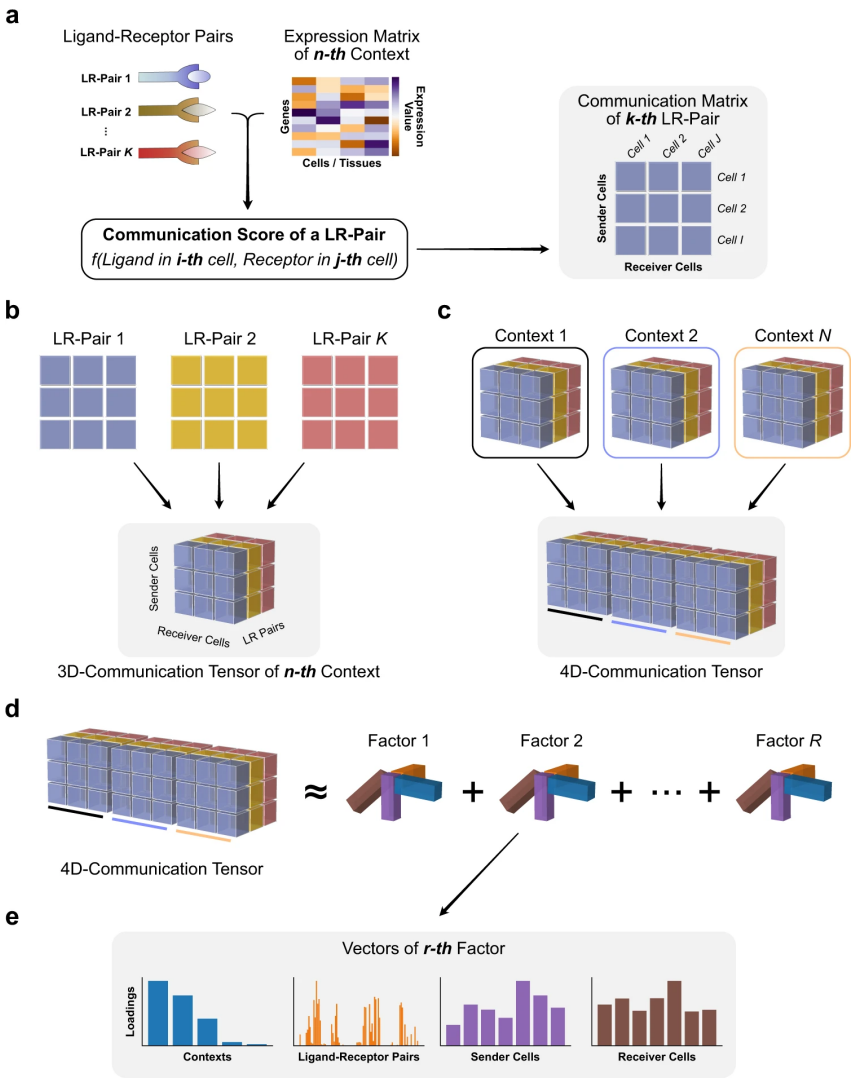


Figure 6.1: Decomposition of communication tensor using Tensor-cell2cell. **a**: for a given sample and ligand-receptor pair, a matrix is built with the score of that interaction for all possible sender-receiver cell types pairs **b**: The matrices for each interaction are stacked together to build a 3-dimensional tensor of the interactions in a sample. **c**: The 3-dimensional tensors are stacked to build the 4-dimensional communication tensor. **d**: Tensor factor decomposition into sum of external products of vectors of loadings. **e**: The non-negative loadings represent consistent interactions between cell type pairs across samples, allowing for biological interpretation. Figure from [36].

6.1.2 Assessing cell-cell communication changes from differential gene expression results

For a more detailed analysis, we used *LIANA* to integrate the results from our differential expression analysis (chapter 5) with its consensus resource. Given source cell type s, target cell type t and ligand-receptor pair lr, we score this interaction by taking the average of the log fold change of l in s and r in t.

For each comparison, we then have a dataset of differential expression of each ligand-receptor pair, for specific sender and receiver cell types, where for each pair expressed in at least 10% of the cells in our scRNA-seq data we take the average of the *LIMMA* test statistics of the differential expression test for the ligand and receptor (see table 6.2).

ligand	receptor	source	ligand_logFC	target	receptor_logFC	interaction_logFC	interaction
FN1	ITGA9	helper-T-cells	2.1042889	Fibroblast	2.4775618	2.290925365	FN1^ITGA9
CALM1	FAS	helper-T-cells	-0.9821922	Fibroblast	-0.8475826	-0.914887425	CALM1^FAS
GNAS	PTGDR	Fibroblast	-0.8079515	helper-T-cells	1.1703396	0.181194028	GNAS^PTGDR
HSPG2	DAG1	Fibroblast	1.0972821	Fibroblast	0.5388009	0.818041535	HSPG2^DAG1
LGALS1	CD69	helper-T-cells	1.2109824	helper-T-cells	-1.1966475	0.007167485	LGALS1^CD69
YBX1	NOTCH1	Fibroblast	-1.1407022	Regulatory-T-cells	0.6005354	-0.270083435	YBX1^NOTCH1

Table 6.2: Example table of altered cell-cell communication events inferred from DGE analysis. As we can see, for each interaction we have the source and target cell types, plus the average log fold change.

We used this dataset to further investigate signals sent from and received by tumor cells in the following way: after choosing whether to investigate incoming or outgoing signals from cancer cells and the other cell type involved, we subset this dataset to those interactions satisfying these criteria. We then rank the interactions by the average log fold change of the ligand-receptor pair and use this ranked list to perform GSEA in the way described in the following section.

Repeating this procedure for both incoming and outgoing signals, and for all cell types of interest, we can capture enriched or depleted events in cell-cell communication between U and HU patients involving tumor cells.

6.1.3 Enrichment analysis of cell-cell communication

We performed enrichment analysis on our results, similarly to what we did for the DGE analysis in chapter 5.

We built interaction sets by integrating the *LIMMA* consensus resource of ligand-receptor pairs with external gene sets databases (specifically, Gene Ontology, Reactome, KEGG and PROGENY).

For the *Tensor-cell2cell* decomposition, for each factor of interest, we computed the z-scores of the ligand-receptor pairs and selected as significant those pairs with z-score > 2. We used these interactions to perform over-representation analysis (ORA), using the interaction sets we defined above.

Over-representation analysis, given a list of significant elements (usually genes), a gene set (usually corresponding to the genes involved in a specific process or pathway) and a background list termed "universe", determines if the overlap between significant genes and the gene set is significant by modeling the number of elements in the intersection using a hypergeometric distribution. We adapted this analysis to ligand-receptor pairs.

Considering how in the previous chapter only the comparison between U and HU patients revealed a significant amount of differentially expressed genes in multiple cell types, we investigated changes in cell cell communication only in this context.

6.2 Altered communication events between U and HU patients

In primary samples, the inferred tensor spans 2136 interactions across 35 samples. The elbow analysis identifies at 5 the optimal number of components (see figure A.8). As we can see in figure 6.2, the decomposition clearly identifies factors corresponding to communication between

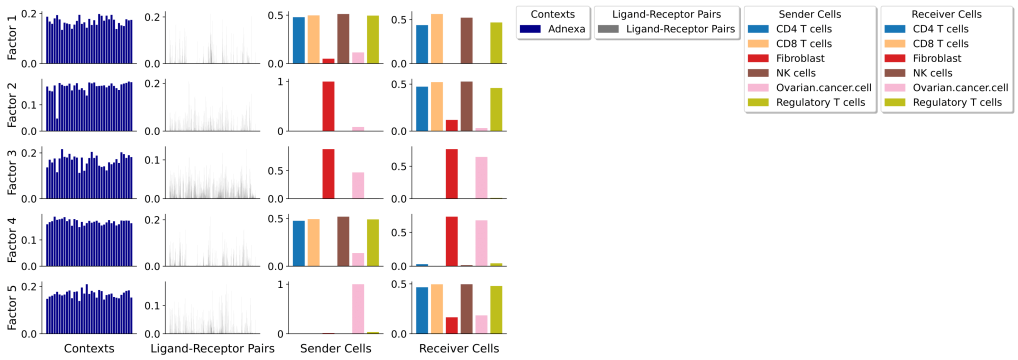


Figure 6.2: Decomposition of communication tensor in primary samples using *Tensor-cell2cell*. Each row describes a different factor, while the columns show the loadings correspinding to samples, ligand-receptor pairs, sender and receiver cell typer respectively.

immune cells, from fibroblasts to immune cells, reciprocal interactions between fibroblasts and cancer cells, from immune cells to fibroblasts and tumor cells and from tumor cells to immune cells respectively.

Characteristic	HU N = 30[1]	U N = 4[1]	p-value[2]	q-value[3]
Factor 1	0.166 (0.155, 0.177)	0.193 (0.165, 0.196)	0.13	0.2
Factor 2	0.173 (0.163, 0.181)	0.164 (0.156, 0.172)	0.2	0.3
Factor 3	0.178 (0.158, 0.184)	0.141 (0.126, 0.157)	0.026	0.066
Factor 4	0.172 (0.165, 0.175)	0.154 (0.151, 0.160)	0.003	**0.015**
Factor 5	0.168 (0.161, 0.178)	0.165 (0.145, 0.183)	0.8	0.8

[1] Median (Q1, Q3)

[2] Wilcoxon rank sum exact test

[3] False discovery rate correction for multiple testing

Table 6.3: Comparison of loadings for each factor between samples from HU and U patients. We used a Wilcoxon ranked sum test, and the Benjamini-Hochberg FDR adjustment, setting the threshold for adjusted p values at 0.05

We identify factor 4 as significantly more active in HU patients (see table 6.3), which we investigated using ORA (see figure 6.3).

We find enrichment in the TGF-β signaling pathway, together with the Notch and Calcium pathways, the epithelial mesenchymal transition and apoptotic signaling. TGF-β plays a dual role as both a tumor suppressor gene in early development, and as a promoter of a more aggressive phenotype in late-stage cancers. In ovarian cancer, it supports metastasis, angiogenesis and immune evasion by recruiting regulatory T cells [37]. Furthermore, Notch signaling has been implicated in tumor progression, metastasis and angiogenesis [38].

We further investigated signaling events involving cancer cells using the method we described in the previous section. A plot of the enriched terms identified by GSEA can be found at figure 6.4.

Some of these terms are worth discussing further:

- There is an enrichment in signals from and to fibroblasts involved in suppressing apoptosis;
- We find upreguated interactions involving the remodeling of the extracellular matrix and adhesion between cancer cells and fibroblasts. Looking at the interactions in these sets they are comprised of interactions between ECM proteins (especially collagen and laminins) and integrins. We can then conclude that in HU patients, there is a higher

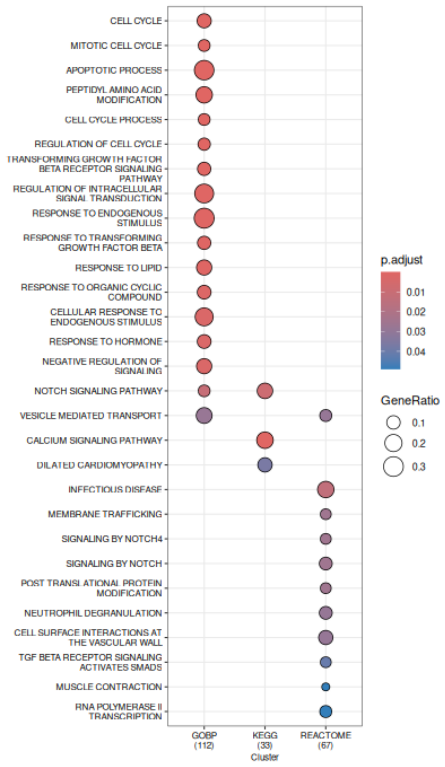


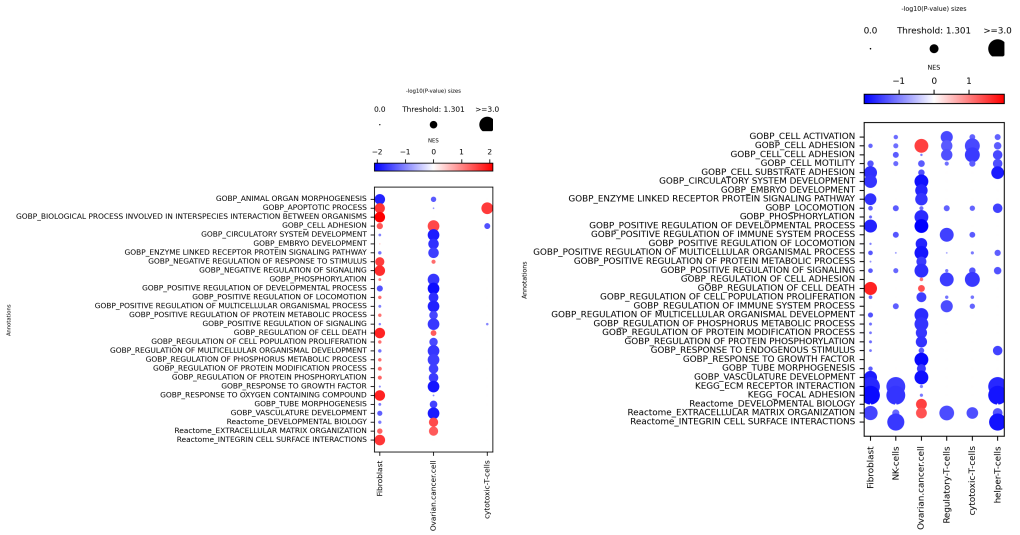
Figure 6.3: Top terms in overrepresentation analysis of the main interactions involved in factor 4, using the interaction sets derived from GOBP, KEGG and Reactome. The color of the dots represents the adjusted p value, while the size is the ratio between the number of interactions in our list of significant ones and all the interactions in the "universe", which we defined as all ligand-receptor pairs present in the tensor.

production of these ECM components by tumor cells and fibroblasts, and increased adhesion of these cells to the ECM via integrins, which regulate the focal adhesion kinase signaling pathway (promoting growth and proliferation), and facilitate migration [39];

- On the contrary, the adhesion of T and NK cells to the ECM and tumor cells is widely diminished. As adhesion is important for many aspects of T cell function, among which activation, proliferation and effector activities [40], this could imply further suppression of immune cell function in HU patients.

In summary, in the TME of primary tumors, HU patients show higher level of TGF-β , Notch and calcium signaling, together with improved adhesion of tumor cells and fibroblasts to each other and the ECM, which is instead downregulated in T and NK cells.

Moving on to metastasis samples, we find again 5 components, which again clearly define communication events between different portions of the TME: signaling between immune



(a) Incoming signals **(b)** Outgoing signals

Figure 6.4: Enriched terms found using GSEA on the ranked list of communication events involving tumor cells in primary lesions. The color of the dots represents the enrichment score, while the size is the adjusted p value (threshold 0.05)

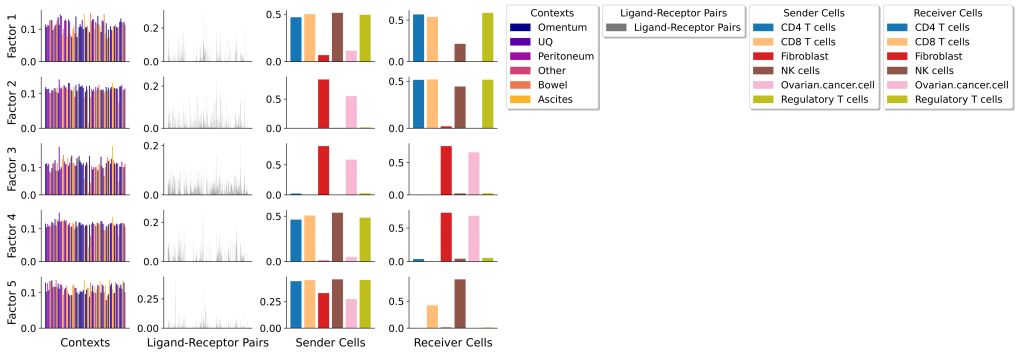


Figure 6.5: Decomposition of communication tesnsor in metastasis samples using *Tensor-cell2cell*. The figure can be interpreted the same way as figure 6.2.

cells, from fibroblasts and tumor cells to immune cells, between cancer cells and fibroblasts, from immune cells to fibroblasts and cancer cells, and from all cell types to cytotoxic immune cells respectively (see figure 6.5).

We tested for up/downregulated modules of cell-cell communication between U and HU patients, and identified factor 1 as more active in HU patients (see table 6.4).

ORA reveals that this factor represents communication between immune cells involved in the activation of the immune response and antigen presentation (see figure 6.6).

As we did for the primary samples, we further investigated communication events involving

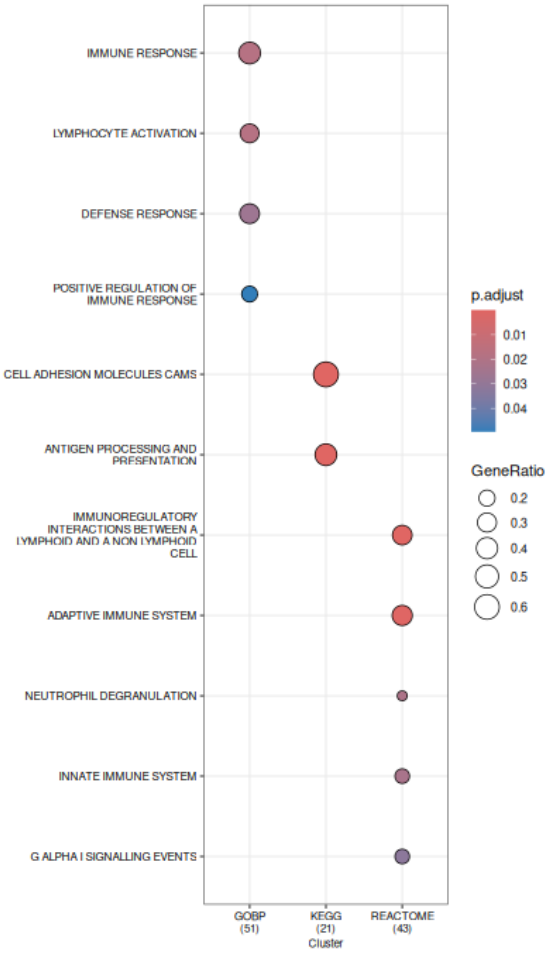
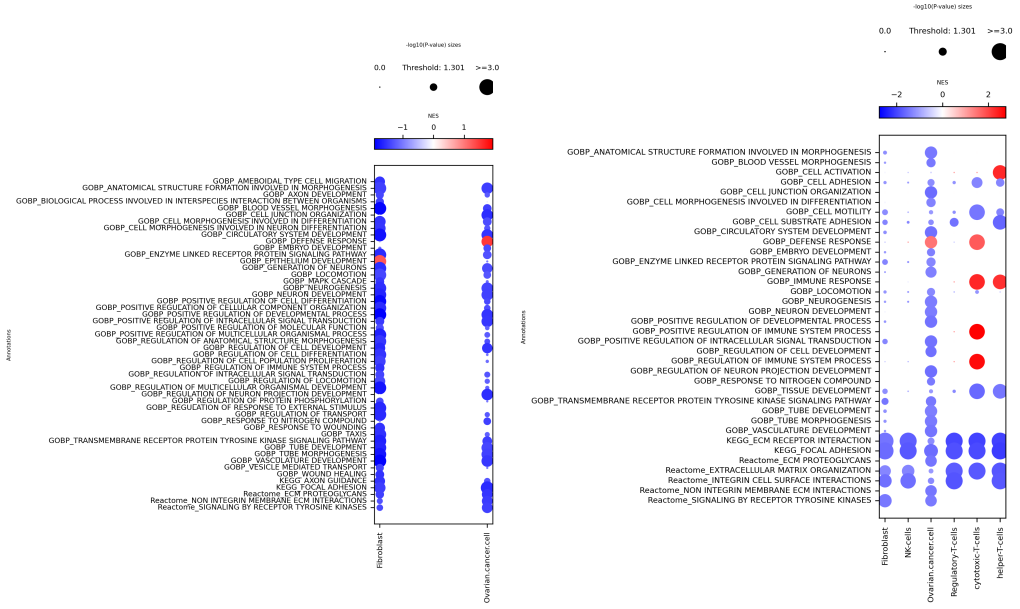


Figure 6.6: Top terms in overrepresentation analysis of the main interactions involved in factor 1, using the interaction sets derived from GOBP, KEGG and Reactome. The figure can be interpreted the same way as figure 6.3

response	Estimate	Std. Error	t value	Pr(>|t|)	q
Response Factor 1	-0.01607112	0.005751032	-2.794476	0.007142311	**0.03571155**
Response Factor 2	-0.002130661	0.003279967	-0.6495982	0.5186557	0.51865568
Response Factor 3	0.01072838	0.007767622	1.381166	0.1728139	0.29274291
Response Factor 4	-0.00369215	0.003449893	-1.070222	0.2891939	0.36149234
Response Factor 5	-0.006162991	0.004492101	-1.371962	0.1756457	0.29274291

Table 6.4: Comparison of loadings for each factor between samples from HU and U patients. For each factor, we used a linear model with covariates the stratification of the patient and the anatomical source of the sample. We then tabled together the coefficients for the mutational covariate from each model and performed FDR adjustment using the Benjamini-Hochberg procedure (threshold at 0.05)

tumor cells using GSEA (see figure 6.7).



(a) Incoming signals **(b)** Outgoing signals

Figure 6.7: Enriched terms found using GSEA on the ranked list of communication events involving tumor cells in metastases. The color of the dots represents the enrichment score, while the size is the adjusted p value (threshold 0.05)

The landscape of altered communication events is widely different from what we found in primary samples:

- Cell adhesion is widely downregulated, even in fibroblasts and cancer cells, in contrast with what we found before;

- We find further confirmation of the ORA results, with increased immune activation signaling from tumor cells to T cells;
- Finally, we find downregulated angiogenetic signaling between tumor cells and fibroblasts.

With this analysis of communication events between cells, we have completed our characterization of the impact of mutational signatures on the tumor microenvironment.

We will conclude this thesis in the next chapter, where we will briefly summarize the results we obtained, highlight some criticalities of this work and propose possible further directions and developments.

7 Conclusion

In this thesis, we carried out an exploration of the effects of copy number variation of the tumor microenvironment in ovarian cancer.

Using the mutational signatures from [16], we classified patients based on the degree of instability of the genome, distinguishing between Unstable and Highly Unstable patients.

Furthermore, the signatures defined in [6] allowed us to stratify the patients based on the degree of homologous recombination impairment, defining a group characterized by its depletion (or *BRCA*-ness), which we termed D2.

We then compared these groups in terms of cell type composition, differential gene expression and cell-cell communication.

Comparing U and HU patients, we characterized the tumor microenvironment in primary lesions of HU patients as involving higher inflammation, evidenced by increased signaling by TNF-α and interferons, together with higher production of complement proteins and widely altered amino acid metabolism. Furthermore, cancer cells and fibroblasts show increased adhesion to other cells and the ECM, which is instead impaired in lymphoid cells. Fibroblasts also show higher metabolism (which is one of the markers of cancer-associated fibroblasts) and expression of gene programs related to the epithelial-mesenchymal transition. Finally, we also find increased activity in multiple signaling pathways, specifically the Notch, TGF-β, SLIT/ROBO and calcium pathways.

All these characteristics converge on an impaired immune response and more aggressive tu-

mor growth.

In metastasis samples instead, while we again find increased inflammation and interferon signaling, we find results that conflict with the previous ones: in particulat, TNF-α signaling is widely downregulated, together with epithelial-mesenchymal transition programs and cell adhesion in tumor cells and fibroblasts. Finally, we also find greater activation across the immune component of the tumor microenvironment.

Concerning instead the comparison between D1 and D2 patients, in primary samples we find increased antigen presentation by tumor cells, inflammation and interferon signaling, with a corresponding decrease in the proliferation of tumor cells which we may impute to a more successful immune response.

We would like to close this work by highlighting some criticalities and possible future developments:

- First, while we have linked CN mutational patterns (specifically, the degree of instability and homologous recombination impairment) with changes in the tumor microenvironment, more research is needed in order to understand the actual causal relationship between these;
- We limited ourserlves to analyzing tumor cells, fibroblasts and T/NK cells: for a more complete characterization of the TME, we could expand this analysis to other cell types. For example, macrophages play an important and dual role in tumors, being able to both promote its progression and contrast it;
- To avoid issues of collinearity and simplify our analysis, we grouped patients based on the activities of mutational signatures: in the future, we could try working with these activities as continuous covariates to establish a more direct link with mutational processes;
- To improve our analysis of metastatic samples, we could separate them based on anatomical source and conduct our analysis separately for each one: however, we could run into issues of sample size

References

[1] M.-A. Lisio, L. Fu, A. Goyeneche, Z. hua Gao, and C. Telleria, "High-grade serous ovarian cancer: Basic sciences, clinical and therapeutic standpoints," *International Journal of Molecular Sciences*, vol. 20, no. 4, 2019.

[2] C. Stewart, C. Ralyea, and S. Lockwood, "Ovarian cancer: An integrated review," *Seminars in Oncology Nursing*, vol. 35, no. 2, 2019.

[3] Y. Hatano, K. Hatano, M. Tamada *et al.*, "A comprehensive review of ovarian serous carcinoma," *Adv Anat Pathol*, vol. 26, no. 5, 219.

[4] E. Schoutrop, L. Moyano-Galceran, S. Lheureux *et al.*, "Molecular, cellular and systemic aspects of epithelial ovarian cancer and its tumor microenvironment," *Seminars in Cancer Biology*, vol. 86, no. 3, pp. 207–223, 2022.

[5] C. D. Steele, N. Pillay, and L. B. Alexandrov, "An overview of mutational and copy number signatures in human cancer," *The Journal of Pathology*, vol. 257, no. 4, 2022.

[6] R. M Drews *et al.*, "A pan-cancer compendium of chromosomal instability," *Nature*, vol. 606, pp. 976–983, 2022.

[7] C. D. Steele, A. Abbasi, S. M. A. Islam *et al.*, "Signatures of copy number alterations in human cancer," *Nature*, vol. 606, pp. 984–991, 2022.

[8] Z. Tao, S. Wang, C. Wu *et al.*, "The repertoire of copy number alteration signatures in human cancer," *Briefings in Bioinformatics*, vol. 24, no. 2, 2023.

[9] B. Arneth, "Tumor microenvironment," *Medicina(Kaunas)*, vol. 56, no. 1, 2020.

[10] P.-H. Li, X.-Y. Kong, Y.-Z. He *et al.*, "Recent developments in application of single-cell rna sequencing in the tumour immune microenvironment and cancer therapy," *Military Medical Research*, vol. 9, no. 52, 2022.

[11] D. C. Hinshaw1 and L. A. Shevde, "The tumor microenvironment innately modulates cancer progression," *Cancer Res*, vol. 79, no. 18, pp. 4557–4566, 2019.

[12] Y. Yang, Y. Yang, J. Yang *et al.*, "Tumor microenvironment in ovarian cancer: Function and therapeutic strategy," *Frontiers in Cell Developmental Biology*, vol. 8, 2020.

[13] I. Vázquez-García, F. Uhlitz, N. Ceglia *et al.*, "Ovarian cancer mutational processes drive site-specific immune evasion," *Nature*, no. 612, p. 778–786, 2022.

[14] H. Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. [Online]. Available: https://arxiv.org/abs/1303.3997

[15] P. Van Loo, S. H. Nordgard, O. C. Lingjærde *et al.*, "Allele-specific copy number analysis of tumors," *PNAS*, vol. 107, no. 39, pp. 16 910–16 915, 2010.

[16] C. Pesenti, L. Beltrame, A. Velle *et al.*, "Copy number alterations in stage i epithelial ovarian cancer highlight three genomic patterns associated with prognosis," *European Journal of Cancer*, vol. 171, pp. 85–95, 2022.

[17] Cristopher D. Steele. Cosmic copy number signatures. [Online]. Available: https://cancer.sanger.ac.uk/signatures/cn/

[18] A. Khandekar, R. Vangara, M. Barnes *et al.*, "Visualizing and exploring patterns of large mutational events with sigprofilermatrixgenerator," *BMC Genomics*, vol. 24, no. 469, 2023.

[19] M. Sìaz-Gay, R. Vangara, M. Barnes *et al.*, "Assigning mutational signatures to individual samples and individual somatic mutations with sigprofilerassignment," *Bioinformatics*, vol. 39, no. 12, 2023.

[20] L. Heumos, A. C. Schaar, C. Lance *et al.*, "Best practices for single-cell analysis across modalities," *Nature Reviews Genetics*, vol. 24, pp. 550–572, 2023.

[21] M. D. Young and S. Behjati, "Soupx removes ambient rna contamination from droplet-based single-cell rna sequencing data," *GigaScience*, vol. 29, no. 12, 2020.

[22] P.-L. Germain, A. Lun, C. Garcia Meixide *et al.*, "Doublet identification in single-cell sequencing data using scdblfinder," *F1000Research*, vol. 10, no. 979, 2021.

[23] A. W. Zhang, C. O'Flanagan, E. A. Chavez *et al.*, "Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling," *Nature methods*, vol. 16, pp. 1007–1015, 2019.

[24] A. Ianevski, A. K. Giri, and T. Aittokallio, "Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data," *Nature communications*, vol. 13, no. 1246, 2022.

[25] C. D. Conde, C. Xu, L. B. Jarvis *et al.*, "Cross-tissue immune cell analysis reveals tissue-specific features in humans," *Science*, vol. 376, no. 6594, 2022.

[26] M. Buttner, J. Ostner, C. L. Muller *et al.*, "sccoda is a bayesian model for compositional single-cell data analysis," *Nature Communications*, vol. 12, no. 6876, 2021.

[27] M. H. Abdolvahab, B. Darvishi, M. Zarei *et al.*, "Interferons: role in cancer therapy," *Immunotherapy*, vol. 12, no. 11, 2020.

[28] A. Macciò, C. Madeddu *et al.*, "Inflammation and ovarian cancer," *Cytokines*, vol. 58, pp. 133–147, 2012.

[29] T. Liu, Y. Li, X. Wang *et al.*, "The role of interferons in ovarian cancer progression: Hinderer or promoter?" *Frontiers in Immunology*, vol. 13, 2022.

[30] M. Zhang, Z. Chen, Y. Wang *et al.*, "The role of cancer-associated fibroblasts in ovarian cancer," *Cancers (Basel)*, vol. 14, no. 11, 2022.

[31] P. M. Szabo, A. Vajdi, N. Kumar *et al.*, "Cancer-associated fibroblasts are the main contributors to epithelial-to-mesenchymal signatures in the tumor microenvironment," *Scientific Reports*, vol. 13, no. 3051, 2023.

[32] Y. Senent, D. Ajona, A. Gonzalez-Martin *et al.*, "The complement system in ovarian cancer: An underexplored old path," *Cancers (Basel)*, vol. 13, no. 15, 2021.

[33] R. K. Gara, S. Kumari, A. Ganju *et al.*, "Slit/robo pathway: a promising therapeutic target for cancer," *Drug Discovery Today*, vol. 20, no. 1, 2015.

[34] S. Jhunjhunwala, C. Hammer, and L. Delamarre, "Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion," *Nature Reviews Cancer*, vol. 21, pp. 298–312, 2021.

[35] D. Dimitrov, P. Sven Lars Schäfen, E. Farr *et al.*, "Liana+ provides an all-in-one framework for cell–cell communication inference," *Nature Cell Biology*, vol. 26, pp. 1613–1622, 2024.

[36] E. Armingol, H. M. Baghdassarian, C. Martino *et al.*, "Context-aware deconvolution of cell–cell communication with tensor-cell2cell," *Nature Communications*, vol. 13, no. 3665, 2022.

[37] B. M. Roane, R. C. Arend, and M. J. Birrer, "Review: Targeting the transforming growth factor-beta pathway in ovarian cancer," *Cancers(Basel)*, vol. 11, no. 5, 2019.

[38] X. Li, X. Yan, Y. Wang *et al.*, "The notch signaling pathway: a potential target for cancer immunotherapy," *Journal of Hematology and Oncology*, vol. 16, no. 1, 2023.

[39] M. Janiszewska, M. C. Primi, and T. Izard, "Cell adhesion in cancer: Beyond the migration of single cells," *Journal of Biological Chemistry*, vol. 295, no. 8, 2020.

[40] H. Harjunpää, M. L. Asens, C. Guenther *et al.*, "Cell adhesion molecules and their roles and regulation in the immune and tumor microenvironment," *Frontiers in Immunology*, vol. 10, 2019.

A

Extra tables and plots

Tao clusters signature enrichment				
	Tao cluster			
Signature	**T1**, N = 27[1]	**T2**, N = 10[1]	**p-value**[2]	**q-value**[3]
PCAWG_CNS1	0.07 (0.02, 0.19)	0.00 (0.00, 0.01)	0.004	**0.010**
PCAWG_CNS2	0.13 (0.08, 0.21)	0.07 (0.03, 0.15)	0.2	0.2
PCAWG_CNS3	0.02 (0.01, 0.04)	0.09 (0.07, 0.13)	<0.001	**<0.001**
PCAWG_CNS4	0.000 (0.000, 0.009)	0.000 (0.000, 0.000)	0.084	0.13
PCAWG_CNS5	0.19 (0.09, 0.25)	0.00 (0.00, 0.01)	<0.001	**<0.001**
PCAWG_CNS6	0.09 (0.03, 0.21)	0.05 (0.00, 0.12)	0.2	0.3
PCAWG_CNS7	0.03 (0.01, 0.06)	0.16 (0.14, 0.26)	<0.001	**<0.001**
PCAWG_CNS8	0.02 (0.00, 0.03)	0.04 (0.01, 0.10)	0.2	0.2
PCAWG_CNS9	0.14 (0.08, 0.19)	0.01 (0.00, 0.02)	<0.001	**<0.001**
PCAWG_CNS10	0.01 (0.00, 0.03)	0.08 (0.01, 0.13)	0.020	**0.040**
PCAWG_CNS11	0.02 (0.01, 0.04)	0.20 (0.12, 0.32)	<0.001	**0.002**
PCAWG_CNS12	0.07 (0.02, 0.11)	0.06 (0.03, 0.08)	0.7	0.7
PCAWG_CNS13	0.00 (0.00, 0.07)	0.00 (0.00, 0.00)	0.2	0.3
PCAWG_CNS14	0.000 (0.000, 0.012)	0.016 (0.000, 0.067)	0.074	0.13

[1] Median (IQR)

[2] Wilcoxon rank sum test

[3] False discovery rate correction for multiple testing

Table A.1: Enrichment in signature activities in Tao clusters

Drews clusters signature enrichment

Signature	Drews cluster: D1 N = 16 [1]	Drews cluster: D2 N = 20 [1]	p-value [2]	q-value [3]
CX1	0.03 (0.00, 0.10)	0.16 (0.09, 0.21)	0.002	**0.007**
CX2	0.17 (0.14, 0.23)	0.13 (0.07, 0.19)	0.034	0.073
CX3	0.14 (0.03, 0.24)	0.42 (0.36, 0.46)	<0.001	**<0.001**
CX4	0.039 (0.020, 0.058)	0.019 (0.000, 0.055)	0.2	0.3
CX5	0.28 (0.24, 0.35)	0.12 (0.08, 0.15)	<0.001	**<0.001**
CX6	0.0000 (0.0000, 0.0000)	0.0000 (0.0000, 0.0012)	0.12	0.2
CX7	0.06 (0.04, 0.11)	0.03 (0.01, 0.05)	0.002	**0.007**
CX8	0.0000 (0.0000, 0.0000)	0.0000 (0.0000, 0.0000)	0.3	0.4
CX9	0.06 (0.03, 0.08)	0.00 (0.00, 0.02)	<0.001	**<0.001**
CX10	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.4	0.5
CX11	0.06 (0.03, 0.07)	0.02 (0.00, 0.05)	0.004	**0.010**
CX12	0.02 (0.00, 0.11)	0.00 (0.00, 0.02)	0.055	0.10
CX13	0.0000 (0.0000, 0.0000)	0.0000 (0.0000, 0.0000)	0.5	0.5
CX14	0.000 (0.000, 0.003)	0.000 (0.000, 0.002)	>0.9	>0.9
CX15	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.12	0.2
CX16	0.002 (0.000, 0.019)	0.021 (0.000, 0.032)	0.2	0.2
CX17	0.00 (0.00, 0.00)	0.00 (0.00, 0.03)	0.005	**0.013**

[1] Median (Q1, Q3)

[2] Wilcoxon rank sum test; Wilcoxon rank sum exact test

[3] False discovery rate correction for multiple testing

(a) Drews clusters comparisons

HUS clusters vs drews signatures

Signature	HUS classification: HU N = 30 [1]	HUS classification: U N = 7 [1]	p-value [2]	q-value [3]
CX1	0.08 (0.02, 0.15)	0.27 (0.19, 0.36)	<0.001	**0.004**
CX2	0.16 (0.13, 0.21)	0.10 (0.05, 0.16)	0.050	0.2
CX3	0.27 (0.13, 0.43)	0.31 (0.21, 0.45)	0.7	0.8
CX4	0.028 (0.009, 0.066)	0.013 (0.011, 0.028)	0.4	0.7
CX5	0.20 (0.13, 0.29)	0.08 (0.06, 0.12)	0.001	**0.010**
CX6	0.0000 (0.0000, 0.0000)	0.0000 (0.0000, 0.0073)	0.044	0.2
CX7	0.05 (0.02, 0.07)	0.03 (0.02, 0.04)	0.2	0.5
CX8	0.0000 (0.0000, 0.0000)	0.0000 (0.0000, 0.0038)	0.2	0.5
CX9	0.02 (0.00, 0.06)	0.02 (0.02, 0.03)	0.9	>0.9
CX10	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.5	0.7
CX11	0.04 (0.01, 0.07)	0.03 (0.03, 0.09)	>0.9	>0.9
CX12	0.01 (0.00, 0.03)	0.00 (0.00, 0.03)	0.5	0.7
CX13	0.0000 (0.0000, 0.0000)	0.0000 (0.0000, 0.0000)	0.4	0.7
CX14	0.000 (0.000, 0.000)	0.000 (0.000, 0.008)	0.2	0.5
CX15	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.5	0.7
CX16	0.019 (0.000, 0.032)	0.000 (0.000, 0.014)	0.2	0.5
CX17	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.6	0.7

[1] Median (Q1, Q3)

[2] Wilcoxon rank sum test; Wilcoxon rank sum exact test

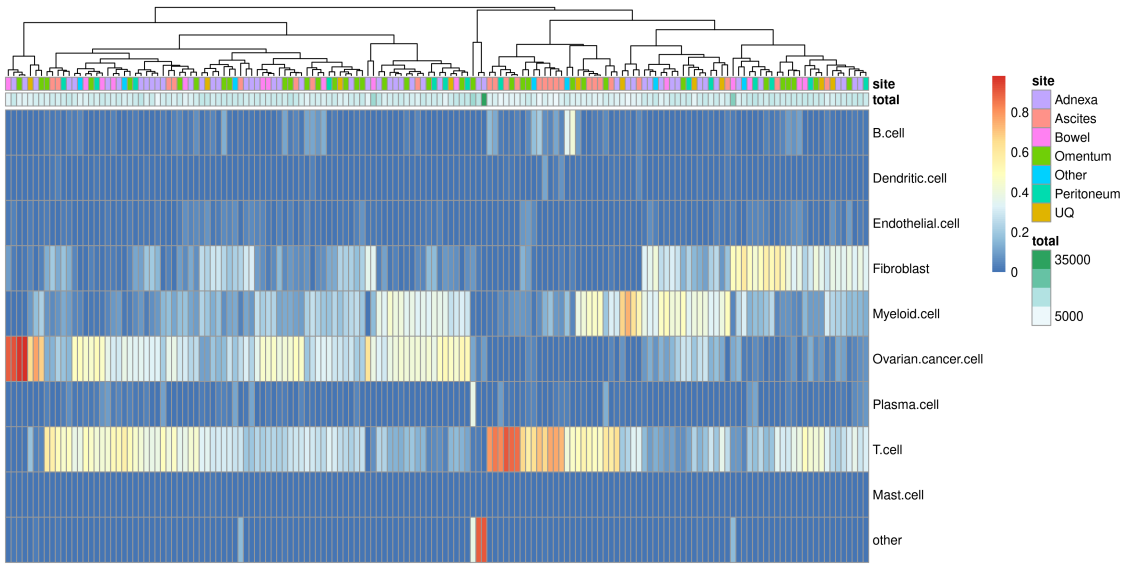
[3] False discovery rate correction for multiple testing

(b) HUS classification comparisons

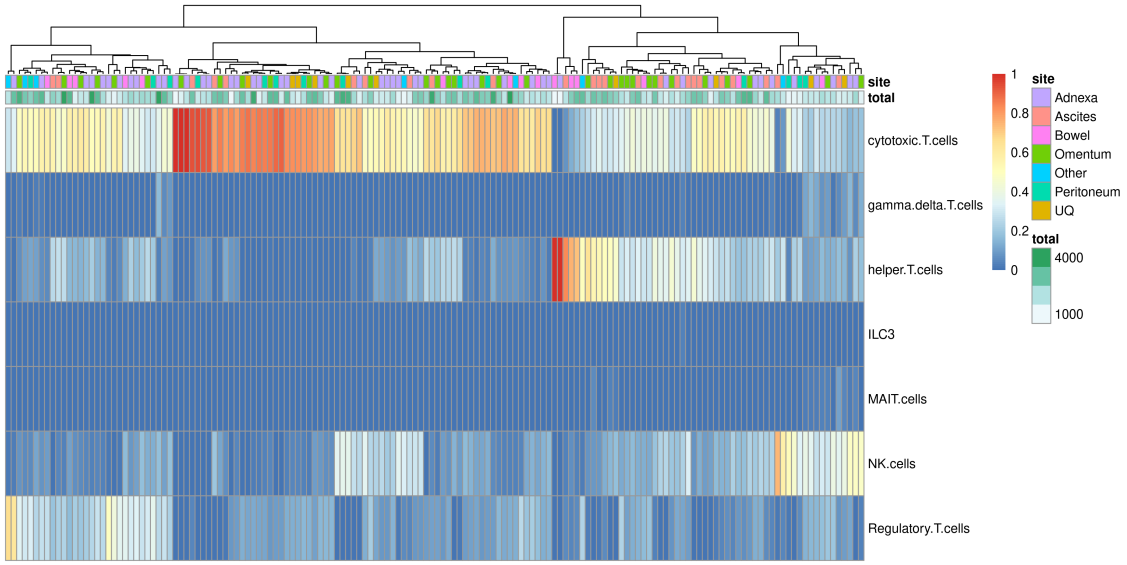
Table A.2: Enrichment in signature activities in Drews cluster

	Anatomical source							
patient	Adnexa	Ascites	Bowel	Omentum	Other	Peritoneum	UQ	total
SPECTRUM-OV-002	4569	0	0	2969	0	0	0	7538
SPECTRUM-OV-003	7903	0	0	4145	0	5444	6751	24243
SPECTRUM-OV-007	6829	2746	4411	6968	4920	5505	0	31379
SPECTRUM-OV-008	0	0	8143	5261	5733	0	0	19137
SPECTRUM-OV-009	14618	719	7325	9410	0	9689	11271	53032
SPECTRUM-OV-014	6843	4437	5590	0	0	7684	4024	28578
SPECTRUM-OV-022	16676	5042	7047	0	0	0	0	28765
SPECTRUM-OV-024	0	4985	0	5153	0	7093	4668	21899
SPECTRUM-OV-025	3223	0	5124	4722	0	0	0	13069
SPECTRUM-OV-026	15660	7692	9339	0	0	0	0	32691
SPECTRUM-OV-031	6111	0	0	0	3413	0	0	9524
SPECTRUM-OV-036	15021	0	0	5886	0	1374	0	22281
SPECTRUM-OV-037	8067	3797	0	7088	0	0	4711	23663
SPECTRUM-OV-041	9023	6345	0	7768	0	0	0	23136
SPECTRUM-OV-042	0	7137	0	8091	0	8807	8340	32375
SPECTRUM-OV-045	19432	0	0	6013	7060	0	0	32505
SPECTRUM-OV-049	7376	0	0	8664	0	0	0	16040
SPECTRUM-OV-050	14838	4896	0	16367	0	0	0	36101
SPECTRUM-OV-051	5897	3524	2662	3115	0	0	0	15198
SPECTRUM-OV-052	7725	0	0	7246	0	0	0	14971
SPECTRUM-OV-053	16203	0	0	8721	0	0	9322	34246
SPECTRUM-OV-054	1386	3266	0	7504	0	0	0	12156
SPECTRUM-OV-065	14434	8422	0	6327	0	0	0	29183
SPECTRUM-OV-067	3986	0	2958	3028	0	0	0	9972
SPECTRUM-OV-068	0	1924	20414	6241	0	0	0	28579
SPECTRUM-OV-070	0	3530	0	6607	8596	7103	0	25836
SPECTRUM-OV-071	19655	5760	0	0	0	6889	0	32304
SPECTRUM-OV-075	14244	4408	0	6709	0	0	0	25361
SPECTRUM-OV-077	7511	4932	9203	4066	0	0	0	25712
SPECTRUM-OV-080	862	8802	0	4085	0	9476	0	23225
SPECTRUM-OV-081	42945	1054	0	6789	0	0	0	50788
SPECTRUM-OV-082	16880	0	16327	5753	0	0	0	38960
SPECTRUM-OV-083	10143	7104	0	10628	0	0	0	27875
SPECTRUM-OV-090	0	0	8057	8746	0	0	0	16803
SPECTRUM-OV-105	0	4121	0	7789	6970	0	5723	24603
SPECTRUM-OV-107	5597	4636	5149	5712	0	0	0	21094
SPECTRUM-OV-110	0	0	9091	6932	0	10365	0	26388
SPECTRUM-OV-112	7901	3607	0	8981	0	2572	0	23061
SPECTRUM-OV-115	17203	0	0	8127	0	7570	0	32900
SPECTRUM-OV-116	6298	7044	0	7996	8134	0	0	29472
SPECTRUM-OV-118	17481	0	0	0	0	8541	0	26022
sum	372540	119930	120840	239607	44826	98112	54810	1050665

Table A.3: Table of number of cells by anatomical source and patient.

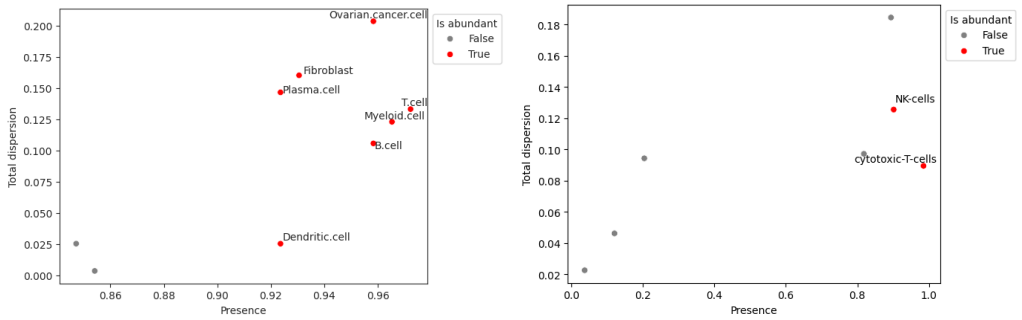


(a) CellAssign-identified clusters



(b) T cells subclusters identified by CellTypist

Figure A.1: Heatmaps of cell type abundances by sample



(a) CellAssign-identified clusters

(b) T cells subclusters identified by CellTypist

Figure A.2: Abundance vs Dispersion plots created usign scCODA. Sufficiently abundant celltypes are labeled and colored in red

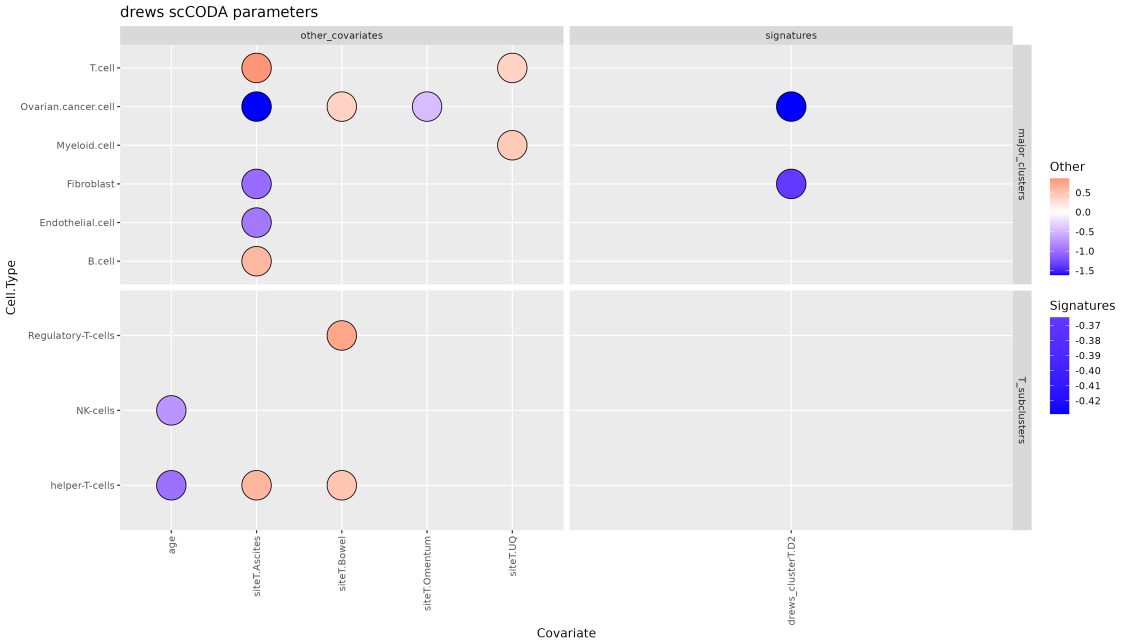


Figure A.3: scCODA parameters inferred for Drews clusters

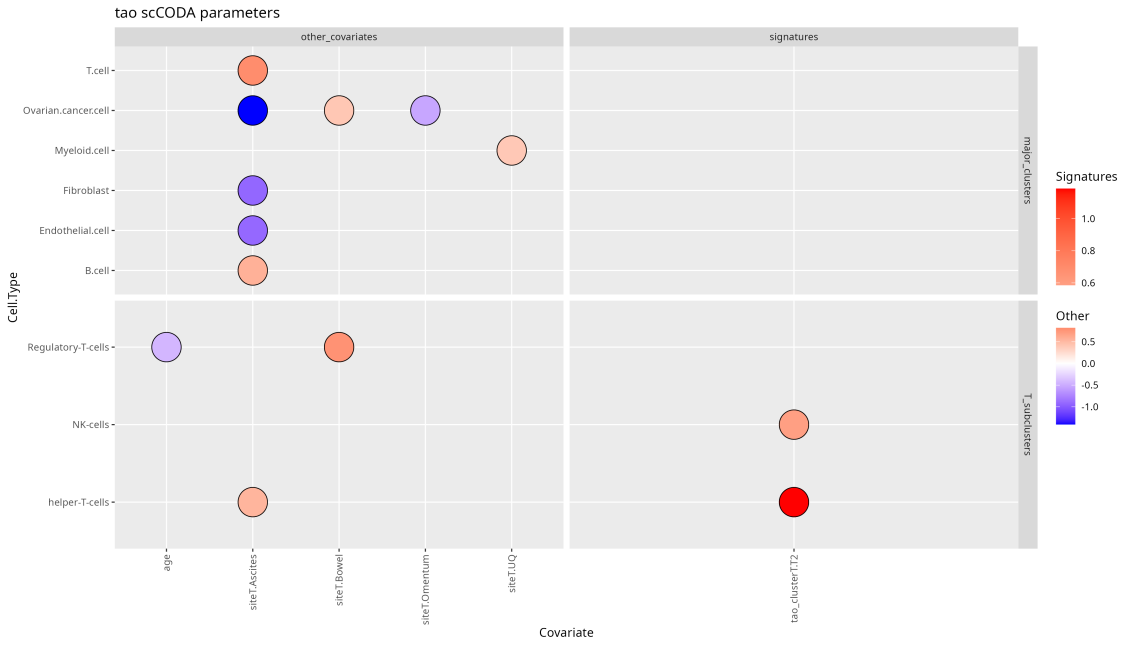
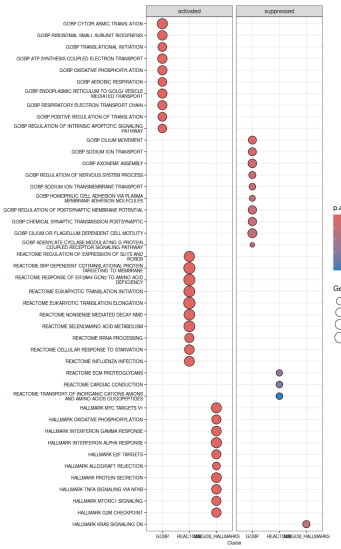
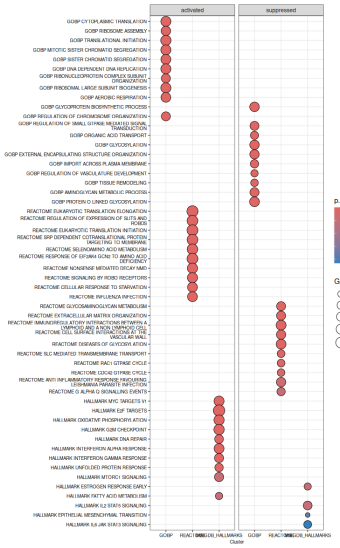


Figure A.4: scCODA parameters inferred for Tao clusters

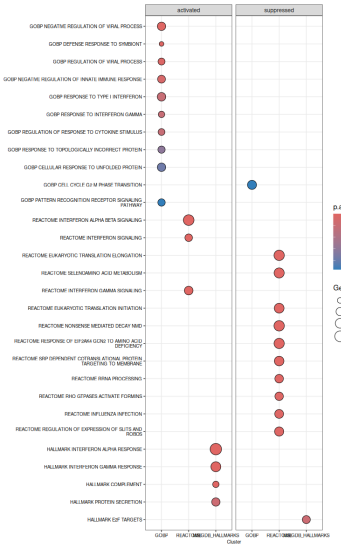
(a) Enriched terms in cancer cells



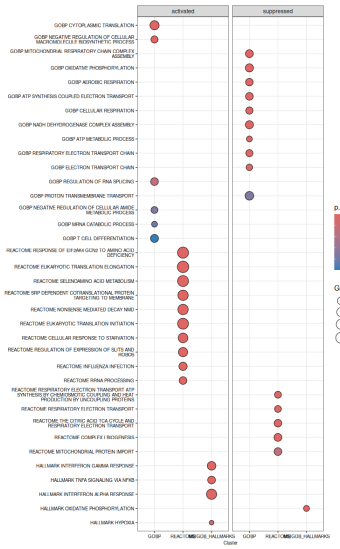
(b) Enriched terms in fibroblasts



(c) Enriched terms in Regulatory T cells



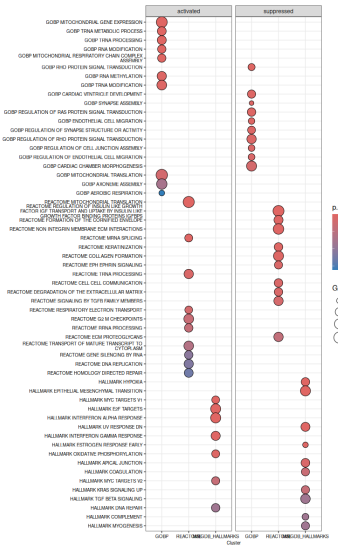
(d) Enriched terms in natural killer cells



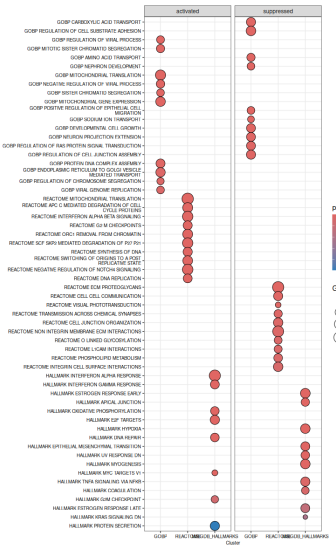
(e) Enriched terms in helper T cells

(f) Enriched terms in cytotoxic T cells

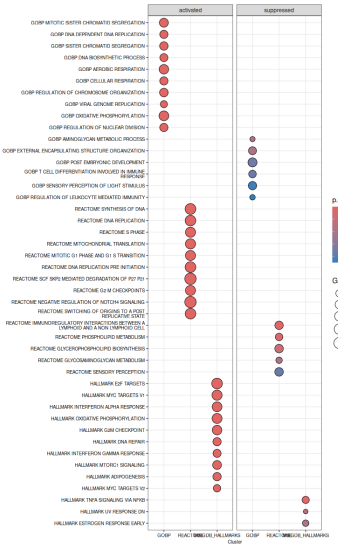
Figure A.5: Top enriched terms in the DGE comparison between U and HU patients, in primary samples. The dots are colored by p value, while the size is based on the overlap between the original gene sets and the genes in our data.



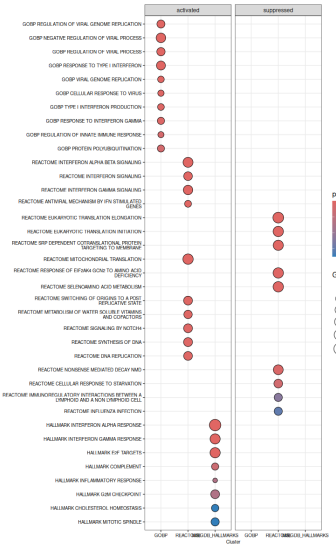
(a) Enriched terms in cancer cells



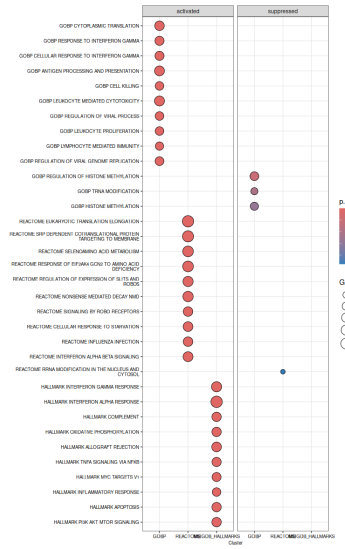
(b) Enriched terms in fibroblasts



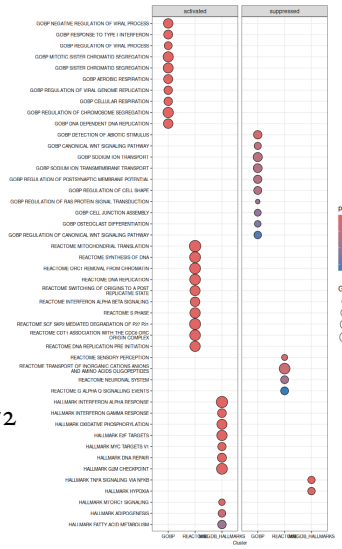
(c) Enriched terms in Regulatory T cells



(d) Enriched terms in natural killer cells



(e) Enriched terms in helper T cells

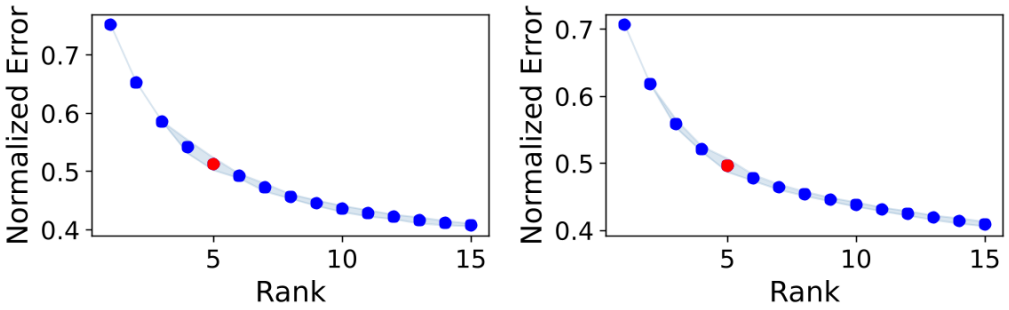


(f) Enriched terms in cytotoxic T cells

Figure A.6: Top enriched terms in the DGE comparison between U and HU patients, in metastasis samples.

(a) Primary samples **(b)** Metastasis samples

Figure A.7: Top enriched terms in the comparison between D1 and D2 patients, in cancer cells



(a) Primary samples **(b)** Metastasis samples

Figure A.8: Elbow plots of the Tensor-cell2cell decompositions. The x axis is the number of factors, while the y axis is the Frobenius distance between the reconstruction and original tensor. The chosen number of factors is highlighted in red

Acknowledgments

First, I would like to extend my thanks to professor Romualdi for giving me the chance to work on this project and for her invaluable guidance.

I would also like to thank everyone else in the Romualdi lab, especially Angelo and Ilaria for answering my (many!) questions and supporting me throughout this project, and Elena for being a great colleague and coworker as we both worked on our theses.

I would not be here without the personal and economical support of my family, which I thank.

Last but not least, I would like to thank all the people and friends I got to know during my university years, especially Lupo and Pizzo for being my friends since my first years here, Alessia for being the best friend someone could ask for, Anna for a brief but wonderful time sharing an apartment and room, and Max for showing me how to be proud of myself. I would not be where I am (and who I am) today without all of you.