Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Digging deep into intratumoral heterogeneity in ovarian cancer using scRNA-seq data

Bachelor Thesis

Antoine Combremont
antoinco@ethz.ch

Boeva Lab, Laboratory for Computational (Epi-)Genetics
Department of Computer Science D-INFK, Institute for Machine Learning
ETH Zürich

Supervisor:
Prof. Dr. Valentina Boeva, Ph.D. Josephine Yates

31/03/2022

Abstract

Research in computational biology boosted discoveries of the underlying mechanisms of cancer. It shed light on the heterogeneous nature of this disease, allowing the identification of cells and their transcriptional activity. This information is key to studying the evolution and resistance of the tumor and provides a solid foundation for the development of treatment.

In this project, we explore tumor heterogeneity of Ovarian Cancer using the database of single-cell RNA-sequences (scRNA-seq). We conduct a scRNA-seq analysis from the raw count to the interpretation of observed biological signals, and use tools developed by the BoevaLab and based on existing technologies. We aim at finding gene signatures to help future work on Ovarian Cancer research.

Contents

Chapter 1

Introduction

The general public is well aware that cancer represents a high risk of mortality. Since it accounts for more than 20% of deaths in the 70+ group in Switzerland according to the Swiss Federal Office of Statistics, patients over a certain age are screened by their physicians to detect some of them as soon as possible. While this group of diseases is a well-known public health issue, it remains difficult to cure for most cancer types and often involves heavy treatments which degrade patients' life quality (e.g., chemotherapy, various types of immunotherapy, radiation therapy). The difficult task of finding efficient treatment lies in the huge complexity of molecular interactions driving cancers and in the heterogeneous nature of this vast family of diseases [1]. Indeed, we count more than 100 types of cancer and each of them may have intrinsic properties which do not translate from one to the other.

Cancer genomics is the field of study of the genetic and epigenetic properties of cancer. It is a promising research trend [2], and it offers the possibility to make observations that go far beyond what is clinically possible. Thanks to advances in sequencing technologies and computer science, it is now possible to capture highly detailed information about cancer tissues for an affordable price. The vast amount of generated data gives the possibility to perform data-driven analysis and enables the generation of results that were previously simply impossible to compute.

The goal of this project is to explore the genetic heterogeneity of Ovarian Cancer by validating known signatures and characterizing new observations of clusters of malignant cells. To achieve this, we performed a complete scRNA-seq experiment on a privately available dataset, starting from raw count matrices of Ovarian Cancer samples, to the interpretation of biological programs. We use a new pipeline developed by the BoevaLab, which integrates a range of computational techniques, and test if it is possible to validate signatures of malignant cells from literature as well as if it is possible to find unique signatures and infer their biological meaning. This work provides the foundation for further research on this private dataset, allowing future projects to refine the observations made here. Finally, this report provides a use case for the pipeline, showcasing of it

can be used to integrate and interpret a cancer scRNA-seq dataset.

1.1 Cancer and ovarian cancer

Cancer is characterized by abnormal cell growth in one tissue, which may spread to other body regions, resulting in what is referred to as metastasis [3].

This disruption leads to the formation of tumors (sometimes also referred to as neoplasm), the generic term for describing a group of cells that have grown without regulation. For readers with little biological background, it should be emphasized that a tumor is the result of an uncontrolled cell growth that can be benign (non-cancerous), or malignant (cancerous). A benign tumor does not spread and remains local, causing at worst local manifestations since it occupies space and potentially compresses its neighborhood. On the opposite, a malignant tumor gains the ability to grow and at the same time the ability to invade its neighborhood (not only compressing) and spread distantly (through blood and lymph vessels) in other body regions.

The loss of a cell's faculty to regulate its growth can be caused by genetic and epigenetic alterations. Genetic alteration is a change in the DNA sequence of the cell. In the case of tumoral cells, the changes occur in genes involved in growth regulatory pathways. We classify the affected genes into two categories. On one hand, oncogenes promote cell growth and reproduction, and on the other hand, tumor suppressor genes inhibit cell division and survival. Epigenetic alterations are another type of cancer driver, allowing (or preventing) certain genes to be expressed (or not) without any alteration in the DNA sequence.

1.1.1 Ovarian Cancer Types

The different types of ovarian tumors originate from the main three types of cells of the ovaries. The cells give the names of the tumor: epithelial tumors, germ cells tumors, and stromal tumors, from highest to lowest incidence rate (figure 1.1). Each type can be benign (non-cancerous), borderline (with low malignant potential), or malignant (cancerous), meaning they will gain the ability to spread to other organs, making it particularly hard to cure and increasing the risk of death.

There exist two additional types of ovarian cancer very close to the epithelial carcinoma, namely Fallopian Tube Cancer and Primary Peritoneal carcinoma. This report will focus on Epithelial Carcinoma and its subtypes, and we will not discuss other types.

1.1.2 Epithelial Carcinomas

Ovarian epithelial cancer tumors are called epithelial carcinomas, and it is by far the most common, with more than 85% of malignant ovarian cancers. There are 4 subtypes of epithelial ovarian carcinomas, based on tumor histology: Serous (52%), Endometroid (10%), Clear cells (6%), and Mucinous (6%) [4].

	Epithelial						
	All epithelial subtypes	**Serous**	**Endometrioid**	**Mucinous**	**Clear cell**	**Sex cord-stromal**	**Germ cell**
All races	9.4	4.9	1.0	0.6	0.6	0.3	0.4
Non-Hispanic white	10.0	5.2	1.1	0.7	0.6	0.2	0.3
Non-Hispanic black	6.9	3.4	0.5	0.4	0.2	0.5	0.4
American Indian/Alaska Native	8.3	4.3	0.9	0.5	0.4	0.3	0.2
Asian/Pacific Islander	7.8	3.4	1.1	0.6	1.0	0.1	0.4
Hispanic	8.1	4.0	0.8	0.5	0.4	0.2	0.5

Figure 1.1: Incidence rates of histological types of ovarian cancer per 100,000, age adjusted to the 2000 US standard population, US, 2010-2014 [5]

We classify each of the subtypes into Grade I and Grade II, low and high, each denoting whether the tumor resembles normal tissue or not, which is an indicator of good or bad survival prognosis. We also distinguish two types, where Type I characterizes a slow tumor growth and causes fewer symptoms but no response to chemotherapy, and Type II characterizes a faster growth and spread, but shows a better response to chemotherapy. It is interesting to note that genetically, the later type shows P53 mutations and genomic instability due to defects in DNA repair mechanisms.

The serous subtype is the most common type of epithelial carcinoma and originates from the tubal epithelial cells. It is important to note that the low-grade and high-grade serous subtypes are understood as different neoplasms with distinct modes of carcinogenesis, molecular-genetic features, and sites of origin.

Type II neoplasm, High-Grade Serous Ovarian Carcinoma (HGSOC) is of particular relevance and we present some relevant details. It accounts for 70–80% of deaths from all forms of ovarian cancer. Because it is commonly only diagnosed at a late stage, HGSOC shows a 5-year survival rate of about 47% in the US [6]. Several investigations have demonstrated that HGSOC is molecularly heterogeneous and comprises at least 4 molecular subtypes [7]. These subtypes have been previously described in The Cancer Genome Atlas (TCGA) and are called differentiated, proliferative, mesenchymal, and immunoreactive according to their associated gene functions [8].

1.2 Tumor Heterogeneity

Because of cancer's heterogeneous nature, it is impossible to establish a valid and more specific characterization among them. Each type (and subtypes as previously seen) of cancer is different to some extent and even varies across patients that show *a priori* the same types of tumor. Further, it evolves and responds and adapts to its environment (for example, when exposed to chemotherapy). We are thus interested in knowing what is going on inside the tumor, and quantifying the different populations of the tumor environment (figure 1.2).

We are particularly interested in observing if certain differences or similarities exist across patients from the same histological subtype, but also between cell populations from the same tumor. We call it intertumoral and intratumoral heterogeneity (or also interpatient and intrapatient heterogeneity). Tumors are complex tissues with local differences, and as was previously mentioned cancer evolves. Therefore we can also observe what is called spatial and temporal heterogeneity.

The general approach for studying tumoral heterogeneity is to find out what genes are expressed by tumor cells since gene expression ultimately translates into modifications of the functionality of cells via interactions between the proteins encoded by those genes (these interactions are also called pathways). Knowing about the dynamics inside tumors finally opens the door to more targeted and efficient treatments. Indeed studies have attributed resistance to therapy and relapse to be caused by only a subset of the entire tumor environment, thus identifying and addressing them directly is a promising research direction.

Tumor heterogeneity in epithelial cancer has been the object of several studies and it has been generally described as an important feature of this type. In particular, it has been emphasized that the current histological classification of subtypes is not sufficient to treat patients effectively, suggesting the need for further understanding of the intratumoral heterogeneity of epithelial ovarian [9].

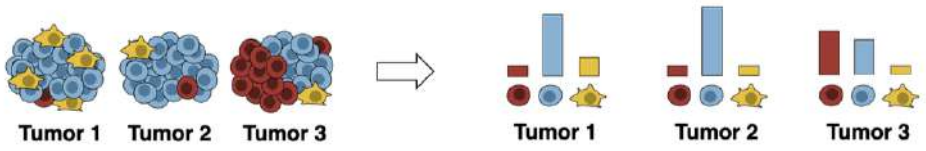


Figure 1.2: Quantifying tumor heterogeneity (figure by Valentina Boeva)

1.3 Batch effects in scRNA-seq Experiments

1.3.1 scRNA-seq

We call transcriptome the set of mRNA sequences that are expressed by a cell. Access to the transcriptome can inform us about which genes are expressed by a cell. After capturing the transcriptome of a significant number of cancer cells from multiple samples and patients, we can then explore expression patterns to understand better the tumoral environment.

We briefly describe the typical process involved in a scRNA-seq experiment [10]. It starts with the extraction of tissue performed by clinicians, before the count of mRNA sequences found in each cell of the sample using specialized sequencing technologies. To count them, specific molecules called unique molecular identifiers (UMI) are appended to the mRNA sequences. The sequences are then amplified using the PCR method and then counted by sequencers. Once digitalized, it is possible to approximate the original transcriptome, using the UMIs as markers. This results in a zero-inflated count matrix, where rows denote cells, columns denote genes, and components are counts. We will not discuss more details of this process typically performed in medical biology laboratories.

The analysis of count matrices is a task of computational biologists and raises many challenges [11]. Many of them stem from the noise involved in the sequencing process. The output for each sample is a very sparse matrix and detecting the biological signals contained in the tissue requires careful choice of processing decisions.

1.3.2 Batch-effect correction

The capture of transcriptomic information is subject to a lot of variabilities, such as equipment, handling personnel, laboratory protocol, and technology platforms to name a few [12]. These factors introduce noise in the datasets, and it is of importance that they are not mistakingly taken as biological signals. This variability is referred to as batch effect and a key component to performing an accurate analysis of a scRNA-seq experiment is to apply batch-effect correction on the dataset before performing downstream analysis.

There exist several tools that will allow correcting for batch effect. The current approach that has shown promising results is to use conditional variational autoencoder (CVAE), a neural-network architecture that allows learning latent distributions of the data, while automatically correcting unwanted sources of variance such as sample-specific profiles and cell cycle effects [13] [14]. In this project, we used the python framework scvi-tools that implements such a method. We provide more details on CVAE, representation learning, and SCVI in section 2.3.

Chapter 2

Methods

2.1 Experimental Setup

We summarize here the origin of data and the coding environment. If not part of the pipeline, or if not specified otherwise, the scripts and interactive notebooks used for processing, file, or figure generation can be found in the shared drive of the BoevaLab at *"/data/projects/Antoine_Tumor_heterogeneity"*. The file "README.md" (also in appendix) serves as a guide for the organization of files in the directory. We used the cluster LeoMed for running our experiments and storing our data. LeoMed is a secure part of ETHZ's Leonhard computing infrastructure, dedicated to confidential data (in our case, personalized medicine data). All data related to this project are exclusively stored on LeoMed (see 2.1.2). The main steps of the analysis were performed using python scripts and interactive Jupyter Notebooks on LeoMed. The scripts mainly interface with the python pipeline developed by the BoevaLab (unpublished at the time of writing) and integrate other open-source libraries, on which the pipeline itself is also built (see 2.1.1).

2.1.1 Code

We used a processing pipeline developed by members of the BoevaLab, which can be found on the GitHub page of the lab. This tool builds upon several pre-existing python libraries widely used and maintained, implementing single-cell experiments methods. This includes scanpy, inferCNVpy which is a python implementation of InferCNV , scvi-tools (which is itself built on the machine learning library pyTorch), and GSEA [15] [16] [17] [18] [19]. Understanding the details of some of those tools is critical to understanding the processing of the dataset, and we will present relevant ones in the next sections.

The goal of the processing pipeline is to be able to quickly and efficiently perform the analysis of scRNA-seq datasets to validate and eventually find *de novo* signatures for ovarian cancer. It provides a user-friendly interface, allowing the modular design of scRNA-seq experiments.

The works in three main phases, first, the preprocessing where cells are filtered based on quality criteria, the modeling where we correct for batch effects, and the postprocessing which consists in obtaining metrics on the gene expression levels to relate them to biological interpretation. We will describe these steps in detail, explaining the different parameters that were used in this project.

2.1.2 Data

We describe the origin of the data that were analyzed in the subsequent steps. It consists of a set of ovarian carcinoma samples provided by the Tumor Profiler study (TuPro), a consortium that aims to exploit the potential of tumor profiling data to enable application in personalized medicine [20].

Sample Selection

A preliminary task consisted in finding which sample to include in the analysis. All the files containing data relative to ovarian carcinoma can be found in the subfolders with names starting with "TP-G". However, due to the vast amount of data generated for each sample and each experiment of TuPro, we used the metadata files *OvarianCancer_Samples.csv*, *scRNA_Analysis.csv* and *Participants.csv* to select the appropriate samples and their location.

We combined these metadata to obtain a consistent metadata file, containing clinical information and the path to the raw count and cell annotation files on the cluster for each sample. We merged metadata by selecting only the latest Run and Sequencing ID available (since we are not interested in having more than one version per sample, we take the most recent possible). The interactive notebook to generate the metadata file can be found under the name "sample_selection.ipynb".

Ovarian Cancer Dataset

We highlight here the main characteristics of the dataset, based on the metadata we generated. There is a total of 48 samples, and we note is that the sample OMABOPY turns out to be annotated as "Tubal", namely Fallopian Tube cancer. Since we are not interested in this type we discarded this sample from the dataset. Finally, the sample OTAMAZA did not pass our preprocessing (explained in section 2.2) because it had too few cancer cells. This leaves a total of 46 samples for analysis.

The average age range of the dataset is 60-69, reflecting the average population. More than half of the samples are known to be metastasis (27 samples) whereas only 4 samples were taken from the primary tumor. The location of the sample

is mainly from the omentum (18 samples), which is formed by two layers of peritoneum that surround abdominal organs. All primary tumors are annotated to originate from the ovaries, except two from the peritoneum. We lack information about the stage (37 unknown samples), but the few annotations show more IIIC (5 samples) and IV (3 samples). The metadata file *"sample_info.csv"* can be found on the cluster in the directory *"/cluster/dataset/boeva/scRNA/raw/ovarian/"*.

NEXUS Annotations

Since any type of cell from the tumor microenvironment can be in our data (remember a tumor or metastasis is not only defined by tumor cells but also by complex interactions with other tissues in its surroundings), we want to annotate them according to their phenotypes to separate healthy cells from tumor cells. This can be done using Copy Number Variation inference (CNV) as we will describe in the next section. The NEXUS Personalized Health Technologies Platforms at ETH provided annotation for the ovarian dataset, using their methods. This will serve as a basis for comparison to filter cells in the preprocessing step.

2.2 Preprocessing

We now have a very large dataset, containing 46 matrices of each sample's *genes*$\times$ *cells* counts. This represents a total of 125'800 cells, with more than 20'000 genes per cell on average. Not all of these cells can be used since the sequencing process has a certain percentage of failure. We thus want to get rid of bad cells to spare computation costs and avoid misleading noise. This is the first step in the pipeline, and more generally the starting point of every scRNA-seq analysis project.

2.2.1 Quality Control

To start with, we can assess cell quality based on a few criteria. Quality criteria include the number of counts per cell, the number of genes per cell and the fraction of mitochondrial genes per cell [21]. For example, it can happen that a cell was counted twice, resulting in an abnormally high number of counts, and eventually higher gene counts as well. Another very important effect is when the cytoplasmic membrane tears apart, resulting in the loss of most of the cytoplasmic content (in particular the mRNA strands floating in it). Then the bigger organelles such as mitochondria would stay in the cell and create a biased high mitochondrial count percentage.

To filter those cases out, we first apply "*scanpy.pp.calculate_qc_metrics*" val-

ues of interest. The scanpy method "*scanpy.pp.filter_cells*" was used, and filtered cells with less than 1500 counts, more than 50′000 counts a percentage of mitochondria genes higher that 35% and less than 700 genes. See figure 2.1 for a visualization.

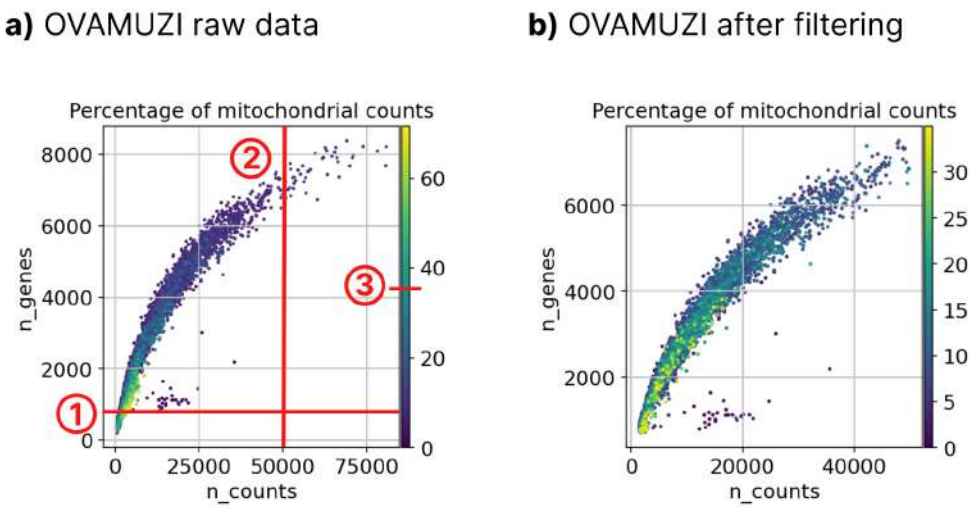


Figure 2.1: On both plots, a cell is represented by a dot, which coordinates represent the number of genes and the count depth and which color represent the percentage of mitochondrial RNA among all genes. The threshold 1, 2 and 3 are applied in prepossessing to filter the cells. There are 5′420 before filtering and 4′546 after.

2.2.2 InferCNV

Our dataset currently contains different cell types, some of them might be healthy, which in our case we want to get rid of. We are indeed interested only in cancer cells, as we want to find cancer-specific signatures in our dataset. Previous studies have shown that genomic instability such as Copy Number Variation (CNV) is correlated with differential gene expression, especially for oncogenes and tumor suppressor genes [22]. Using the information on the CNV of cells, we can find subclones in our dataset, and those with important CNV are likely candidates for being cancer subclones.

We use a probabilistic tool called inferCNV (a package for the language R) for which a python version called "inferCNVpy" exists. InferCNV was developed based on studies, which confirmed the potential of using CNV to infer subclonal structure [23]. From a given single-cell sample, the program is capable of reconstructing the CNV profile of that sample. We can then annotate

cells according to their subclone, where we consider cells to be "unhealthy" when they belong to subclones with high CNV. This annotation can then be compared against NEXUS, to finally decide if the cells belong to the category "unhealthy", "healthy" (when both annotations agree on the label), or "undecided" (when there is an annotation conflict).

InferCNV uses a corrected moving average of gene expression data to determine CNV profiles. Genes are sorted by absolute genomic position, which is passed to the program using a file encoding the positions. To further refine the CNV profile of tumor cells, InferCNV constructs the CNV profile of a reference cell, and then for each gene of each other cell, the reference profile is subtracted to determine the final CNV profile (if no reference cells are provided it will simply use the average of all cells as the reference profile).

2.2.3 Gene Signatures Scoring

The last step included in preprocessing of the pipeline is to score gene signatures that were passed as arguments. Those signatures are lists of genes that were found by other studies as related to certain programs in cancer. They are used as a comparison point for downstream analysis, and we usually inspect them visually (using UMAP representation). They provide a reference to assess if there is a corresponding organization of the space.

To interpret the organization of cells based on their transcriptome, we need a way to measure the expression of genes from a given signature for each cell. A signature score is an overall measure of expression for each cell in the dataset, and it can be computed in various ways. A naive approach is to take the sum or average gene expression in each cell, but this would suffer from bias introduced by technical noise. Here we use the method *"scanpy.tl.score_genes"*, which implements a more advanced approach [24]. It places all expressed genes in a certain number of bins. Then, for each gene in the signature, a set of control genes from the same bin are selected. The average expression of the signature genes is then subtracted by the aggregated expression of control genes. This reduces the effects of cell quality and complexity on the scoring.

In this work, we scored signatures from two different sources to compare our observations. The first comes from unpublished work from Ph.D. Agnieszka Kraft, which methods aim at predicting the shared intratumor transcriptional heterogeneity on bulk RNA data (another very popular way to capture transcriptional information). In her work, three distinct signatures were discovered on ovarian cancer data from The Cancer Genome Atlas program (TCGA). We also considered the signatures provided by a recent scRNA-seq study, the "Hao study" [25]. It characterizes different categories of cells, including malignant epithelial cells, which are of particular interest in the study of Epithelial Ovarian Cancer. We will introduce these signatures in more detail in chapter 3.

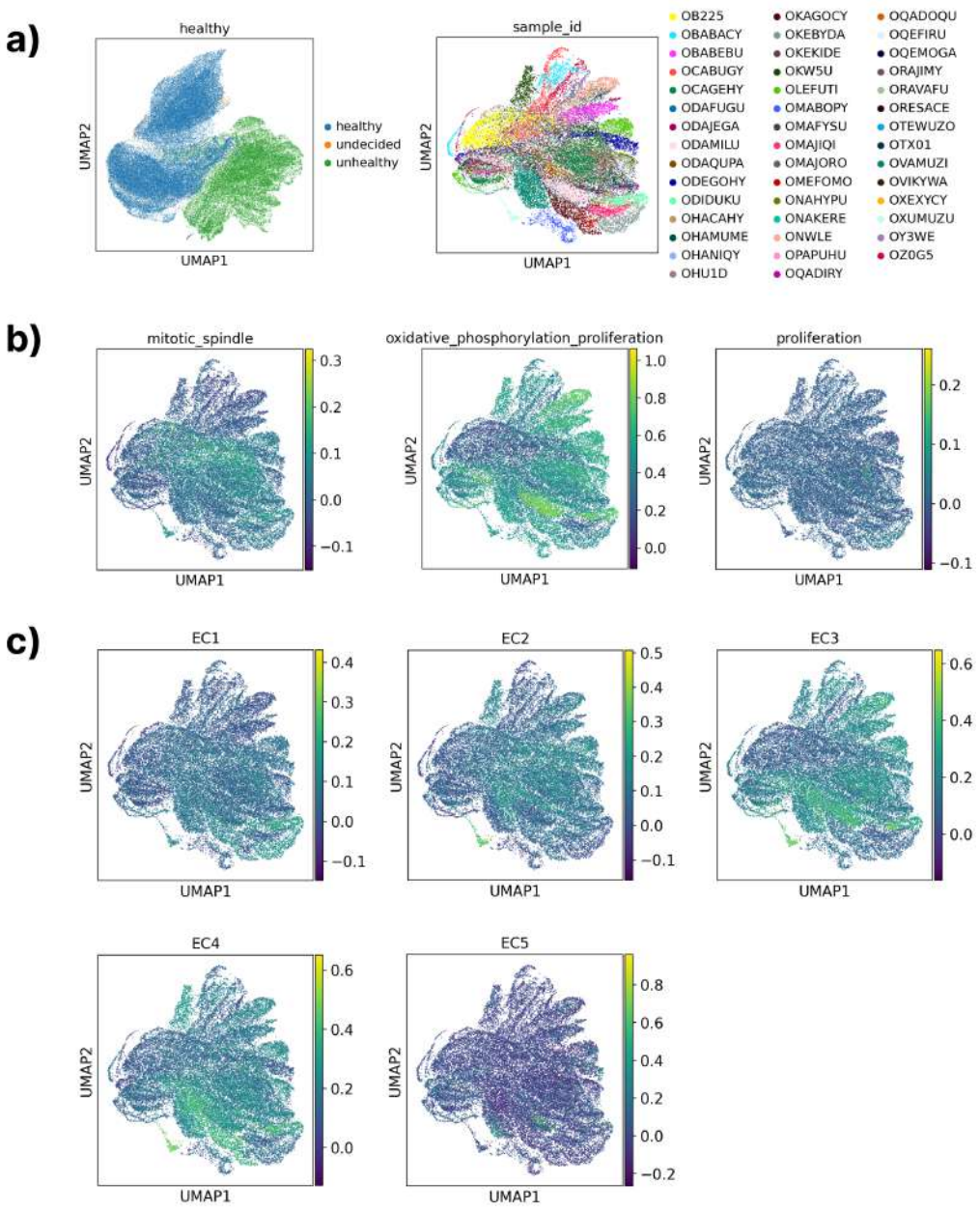


Figure 2.2: **a)** We show the entire dataset after preprocessing (label "healthy") with healthy cells in blue, unhealthy cells in green and undecided cells in orange. The rightmost plot shows the unhealthy cells, colored by sample id (note that the cells are grouped according to sample id in the UMAP representation). **b)** The unhealthy cells with the score of the signatures from Agnieszka's work (the color intensity represents the score of a cell). **c)** Similarly, we show a UMAP representation of the signatures from the Hao study.

2.3 Modeling

Modeling is a crucial part of the pipeline, in this step we seek to reduce batch effects to avoid misleading trends to propagate to downstream analysis. The approach we are using here is to learn a low dimensionality representation of the dataset. This means to find a space of size k such that $k << n_{genes}$ (where n_{genes} is the input dimension, equal to the number of genes expressed per cell). From this space, we can extract the signals from technical noise and batch effects using probabilistic methods. We outline here the theory as well as the tool used to perform this step.

2.3.1 Conditional Variational Autoencoder

The introduction in the previous paragraph intuitively describes the model of Conditional Variational Autoencoder (CVAE). We assume the reader has some familiarity with Autoencoders (AE) and give a brief refreshing of Variational Autoencoder (VAE) since it takes only one step to derive CVAE from VAE.

We recall that the architecture of VAE is similar to the one of AE, where an intermediate representation of the data is learned by the encoder in a bottleneck, and the input is reconstructed with minimal loss by the decoder. However, the VAE is interested in learning probabilistic models of the distributions of datasets [26]. That is, while AE will reduce the representation to arbitrary dimensions (often overfitting), the VAE bottleneck is organized into meaningful dimensions. The input of the encoder is an observation, defined by the parameters for specifying the posterior distributions that were passed in the training process. The VAE encodes the input as parameters of the distribution of the data, yielding the latent distribution $p(z|x)$ (usually a Gaussian). The decoder in VAE is $p(x|z)$, namely, the description of the decoded variable is given the encoded one. Finally in VAE we assume z to follow a (Gaussian) distribution $p(z)$, so using Bayes we see that we only lack $p(x)$ to obtain $p(z|x)$:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int p(x|t)p(t)dt}$$

Since the denominator is intractable, we use the technique of variational inference to find an approximation (hence the name). Using such method, we can approximate $p(z|x)$ with Gaussian $q_\lambda(z|x)$ with $\lambda_i = (\mu_i, \sigma_i^2)$, minimizing the Kullback-Leibler (KL) divergence to constrain the distribution found by the encoder to be close to the "real" one. The regularization term using the KL divergence is:

$$q_\lambda(z|x) = \arg\min_\lambda \mathrm{KL}(p(z|x) \parallel q_\lambda(z|x))$$

Since the neural network used for the approximation requires backpropagation for training, it is important to mention that the "reparametrization trick" is used, thus separating the stochastic node with the bottleneck (see figure 2.3).

After some derivation the loss function of VAE can be written as:

$$\mathcal{L}_{VAE}(x;\lambda) = \mathrm{KL}(q_\lambda(z|x) \parallel p(z)) - \mathbb{E}_q[logp(x|z)]$$

where the expectation term is the reconstruction error, and the KL is the regularization.

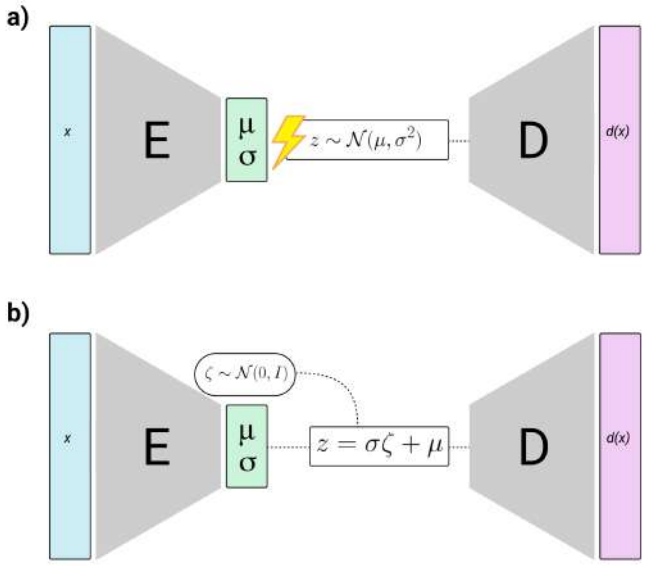


Figure 2.3: Reparametrization trick: **a)** Sampling without reparametrizing the stochastic process in a separate node does not allow for back propagation **b)** With a separate node for generating the (here Gaussian) stochastic value allows to use backpropagation on neural networks E and D

We now have a probabilistic low dimension latent representation of our data, what we want next is to add conditions to the latent distribution as well as on the output [27]. Conditioning the VAE on a set of conditions c means that the latent variable is distributed following $p(z|x)$ instead of $p(z)$ only and thus the encoder doesn't learn about c. The loss is adapted from the VAE:

$$\mathcal{L}_{CVAE}(x,c;\lambda) = \mathrm{KL}(q_\lambda(z|x,c) \parallel p(z|c)) - \mathbb{E}_q[logp(x|z,c)]$$

With this approach, we can force our network not to model the different samples id's, or any variable we know might introduce bias in our data. The latent space that results is thus organized and also not disturbed by those technical variables we take as unwanted in the context of our experiment.

2.3.2 SCVI

We used the model Single Cell Variational Inference (scVI) from the library scvi-tools for the modeling task [17]. What scVI does is to model the expression x_{ng} of each genes in each cells as sampled from a zero-inflated negative binomial distribution (ZINB) $p(x_{ng}|z_n, s_n, l_n)$ conditioned by variables s_n and l_n. The variable s_n denotes the batch annotation of the cell (in our case the sample id), and l_n a Gaussian simulating noise due to technical variations during measurements.

We use scVI because it is highly scalable, provides in-depth documentation, and also has a modular API that allows the creation of new probabilistic models. While there exist other batch-effect correction techniques, the advantage of conditional variational inference is the performance and flexibility it offers. The goal for the pipeline is also to provide a platform for implementing, adding, and testing new models, and the conditional algorithms offer a broad range of possible innovations.

In our experimental setting, we modeled the data with three different configurations, with 4 to 10 latent dimensions respectively (the configuration files are named "latent4.yaml" to "lantent10.yaml" and can be found in the shared drive). Since the learning process is not deterministic, we compute the latent spaces of each different configuration with 5 random seeds. This allows us to assess if the compression is meaningful if we have a similar organization of the latent space.

2.4 Postprocessing

This is the last step in the pipeline, where we finally start applying methods to get quantitative results about gene expression. We first perform unsupervised clustering on the latent representation of the dataset and then compute differential gene expression between each cluster to be able to characterize the clusters, and eventually compare them with other studies. We perform gene set enrichment analysis (GSEA), a method that will compute the statistical significance of pre-defined sets of genes, allowing assessment of concordance with biological characteristics [19]. We present here the methods and will discuss the results in the section that follows.

2.4.1 Clustering

Clustering is a technique that aims at grouping data points in the space they lie in according to some metrics. This partitioning finds use for the interpretation of high-dimensional data and highlights the similarities between the groups of data points. To achieve good clustering we must find the best way to capture the underlying structure of the data points, which depends on the nature of the data. This is an example of unsupervised learning, as the goal is to infer a label (i.e.

cluster) to each data point without prior knowledge, only by detecting similarities between the points. Defining the appropriate similarity criteria is important and there exists a few popular types of algorithms such as connectivity-based, centered-based, distribution-based, and density-based algorithms.

In our task of discovering tumor markers and gene signatures in ovarian cancer, we want to see how the single cells organize in the latent space we modeled using methods described in previous sections. By carefully choosing algorithms and parameters, applying unsupervised clustering methods on our dataset will show us communities of cells that would have been difficult or impossible to distinguish otherwise.

We use the Leiden clustering algorithm to find clusters in the latent space of our dataset which was obtained after modeling. The Leiden algorithm is similar to Louvain but performs better at detecting non-badly connected communities in large graph structure [28]. We construct the graph from the latent space representations using the method *scanpy.pp.neighboors* in scanpy, which is itself heavily relies on UMAP [29]. We then perform Leiden clustering with automatic resolution search, implemented in the pipeline by a binary search for the optimal value. The underlying implementation of Leiden is provided by the method *scanpy.tl.leiden* in scanpy.

We evaluated different sizes of clusters, and thus for each latent representation, we ran clustering with three different numbers of clusters, 6, 8, and 10. These numbers allow different levels of granularity for later interpretations and are relatively low since we expect a large number of cells to express similar pathways. There is indeed a tradeoff between a higher and lower number of clusters, as a too high number of clusters would miss more global trends which we are more interested in.

Community detection-based clustering is fast and preserves the structure of the high-dimensional data. It has also been shown that this approach appears to perform among the best in benchmarks [30]. Finally, the *scanpy* API makes it extremely simple to use, thus justifying our use of Leiden.

2.4.2 Differential Gene Expression

We have clusters that were found in an unsupervised way. We now want to find out the meaning behind this organization of the latent space, and a classical method to do so is to analyze differential gene expression (DGE), which consists in quantifying the difference in gene expression in the transcriptome of different sets or different cells.

DGE analysis has previously been used in bulk-RNA and microarray studies and as a standalone tool for interpreting gene expression and cell-type annotation in scRNA-seq analysis experiments. The major difference in single-cell data is

that the cell-level measurements are much more variable than in bulk (since there can be different types of cells, each expressing their particular transcriptome), and methods model the counts as distributions. DGE tests for statistical differences between two groups. For example deciding whether, when comparing a given gene in a cell to others, an observed difference in read counts is greater than what would be expected just due to natural random variation. This can be done for each gene in a cluster against the rest of the genes in other clusters, resulting in a list of differentially expressed genes (DEG).

After having normalized the data, we use the scanpy implementation *"scanpy.tl.rank_genes_groups"* based on [31], to compute DGE on the cluster obtained in the previous section. We obtain lists of genes, which are statistically differentially expressed in each cluster compared to the others, as explained in the above paragraph.

2.4.3 Gene Set Enrichment Analysis

Because of the high number of DEG, we need a technique to understand it as a bigger system. We use a method is called gene set enrichment analysis (GSEA), where we try to determine if a predefined group of genes show a statistically significant and concordant difference between two biological states. That means, for each cluster we use the previously computed lists of DEG, and run the GSE algorithm to compute concordance between the list and the pathways we want.

Intuitively, the method consists in measuring the distribution of the genes from the pathways over the list of DEG. If most of them appear at the top or the bottom of the list, the pathway is considered to be phenotypically related to the gene expression of the cells in the cluster corresponding to the list (see figure 2.4).

The first step of the method is to compute the Enrichment Score (ES) for a given list of DEG pathways, which indicates if the genes in the pathway are either over-represented at the top or bottom of the list of DEG. To compute the score, the algorithm walks across the ranked list and increments a running-sum statistic when a gene from the pathway is encountered in the list. The final result corresponds to the maximum deviation from zero encountered genes. We can then calculate the Normalized Enrichment Score (NES), which accounts for differences in gene set size and correlations between gene sets and the expression dataset:

$$NES = \frac{Actual\ ES}{mean(ESs\ against\ all\ permutation\ of\ the\ dataset)}$$

The permutation creates a null distribution (intuitively, the ES is normalized by a truly random list, namely the null distribution that does not correlate with genes). The null distribution serves to compute the nominal P-value of the ES, which

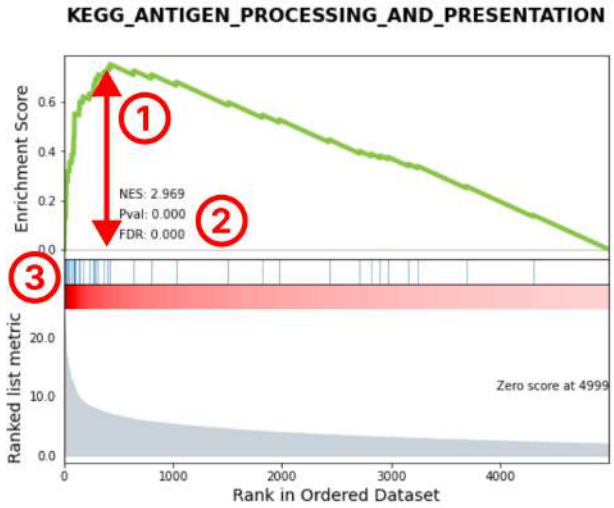


Figure 2.4: This plot illustrate an output of GSEA (the figure is generated using the plot function from gseapy). We see in (1) the value of the Enrichment Score, corresponding to the maximal difference between 0 and the green curve (running-sum). The other values of interest are reported in (2). In (3) we see vertical lines representing the occurence of a gene of the gene set in the list of ranked genes (note that an occurence increments the running-sum).

is finally also a statistic of interest, which indicates the statistical significance of the ES.

We used the method *"gseapy.prerank"* and supply the list of DEG we computed earlier. We perform GSEA if the KEGG and Hallmark database, which can be downloaded from MSigDB [32]. We used the KEGG and Hallmark databases because they are well known and used in many studies, including the work of Agnieszka Kraft and Hao, which will be our comparison points for interpreting the final results.

Chapter 3

Results

In this chapter we will analyze the output of postprocessing and interpret the results, to see if some known signature can be found again to validate the method and assess the potential of new signals in the context of Ovarian Cancer. Note that all the results used in this section (latent representations, GSEA scores for each run) can be found on the shared drive.

3.1 Evaluation of the Latent Space

We visually explored the modeling generated by scVI to make sure we do not observe clusters or outliers, which would be interpreted as unwanted artifacts possibly caused by batch-effect or other noise. We validated that no cluster according to sample identifier can be observed, nor any cluster according to the location of the sample (primary tumor vs metastasis).

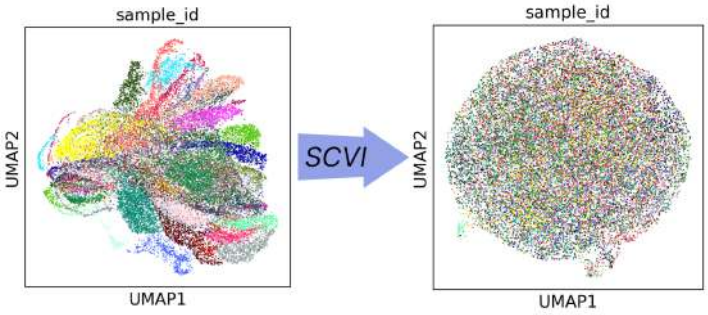


Figure 3.1: UMAP of dataset colored by sampled ID before modeling (left) and after (right). Cells are not mixed by sample ID anymore, and the organization of signatures (see fig. 3.6, fig. 3.7) and scored pathways (see appendix A.1 A.2) indicates that SCVI captures some biological information in the latent space.

However, we noticed two clusters of low count cells. After inspection, we noticed they were almost exclusively cells from samples OLEKESU and OCANYTE.

An additional low count cluster with mixed sample origin was also found. To exclude the hypothesis of low-quality cells, we ran the preprocessing with different count thresholds, once with a minimum count of 1500 and once with a minimum count of 3000. We observed no change in the latent space embedding, suggesting that the cell quality in those clusters is sufficient.

What sets the cluster for OLEKESU and OCANYTE apart, however, is their high score from the T-cells and B-cells signatures (from the signatures provided by the Hao study). We suggest this is caused by contamination of the sample, whose content was mixed with genes from immune cells wrongly. Therefore, we decided to discard these two samples for later analysis. The final modeling contains 44 samples, and the corresponding UMAP visualization shows good mixing, which indicates the batch effect is corrected effectively (figure 3.2).

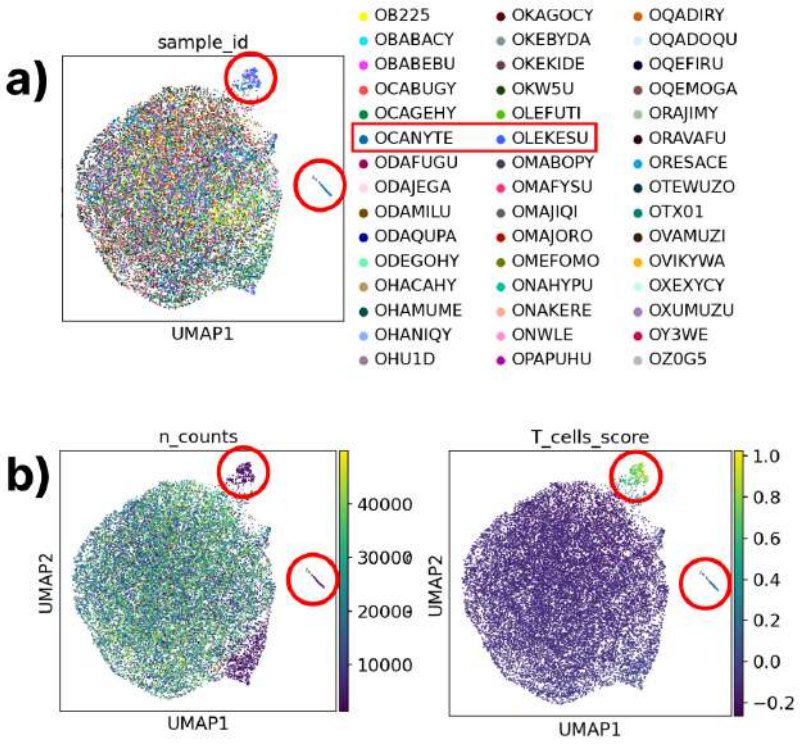


Figure 3.2: **a)** Dataset in 4 dimensional latent space embedding, colored by sample ID, **b)** Same representation colored by count depth and T-cells score (from Hao study). In red we highlight artifact samples, which are thought to be contaminated samples.

3.2 Unsupervised Signature Discovery

Because of the stochastic nature of the modeling process, we computed numerous different runs with the same parameters, but with a different random seed. We explain here the approach and describe the general trend we can observe after all those different runs in the stability analysis.

3.2.1 Stability Analysis

For our experiment, we compared latent spaces with dimension $d \in \{4, ..., 10\}$ and ran clustering with a number of cluster $k \in \{4, 6, 10\}$. For each run of the latent space modeling phase, we used 5 random seeds, this makes a total of 105 configurations. We computed GSEA for each cluster, using the database Hallmark Pathways and KEGG as gene sets.

To gain intuition on this large amount of information, members of the project developed a methodology we refer to as "stability analysis". This consists in counting for each latent dimension and each number of clusters, the number of pathways with $NES > 2$ (independent from which cluster the pathway was found). We can then observe the top-scored pathways across all results as in figure 3.3. The heatmap representation gives us information at a glance: each row contains the NES of the corresponding pathway for the parameters described by the respective coordinates. The pathways appear in decreasing order, from the overall top-scored pathway to the least scored pathway.

In figure 3.3 we observe that the count of NES is systematically higher than 2 for low latent dimensions and a high number of clusters for each pathway. This suggests that a high latent dimension fails at modeling relevant signals, and conversely, a low number of clusters does not offer enough granularity to capture the underlying biological trends. We thus choose to manually inspect in priority runs with parameters $d = 4$ and $k = 10$ (where d is the number of latent dimension and k the number of cluster). The other parameters were not considered as containing significantly more interpretable results, we include more visualizations in the appendix (figure A.4 A.3).

3.2.2 Characterization of Clusters

In this section, we want to showcase a specific result obtained after running postprocessing on a 4 dimensional latent space with 10 clusters found using the methods described in previous sections. We considered here the results of an arbitrary random seed, which latent representations can be found on the cluster at *"/cluster/dataset/boeva/antoinco/modelling/latent4/ovarian"* under the name *"20220320-095825-awake-magpie"*.

The top five results of GSE for the pathway databases KEGG and Hallmark

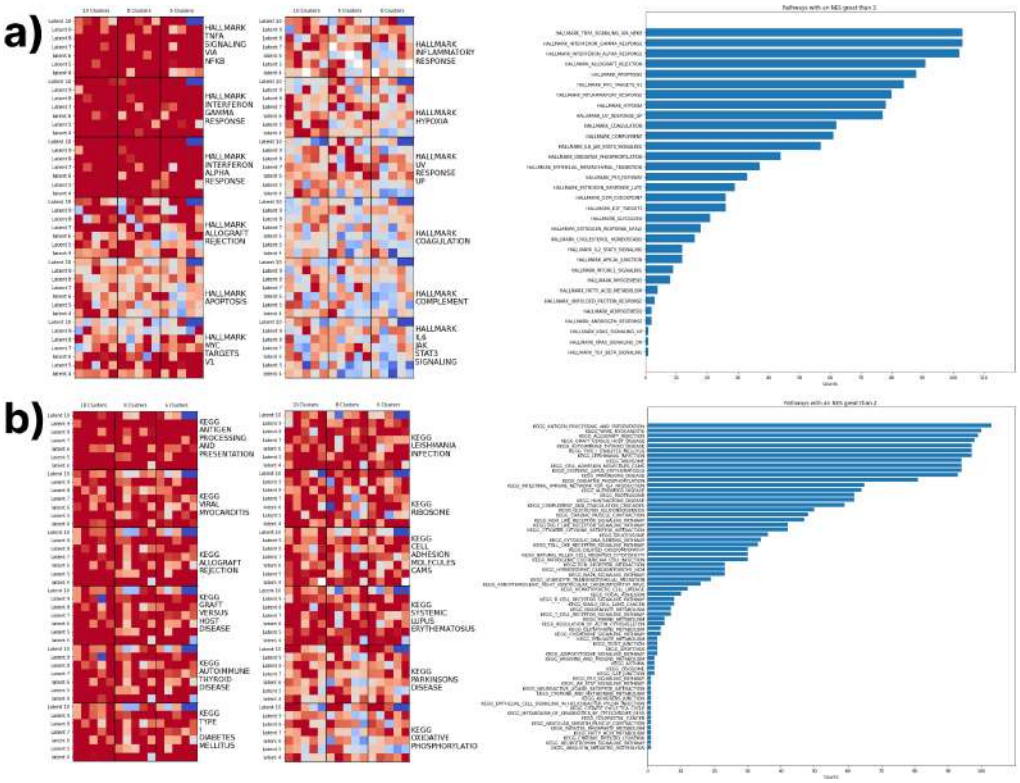


Figure 3.3: **a)** The heatmap (left) of the 12 Hallmark pathways with most count, a square represent a run with a certain random seed, in the corresponding latent space and cluster (row, column) and the intensity is the NES score for the corresponding parameters. Bar plot of the count of $NES > 2$ for each pathways among all runs. **b)** Similar plots, for the KEGG database. Higher resolution images can be found in the appendix (figures A.5 A.6)

on the 10 clusters of this run can be found in the appendix (tables A.1, A.2). The cluster with pathways that scored a NES higher than 2 are $C2, C5, C6, C7, C8, C9$. The fact that both KEGG and Hallmark pathway databases agree confirms the validity of our observation. The most significant pathways for each cluster are consistent between the two databases, in the following we describe the clusters according to their associated pathways.

We find $C2$ to be characterized by enrichment of the glycolysis pathway, which describes the energy-yielding process of converting glucose into pyruvic acid. The enrichment of *KEGG Ribosome* and *Hallmark Myc Target* suggests a proliferation of the cells in this cluster since myc acts as a transcription factor thus stimulating the activity of ribosomes. This cluster can therefore be interpreted as cells in a proliferative state.

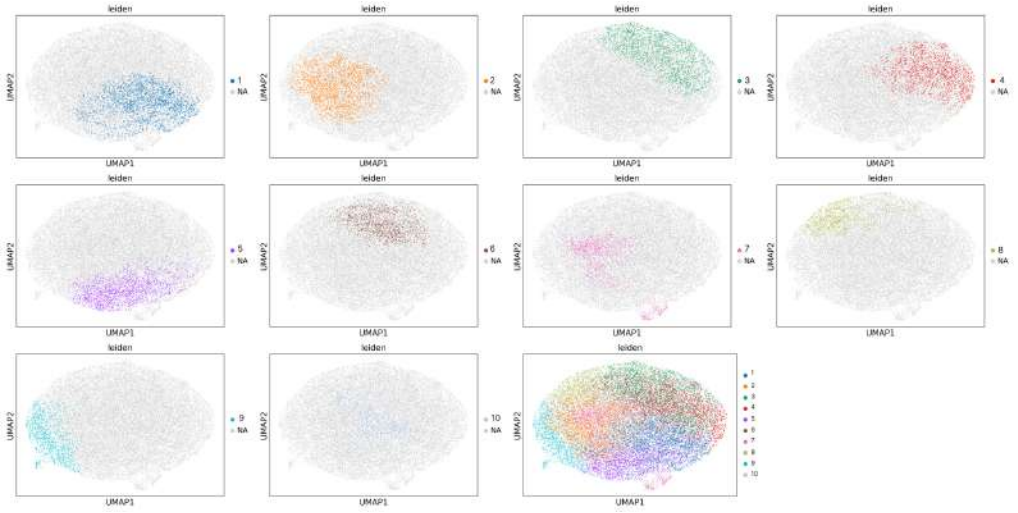


Figure 3.4: UMAP embedding of the 10 cluster found in the 4 latent dimensions representation.

The cluster $C5, C7$, and $C8$ share antigen and allograft rejection KEGG pathways, and interferon-alpha and beta Hallmark pathways. They highlight immune response associated activity, and we note that the presence of interferon can be linked with apoptosis pathways (also enriched in these clusters). Indeed interferon act on P53 proteins, which will normally trigger apoptosis. But it is interesting to note that TP53 (the gene coding for P53) is the most commonly mutated gene in Ovarian Cancer [33], the interactions of cells in these clusters could thus be linked to abnormal apoptosis behavior.

$C6$ shows enrichment in myc, ribosome, and oxidative phosphorylation pathways. This suggests intense cell respiration, indicating stress or cycling cells. Finally the cluster $C9$ shows a mix of different pathways, including TNF-alpha via NF-kB, infection, myc, ribosome, and fatty-acid metabolism. While it is not clear how these are related, the TNF-alpha and infection pathways could indicate an inflammatory response [34].

We provide plots of the scored pathways in the same latent space as we presented throughout this section to allow visual exploration of the organization of the latent space for the pathways.

3.3 Comparison and interpretation

We introduced reference signatures in section 2.2.3, where we mentioned their origin. Here we want to compare the score of these signatures, and visually associate them to the clusters that were found in postprocessing. This serves as

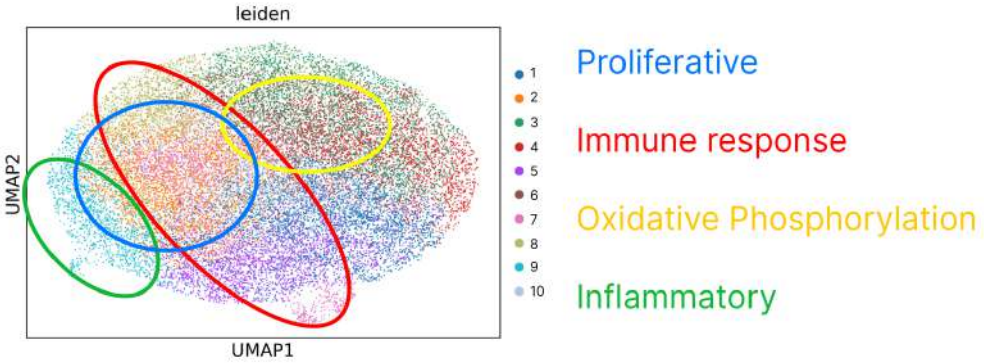


Figure 3.5: UMAP embedding of the dataset, with our characterization summarized graphically. Cluster $C2$ is named "Proliferative", clusters $C5, C7, C8$ are named "Immune response", cluster $C6$ is named "Oxidative Phosphorylation" and cluster $C9$ is named Inflammatory.

a basis for validating the known signatures and allows us to identify if the clusters we found suggest new ones.

3.3.1 Unsupervised Bulk RNA Signatures

In her unpublished work, Agnieszka Kraft describes 3 distinct signatures in Ovarian Cancer, named after the enriched pathways representing them. These are "Mitotic Spindle", "Oxidative Phosphorylation Proliferation" and "Proliferation". We first observe their organization in the latent space, which shows that mitotic spindle and oxidative phosphorylation are complementary. We observe a very light gradient for proliferation, which score is higher the further we move away from the mitotic spindle program.

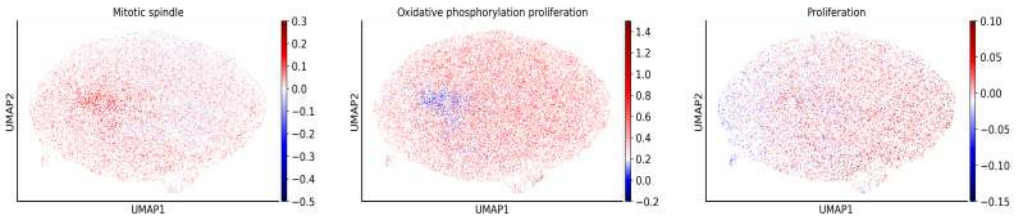


Figure 3.6: Score for the signatures found in TCGA Ovarian, using Agniezska Kraft's method.

We can only match visually and with good confidence the signatures for mitotic spindle and proliferation to the cluster that were found with our method. In particular, mitotic spindle shows the most overlap with cluster $C7$ which we found

to be enriched in immune response-related pathways. This does not seem to correspond and it is, therefore, unclear which characterization we should trust. The proliferative state shows overlap with clusters $C1, C3, C4$. However, we found that the NES for these clusters is relatively low suggesting the organization of cells is not representing significant biological information (at least in the KEGG and Hallmark databases). Finally, the score of oxidative phosphorylation proliferation does not give much more information, as it shows a high score everywhere except in $C7$.

3.3.2 Hao Study Signatures

Researchers analyzed the transcriptomic profiles of 13 571 cells from four paired primary and metastatic High-Grade Serous Ovarian Cancer (HGSOC) samples [25]. They identify a group of aggressive epithelial (cancer) cells, among which they identified five distinct clusters, functionally annotated using Gene Set Variation Analysis and the KEGG database after computing DGE:

- $E1$: is enriched in glycolysis, citrate cycle, extracellular matrix interaction, and focal adhesion
- $E2$: the differentially expressed genes in EC2 are involved in cytokine-cytokine receptor interaction and the neuroactive-related pathways
- $E3$: exhibited higher expression of genes associated with nucleotide and amino acid metabolism, as well as Fanconi Anemia and ABC transporter pathways
- $E4$: was characterized by the immune response-related pathways, coagulation cascades, and antigen processing and presentation
- $E5$: finally shows enrichment for pathways associated with cell cycle, DNA replication, DNA repair, and drug metabolism

Based on the description of the study we recognize some similarities in the pathways that are expressed. For example, one can associate the description of $E1$ with $C2$ based on the enrichment in glycolysis pathway. The neuroactive pathways pointed out in $E2$ could have a link with the neurological disease found in $C6$. $E3$ nucleotide and amino acid metabolism, and the ABC transporter pathways can suggest a proliferative, energy-yielding state of the cells. A cluster showing similar characteristics could be given by $C2$. $E4$ is enriched in immune-response related pathways and shares a significant number of traits with the group $C5, C7, C8$ found in our experiment. $E5$ does not correspond clearly to any cluster we described previously.

While some pathways are shared between clusters of the study and our experiment, we cannot establish clear connections when we visually compare the scores

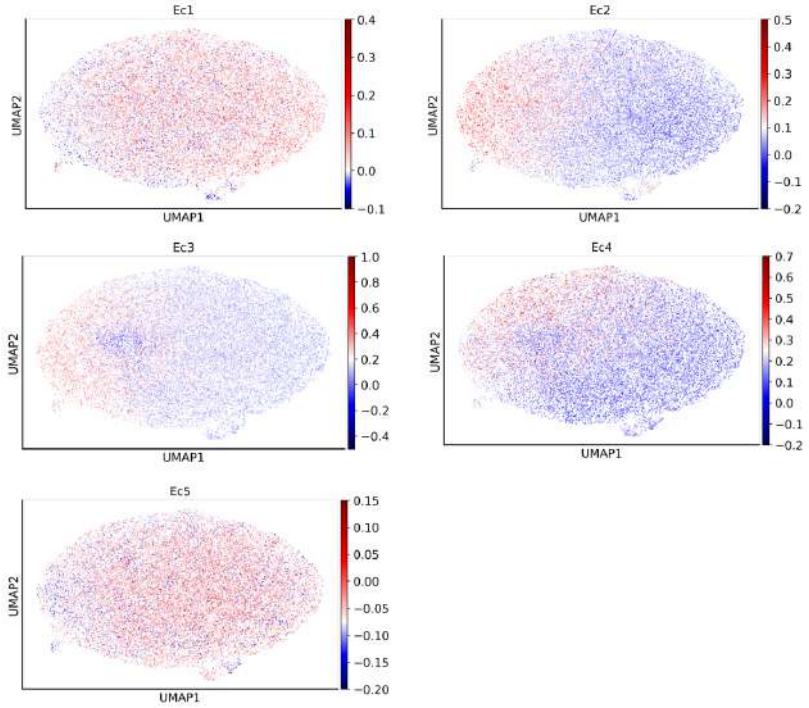


Figure 3.7: Score for the signatures found by the Hao study.

of the signatures to the cluster in the latent space embedding. Example of such mismatch include the association of $E1$ to $C2$, $E2$ to $C6$, $E3$ and $C2$. The only signature that shows interesting overlap is the $E4$, which we see is particularly overlapping $C7$ (it is also sharing the most pathways with $C7$ than other clusters). We note that the score of $E5$ is relatively homogeneous, which makes it difficult to associate it with any cluster of our experiment.

We thus can only confirm similarities between the results of the Hao study and our experiment in the immune-related cell clusters. A lack of evidence prevents us from making any further assumptions based on these observations only. We might however suggest reasons as to why it is not possible to have obvious validation. These include the lack of batch effect correction, different techniques of clustering, different gene set enrichment method, and finally a relatively low number of cells (there is indeed a ten-fold difference in the number of cells present in the datasets).

Chapter 5

Conclusion

During this project, we performed a complete scRNA-seq analysis on a new private dataset of Ovarian Cancer samples. We used advanced methods and tools, integrated into a pipeline that was built by members of the BoevaLab. This report provides an additional case study of the pipeline, showing the potential of the pipeline for simplifying and accelerating the discovery of *de novo* shared transcriptional signatures.

From the raw counts found on the cluster LeoMed, we assembled a dataset containing more than 100'000 cells meeting our quality standards. We found that their embedding in latent spaces is relatively stable in terms of dimension (i.e., we observe the same organization between low dimension embedding and high dimension embedding). We identified contaminated samples and rejected them, keeping the purest samples for our analysis. We made sure there were no unexpected clusters of cells by visually inspecting the UMAP representation colored with sample ID, count-depth, location of the sample.

We validated the signature $E4$ from the Hao paper. Indeed, we identified 3 clusters ($C5$, $C7$, and $C8$) enriched in the same pathways, and visually overlap with $E4$ in the UMAP representation of the latent space. We identified 3 other distinct signatures, namely the Proliferative, Oxidative Phosphorylation, and Inflammatory signatures. They share some pathways with our references (e.g., $C2$ and $E1$ are enriched in glycolysis pathways) but we could establish clear correspondence due to the bad overlap of their visual representation in UMAP.

Finally, we provide a foundation for future work on Ovarian Cancer, including a list of usable samples and their corresponding information, a preprocessed dataset, latent representations, and gene set enrichment analysis results for each cluster of each run of each latent dimension we analyzed. This report also provides a guide for future students, with a clear and concise description of each step of a scRNA-seq experiment with the pipeline, as well as its design choices.

Further Work

We note that, while we were able to link existing immune response signatures to the one we discovered, we lack validation. Thus there is a need for further comparison points, as well as a deeper biological interpretation of the results. While GSEA remains a solid basis that has been used for many years for characterizing lists of deferentially expressed genes, it is an old method and alternatives may provide at least a basis for comparison.

Acknowledgements

I hereby want to thank Prof. Valentina Boeva, as well as core members of the "single-cell team" Josephine Yates, Florian Barkmann, and Paweł Czyż, for their supervision during my project, and for their insights on the development of software for computational biology in an academic setting.

Bibliography

[1] S. Chakraborty and T. Rahman, "The difficulties in cancer treatment," *ecancermedicalscience*, vol. 6, 2012.

[2] "Advancing cancer genomics," *Nature Genetics*, vol. 51, no. 5, pp. 767–767, May 2019. [Online]. Available: https://doi.org/10.1038/s41588-019-0419-6

[3] N. C. Institute. (2022) What is cancer. [Online]. Available: bit.ly/3ptHlzb

[4] A. C. Society. (2022) What Is Ovarian Cancer? [Online]. Available: https://bit.ly/what-is-oc

[5] L. A. Torre, B. Trabert, C. E. DeSantis, K. D. Miller, G. Samimi, C. D. Runowicz, M. M. Gaudet, A. Jemal, and R. L. Siegel, "Ovarian cancer statistics, 2018," *CA: a cancer journal for clinicians*, vol. 68, no. 4, pp. 284–296, 2018.

[6] M.-A. Lisio, L. Fu, A. Goyeneche, Z.-h. Gao, and C. Telleria, "High-grade serous ovarian cancer: basic sciences, clinical and therapeutic standpoints," *International journal of molecular sciences*, vol. 20, no. 4, p. 952, 2019.

[7] W. Kassuhn, O. Klein, S. Darb-Esfahani, H. Lammert, S. Handzik, E. T. Taube, W. D. Schmitt, C. Keunecke, D. Horst, F. Dreher *et al.*, "Classification of molecular subtypes of high-grade serous ovarian cancer by maldi-imaging," *Cancers*, vol. 13, no. 7, p. 1512, 2021.

[8] R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro *et al.*, "Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome," *Clinical cancer research*, vol. 14, no. 16, pp. 5198–5208, 2008.

[9] M. Kossaï, A. Leary, J.-Y. Scoazec, and C. Genestie, "Ovarian cancer: a heterogeneous disease," *Pathobiology*, vol. 85, no. 1-2, pp. 41–49, 2018.

[10] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, "A practical guide to single-cell rna-sequencing for biomedical research and clinical applications," *Genome medicine*, vol. 9, no. 1, pp. 1–12, 2017.

[11] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz *et al.*, "Eleven grand challenges in single-cell data science," *Genome biology*, vol. 21, no. 1, pp. 1–35, 2020.

[12] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen, "A benchmark of batch-effect correction methods for single-cell rna sequencing data," *Genome biology*, vol. 21, no. 1, pp. 1–32, 2020.

[13] L. Dony, M. König, D. Fischer, and F. J. Theis, "Variational autoencoders with flexible priors enable robust distribution learning on single-cell rna sequencing data," in *ICML 2020 Workshop on Computational Biology (WCB) Proceedings Paper*, vol. 37, 2020.

[14] M. Gletting, "Developing methodology for the detection of shared transcriptional signatures from single cell rna-seq data," in *Master Thesis, EPFL, ETH Zurich Boevalab*, 2021.

[15] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, no. 1, pp. 1–5, 2018.

[16] Broadinstitute. (2019) infercnv of the trinity ctat project. [Online]. Available: https://github.com/broadinstitute/inferCNV

[17] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature methods*, vol. 15, no. 12, pp. 1053–1058, 2018.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[19] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[20] A. Irmisch, X. Bonilla, S. Chevrier, K.-V. Lehmann, F. Singer, N. C. Toussaint, C. Esposito, J. Mena, E. S. Milani, R. Casanova *et al.*, "The tumor profiler study: integrated, multi-omic, functional tumor profiling for clinical decision support," *Cancer Cell*, vol. 39, no. 3, pp. 288–293, 2021.

[21] M. D. Luecken and F. J. Theis, "Current best practices in single-cell rna-seq analysis: a tutorial," *Molecular systems biology*, vol. 15, no. 6, p. e8746, 2019.

[22] X. Shao, N. Lv, J. Liao, J. Long, R. Xue, N. Ai, D. Xu, and X. Fan, "Copy number variation is highly correlated with differential gene expression: a pan-cancer study," *BMC medical genetics*, vol. 20, no. 1, pp. 1–14, 2019.

[23] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza *et al.*, "Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, 2014.

[24] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature biotechnology*, vol. 33, no. 5, pp. 495–502, 2015.

[25] Q. Hao, J. Li, Q. Zhang, F. Xu, B. Xie, H. Lu, X. Wu, and X. Zhou, "Single-cell transcriptomes reveal heterogeneity of high-grade serous ovarian carcinoma," *Clinical and Translational Medicine*, vol. 11, no. 8, p. e500, 2021.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[27] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf

[28] V. A. Traag, L. Waltman, and N. J. Van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[29] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[30] L. Yu, Y. Cao, J. Y. Yang, and P. Yang, "Benchmarking clustering algorithms on estimating the number of cell types from single-cell rna-sequencing data," *Genome biology*, vol. 23, no. 1, pp. 1–21, 2022.

[31] X. Li, H. Cai, X. Wang, L. Ao, Y. Guo, J. He, Y. Gu, L. Qi, Q. Guan, X. Lin *et al.*, "A rank-based algorithm of differential expression analysis for small cell line data with statistical control," *Briefings in bioinformatics*, vol. 20, no. 2, pp. 482–491, 2019.

[32] U. S. Diego and B. Institute. (2022) MSigDB. [Online]. Available: https://www.gsea-msigdb.org/gsea/msigdb/

[33] R. J. Yamulla, S. Nalubola, A. Flesken-Nikitin, A. Y. Nikitin, and J. C. Schimenti, "Most commonly mutated genes in high-grade serous ovarian carcinoma are nonessential for ovarian surface epithelial stem cell transformation," *Cell reports*, vol. 32, no. 9, p. 108086, 2020.

[34] G. M. Chen, L. Kannan, L. Geistlinger, V. Kofia, Z. Safikhani, D. M. Gendoo, G. Parmigiani, M. Birrer, B. Haibe-Kains, and L. Waldron, "Consensus on molecular subtypes of high-grade serous ovarian carcinoma," *Clinical Cancer Research*, vol. 24, no. 20, pp. 5037–5047, 2018.

Appendix A

Appendix

The following is the README for the archive folder on the shared drive of BoevaLab.

Running the experiment

Paths

In the scripts, the path will save the result of preprocessing on LeoMed in `/cluster/dataset/boeva/scRNAdata/preprocessed/ovarian`, the latent `/cluster/dataset/boeva/antoinco/modeling`, and the result of GSEA in `/cluster/dataset/boeva/antoinco/postprocessing`. The corresponding variables must be edited in the corresponding files.

The paths for the metadata used in the preprocessing correspond to paths in LeoMed. We provide the files in this directory in the folder "metadata".

Run

To recreate the scRNA-seq analysis, run the following in this order:

- First, install all dependencies needed for the pipeline
- Run the bash script "preprocessing.sh" in `/scripts`
- Run the python script "OC_get_location.py" in `/script`
- Run the python script "OC_run_latent_experiment.py" in `/scripts`
- Run the python script "OC_run_stability_analysis.py" in `/scripts`

Note that it is important to change the paths, depending on the execution environment. As is, the script work from in LeoMed, after having added the pipeline

to the system path: `export PYTHONPATH=/path/to/scRNA_shared_signatures/src`. To run on new preprocessing, the paths inside "OC_run_latent_experiment.py" and "OC_run_stability_analysis.py" will have to be edited: the path to "unhealthy.h5ad" in "OC_run_latent_experiment.py" must be changed to the latest run and the time-stamp in "OC_run_stability_analysis.py" to identify the run must be replaced.

Content of folders

metadata

- metadata.csv : contains the sample_id, data paths and annotation paths for each 48 samples included at the beginning of the dataset (note that we use the list "excluded_samples" in the preprocessing config file `scRNA_shared_signatures/src/preprocessing/conf/cancer/ovarian.yaml` to exclude relevant samples from the experiment).
- all_info.csv : contains the aggregated information about all samples, generated by the notebook `notebooks/OC_generate_metadata.ipynb`

notebooks

To build intuition on the different frameworks, libraries, and the pipeline itself, Jupyter Notebooks were used. Some relevant ones are in this folder (starting with "OC"). We used them mainly to generate visualizations and explore the results.

archive

In this subfolder are all the Jupyter notebooks that were not used directly to generate or interpret results, but they were useful to learn about the libraries and frameworks, and experiment in general with scRNA-seq data to get familiar with it.

results

In the subfolder "modeling" are all the latent representations for each run of each latent space that was modeled using SCVI.

In the subfolder, "postprocessing" are all the results of GSEA for each run of each latent space and each number of clusters, for KEGG and Hallmark databases

scRNA_share_signatures

This is the version of the pipeline that was used to perform the experiment. All modules are contained in `src`

signatures

- agnieska_signatures: contains the 3 signatures that Agnieszka discovered in her work
- Hao_signatures: for each cluster identified in the Hao study, we kept the top 150 differentially expressed genes in the clusters
- misc: contains signatures we investigated but didn't end up using
- olbrecht_signatures: contains signatures we investigated but didn't end up using
- tothill_signatures: contains signatures we investigated but didn't end up using
- cc_genes_2.csv : is a cell cycle annotation file, used to score the cell-cyle

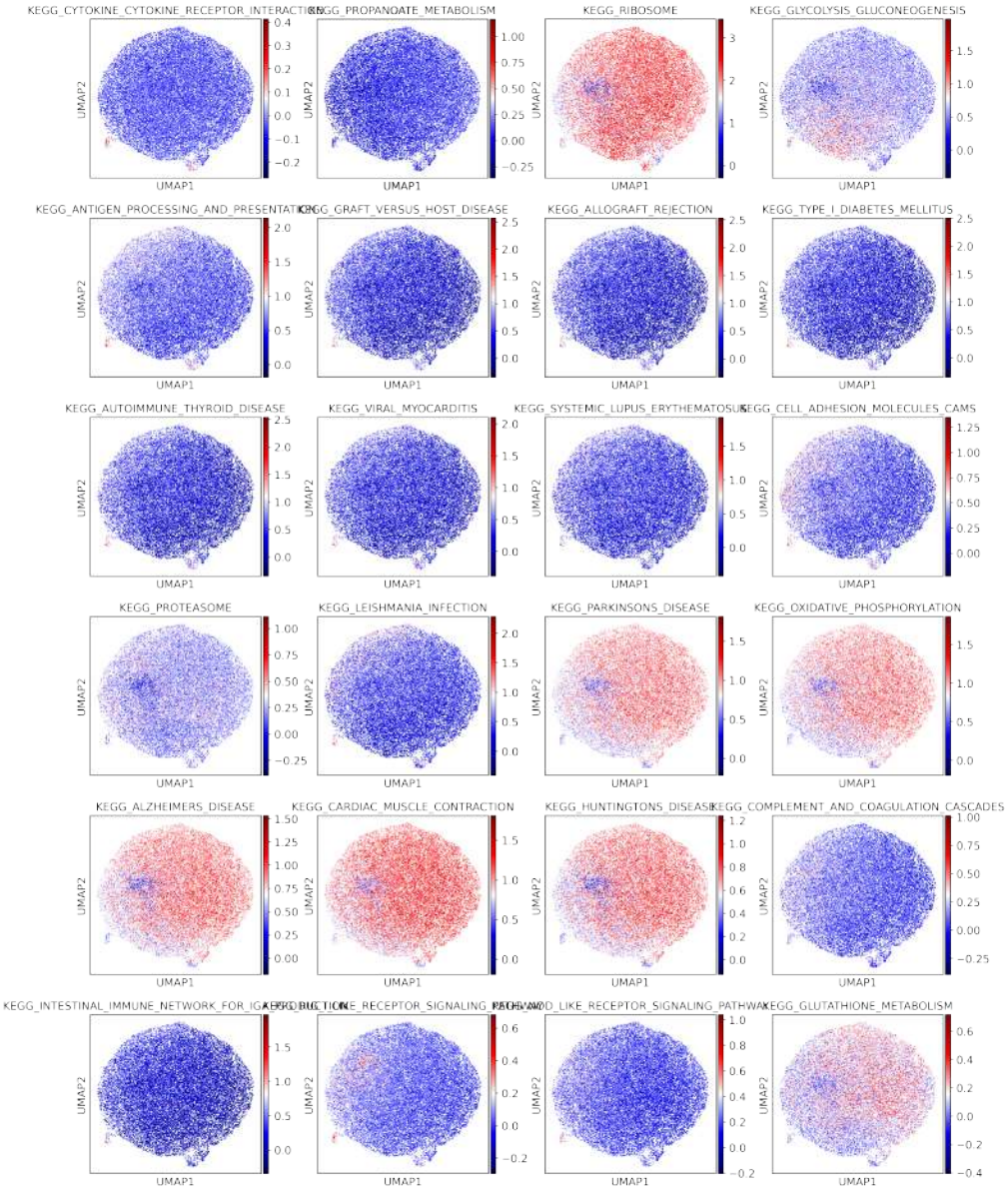


Figure A.1: KEGG pathways found as top pathway in the run inspected in section 3.2., scored and plotted in the corresponding latent space embedding.

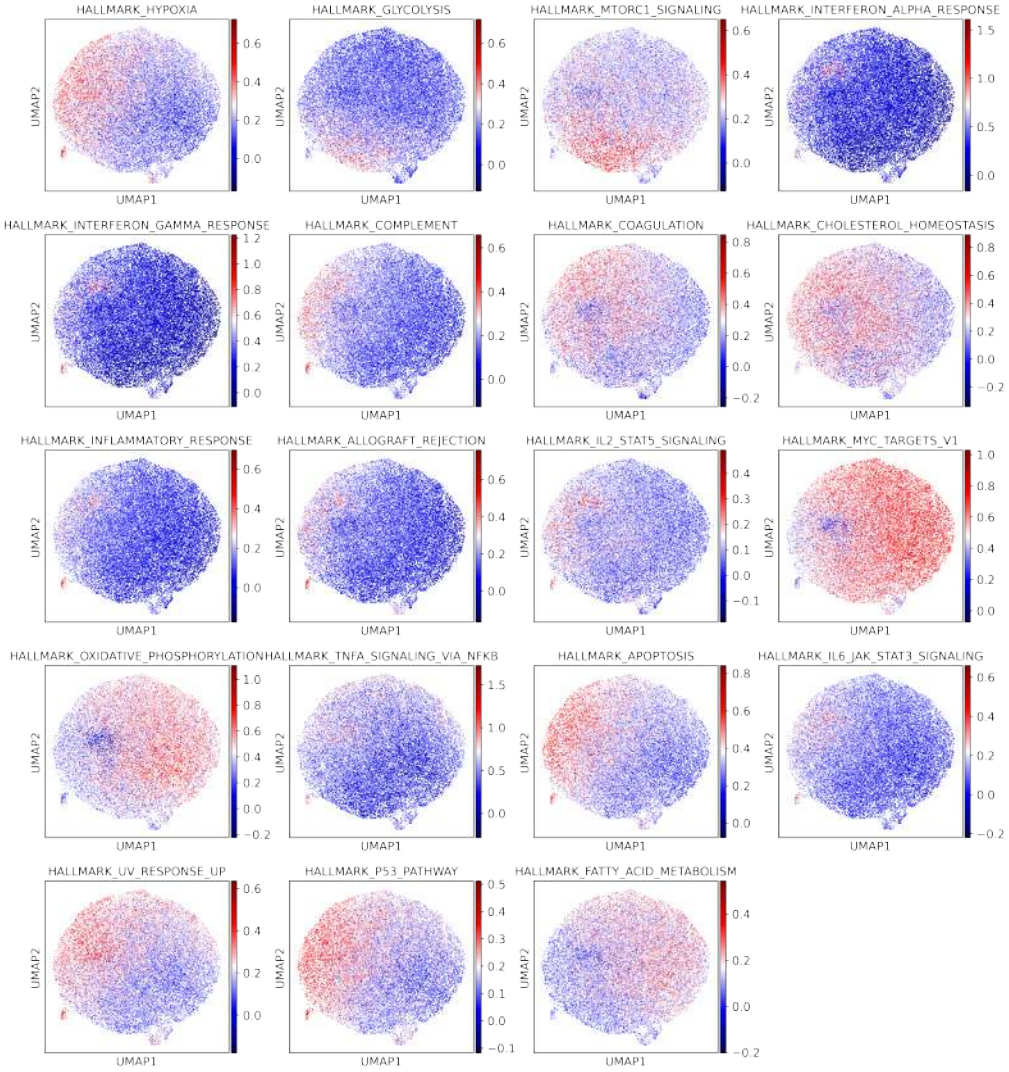


Figure A.2: Hallmark pathways found as top pathway in the run inspected in section 3.2., scored and plotted in the corresponding latent space embedding.

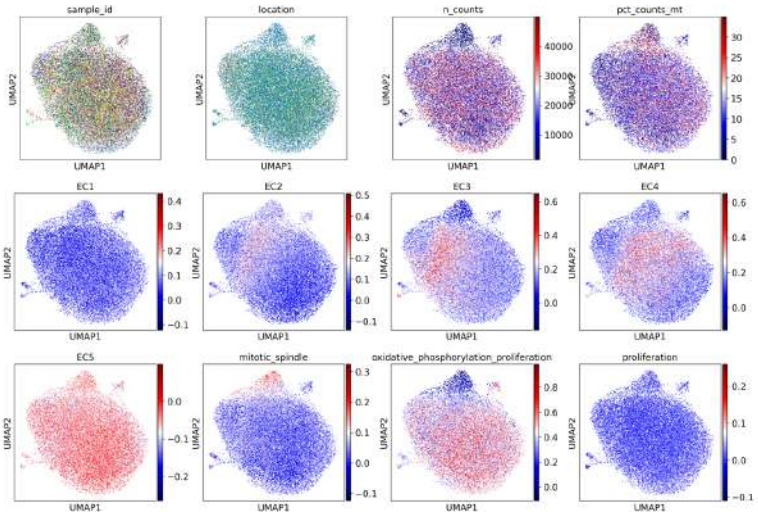


Figure A.3: UMAP of dataset embeding of the runs "*20220320-095832-usable-slug*" in a 6 dimensional latent space. Note that visually in UMAP, we can only draw similar conclusion on the organization of the space than for the 4 dimensional latent spaces.

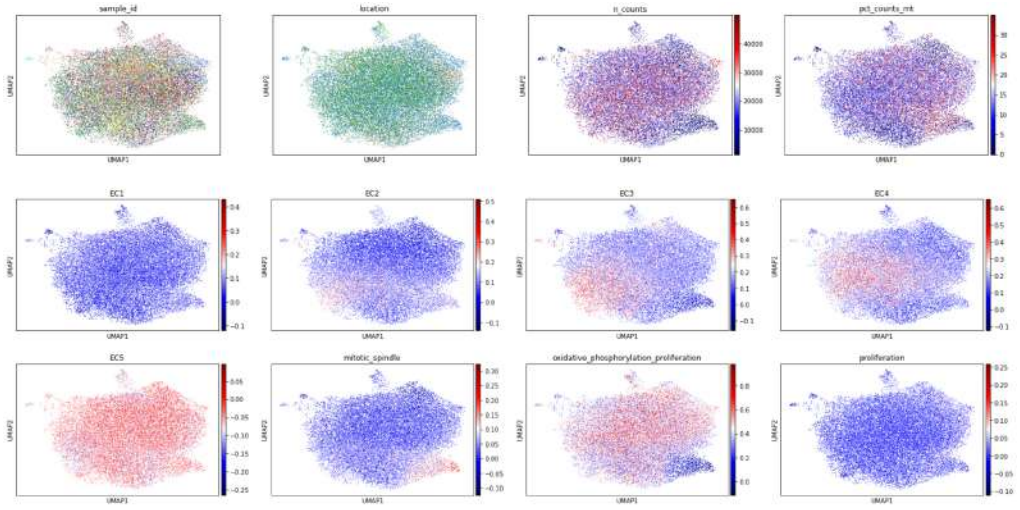


Figure A.4: UMAP of dataset embeding of the runs "*20220320-095822-superb-quail*" in a 10 dimensional latent space.

Pathway	es	nes	pval	fdr	Cluster
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	-0.1513	inf		0	
KEGG_RIBOSOME	0.9152	1.4332	0	0	
KEGG_ALZHEIMERS_DISEASE	0.8585	1.3445	0	0.0191	1
KEGG_OXIDATIVE_PHOSPHORYLATION	0.8492	1.3401	0	0.0198	
KEGG_PARKINSONS_DISEASE	0.8491	1.3292	0	0.0218	
KEGG_RIBOSOME	0.7565	2.5161	0	0	
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	0.7712	2.1466	0	0	
KEGG_P53_SIGNALING_PATHWAY	0.5253	1.5130	0.0327	0.4939	2
KEGG_VIBRIO_CHOLERAE_INFECTION	0.5014	1.4892	0.0224	0.4711	
KEGG_SPLICEOSOME	0.4471	1.4812	0.0020	0.4037	
KEGG_RIBOSOME	0.9105	1.5089	0	0	
KEGG_HUNTINGTONS_DISEASE	0.8241	1.3878	0	0.0110	
KEGG_PARKINSONS_DISEASE	0.8365	1.3874	0	0.0073	3
KEGG_ALZHEIMERS_DISEASE	0.8263	1.3757	0	0.0110	
KEGG_OXIDATIVE_PHOSPHORYLATION	0.8282	1.3733	0	0.0100	
KEGG_VIRAL_MYOCARDITIS	0.7465	1.5993	0	0.0180	
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	0.7574	1.5992	0	0.0090	
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	0.7524	1.5969	0	0.0067	4
KEGG_AUTOIMMUNE_THYROID_DISEASE	0.7653	1.5729	0	0.0085	
KEGG_ALLOGRAFT_REJECTION	0.7474	1.5519	0	0.0136	
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	0.7527	2.9695	0	0	
KEGG_GRAFT_VERSUS_HOST_DISEASE	0.8174	2.6549	0	0	
KEGG_ALLOGRAFT_REJECTION	0.7880	2.6313	0	0	5
KEGG_TYPE_I_DIABETES_MELLITUS	0.7921	2.6005	0	0	
KEGG_AUTOIMMUNE_THYROID_DISEASE	0.7880	2.5979	0	0	
KEGG_RIBOSOME	0.8061	2.8990	0	0	
KEGG_PARKINSONS_DISEASE	0.6669	2.4410	0	0	
KEGG_OXIDATIVE_PHOSPHORYLATION	0.6646	2.4140	0	0	6
KEGG_ALZHEIMERS_DISEASE	0.6373	2.2986	0	0	
KEGG_CARDIAC_MUSCLE_CONTRACTION	0.6722	2.1860	0	0	
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	0.6531	2.0179	0	0	
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	0.7138	1.9255	0	0.0020	
KEGG_VIRAL_MYOCARDITIS	0.6217	1.8856	0	0.0020	7
KEGG_GRAFT_VERSUS_HOST_DISEASE	0.6903	1.8770	0	0.0020	
a KEGG_AUTOIMMUNE_THYROID_DISEASE	0.6944	1.8641	0	0.0016	
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	0.8172	2.5690	0	0	
KEGG_ALLOGRAFT_REJECTION	0.8846	2.5479	0	0	
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	0.8635	2.4843	0	0	8
KEGG_AUTOIMMUNE_THYROID_DISEASE	0.8812	2.4681	0	0	
KEGG_GRAFT_VERSUS_HOST_DISEASE	0.8662	2.4626	0	0	
KEGG_RIBOSOME	0.6232	2.7573	0	0	
KEGG_LEISHMANIA_INFECTION	0.6843	2.2849	0	0	
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	0.5467	2.2107	0	0.0007	9
KEGG_VIRAL_MYOCARDITIS	0.5988	2.0905	0	0.0026	
KEGG_GLUTATHIONE_METABOLISM	0.5784	2.0709	0	0.0029	
KEGG_SPLICEOSOME	0.8197	1.3213	0.0040	0.7562	
KEGG_MTOR_SIGNALING_PATHWAY	0.8026	1.2726	0.0630	0.9738	
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	0.7969	1.2717	0.0544	0.6612	10
KEGG_ACUTE_MYELOID_LEUKEMIA	0.8057	1.2704	0.0643	0.5089	
KEGG_CHRONIC_MYELOID_LEUKEMIA	0.7613	1.2220	0.1109	0.8865	

Table A.1: List of the top 5 KEGG pathways for each cluster of the run inspected in section 3.

Pathway	es	nes	pval	fdr	Cluster
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	0.8226	1.2638	0.0141	0.1201	
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.7849	1.2460	0	0.1131	
HALLMARK_HYPOXIA	0.7905	1.2216	0.0100	0.1347	1
HALLMARK_ALLOGRAFT_REJECTION	0.7793	1.2181	0.0020	0.1131	
HALLMARK_MYC_TARGETS_V1	0.7752	1.2174	0	0.0929	
HALLMARK_HYPOXIA	0.6430	2.2165	0	0	
HALLMARK_GLYCOLYSIS	0.6234	2.0891	0	0	
HALLMARK_MTORC1_SIGNALING	0.5740	1.9329	0	0	2
HALLMARK_MYC_TARGETS_V1	0.5371	1.8287	0	0.0020	
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.5330	1.7820	0	0.0032	
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.8002	1.3485	0	0.0080	
HALLMARK_MYC_TARGETS_V1	0.7669	1.2788	0	0.0510	
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	0.8040	1.2774	0.0160	0.0367	3
HALLMARK_ALLOGRAFT_REJECTION	0.7710	1.2608	0	0.0400	
HALLMARK_ADIPOGENESIS	0.7020	1.1696	0.0040	0.3196	
HALLMARK_HYPOXIA	0.6613	1.4890	0	0.0160	
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	0.7078	1.4826	0	0.0080	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.7033	1.4821	0	0.0053	4
HALLMARK_APOPTOSIS	0.6490	1.4682	0	0.0065	
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.6359	1.4325	0	0.0104	
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.8106	3.5562	0	0	
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.7447	3.3609	0	0	
HALLMARK_COMPLEMENT	0.5178	2.2269	0	0	5
HALLMARK_COAGULATION	0.5589	2.2150	0	0	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.5992	2.0923	0	0.0004	
HALLMARK_MYC_TARGETS_V1	0.5733	2.1820	0	0	
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.5705	2.1396	0	0	
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.5883	2.0422	0	0	6
HALLMARK_ALLOGRAFT_REJECTION	0.6332	1.9660	0	0	
HALLMARK_UV_RESPONSE_UP	0.5001	1.7411	0.0020	0.0138	
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.6642	2.1872	0	0	
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.6259	2.1475	0	0	
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.5688	1.9510	0	0	7
HALLMARK_APOPTOSIS	0.5581	1.8204	0	0.0010	
HALLMARK_ESTROGEN_RESPONSE_LATE	0.5321	1.7797	0	0.0008	
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.7703	2.6248	0	0	
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.8092	2.6165	0	0	
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.7477	2.5332	0	0	8
HALLMARK_APOPTOSIS	0.6913	2.2848	0	0	
HALLMARK_ALLOGRAFT_REJECTION	0.7147	2.2509	0	0	
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.5568	2.4037	0	0	
HALLMARK_MYC_TARGETS_V1	0.4701	2.2550	0	0.0010	
HALLMARK_FATTY_ACID_METABOLISM	0.4433	1.9552	0	0.0061	9
HALLMARK_ALLOGRAFT_REJECTION	0.4389	1.7932	0	0.0224	
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.3745	1.7880	0	0.0188	
HALLMARK_MYC_TARGETS_V1	0.8978	1.4342	0.0020	0.0080	
HALLMARK_UV_RESPONSE_UP	0.8124	1.3177	0.0080	0.1031	
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.8255	1.3151	0.0140	0.0761	10
HALLMARK_APOPTOSIS	0.8021	1.3071	0.0040	0.0656	
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.7654	1.2719	0	0.1157	

Table A.2: List of the top 5 Hallmark pathways for each cluster of the run inspected in section 3.

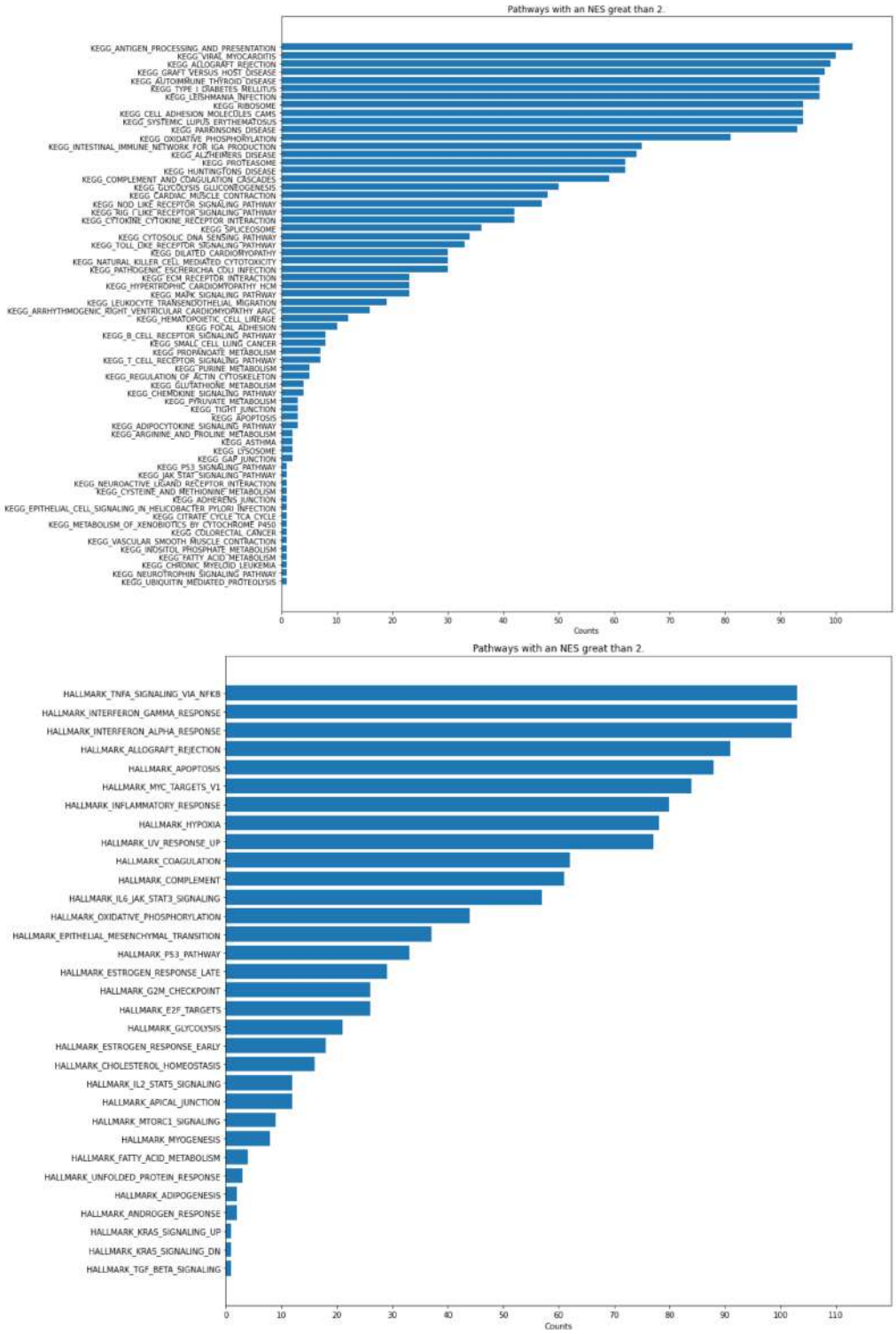


Figure A.5: Number of occurrences of NES higher than 2, for each pathway and over all runs, for KEGG and Hallmark databases.

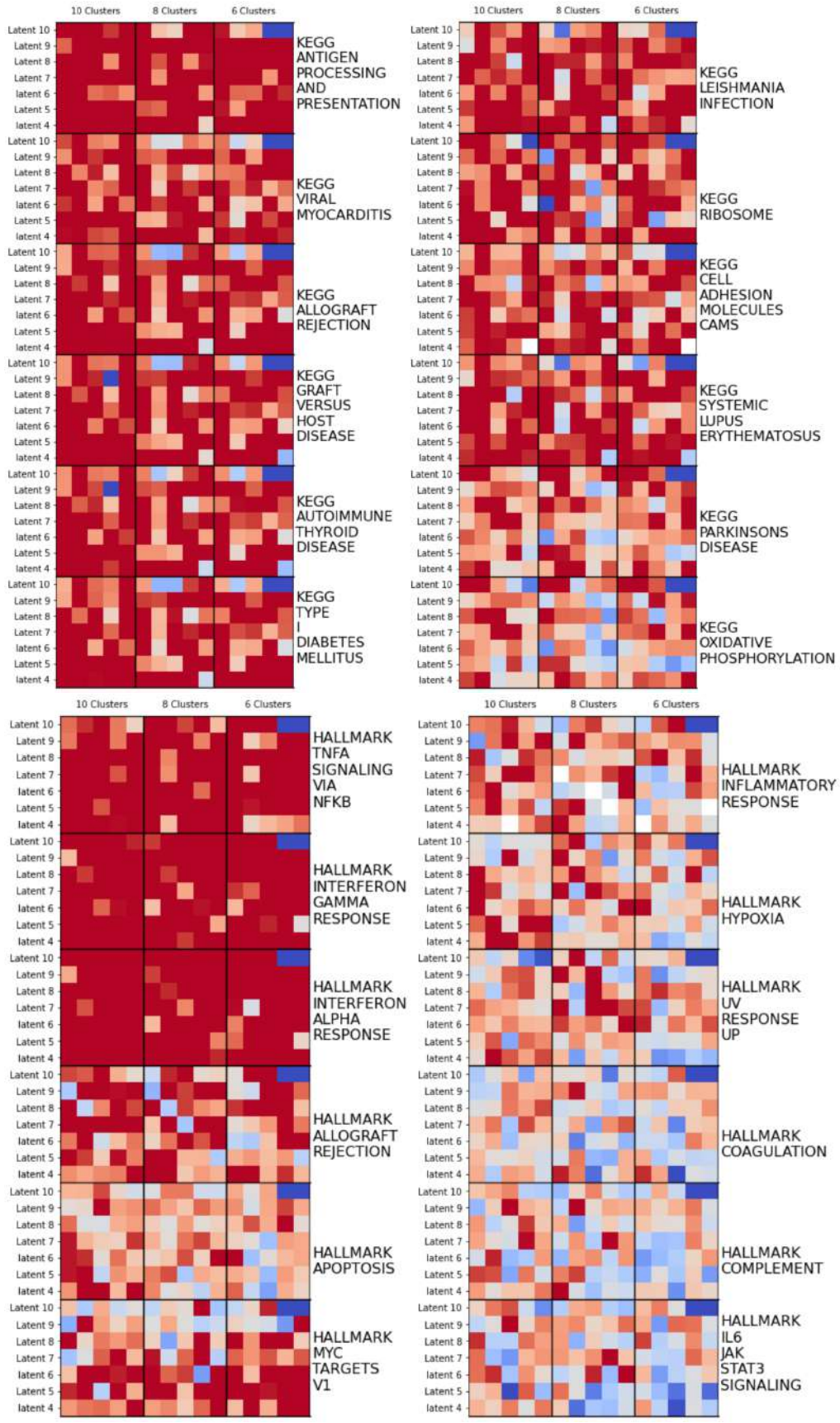


Figure A.6: Heatmap with the top 12 pathways (in term of NES) over all runs, for KEGG and Hallmark databases.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

DIGGING DEEP INTO INTRATUMORAL HETEROGENEITY USING SCRNA-SEQ DATA

Authored by (in block letters):

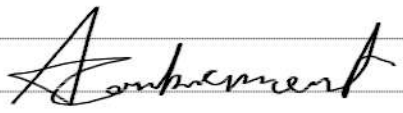
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
COMBREMONT	ANTOINE

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
EDINBURGH 30.03.2022	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.