ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Supervised and Unsupervised Discovery of Intratumor Heterogeneity from Single Cell RNA-seq Data

Bachelor Thesis

Philip Toma

October 12, 2022

Advisors: Prof. Dr. Valentina Boeva, Agnieszka Kraft
Department of Computer Science, ETH Zürich

Abstract

Research in the field of biomedicine and bioinformatics has shown a significant degree of intratumor heterogeneity across cancer types. The degree of differentiation within a tumor severely impacts treatment success of cancer patients, due to drug-persistent cancer cells present within tumors. It follows, that progress in the effective treatment of cancers depends on the development of feasible methods for the analysis of intratumor heterogeneity.

In this work, we establish a pipeline for the study of intratumor transcriptional heterogeneity and its association with copy number aberration based subclonal structures in malignant skin melanoma.

First, we explored the utility of existing semi-supervised cell type annotation programs for the annotation of melanoma transcriptional states. We generated transcriptional state labels for skin melanoma data from the Tumor Profiler Study and evaluated annotations of both models.

Furthermore, we generated labels to quantify genetic differences within tumors. To achieve this, we implemented a method to automatically infer subclonal structures based on predicted copy number aberrations. Finally, we analysed the association of transcriptional state and subclonal structures, extracting a list of genes for future study of cancer cell differentiation in melanoma.

Contents

Chapter 1

Introduction

If asked in the year 2020 which public health problem is the most prevalent in developed countries, it is likely that a common answer would be Covid-19. While the global pandemic plays a major role in the eye of the public, cancer continues to be a major cause of premature death which has continued to affect millions of lives across the globe. According to the Swiss Federal Office for Statistics, one in five persons will develop a cancer before the age of 70.

Cutaneous malignant melanoma, of which we will in the following simply refer to as melanoma, is a form of skin cancer that develops in melanocytes. It is known for its tendency to metastasise and thereby spread to other tissues in the body.

In a 2020 study, the American Cancer Society estimated melanoma to contribute by around 6% to the number of new cancer cases, and by nearly 2% to the predicted number cancer-caused deaths in the US [1]. While the relatively low mortality rate seems comforting, research efforts regarding melanoma may be more important than ever. According to the German Cancer Society, the prevalence of melanoma in Germany is steadily increasing: it is expected that the number of new cases will double in the coming 20 to 30 years for men and women.

In 2018 and 2020, studies identified transcriptional states of cancer cells in melanoma [3, 4]. According to their discoveries, specific states showed a higher drug-tolerance than others. In particular, specific cells showed increased proliferative characteristics, while others were more likely to possess invasive properties. This discovery is in line with a previous review, which established that aside from genetic heterogeneity, novel sequencing methods would allow transcriptional and epigenetic heterogeneity to play a major role in the understanding of cancer cell differentiation and the resulting gain of drug tolerance capabilities [5].

Taking into account the inherently invasive nature of melanoma compared to other skin cancers, the development of effective treatment methods is urgent [6]. Fundamentally relying on the understanding of cells within the tumor, study of intratumor heterogeneity plays a major role in this journey.

1.1 Background

While the words *tumor* and *cancer* should be familiar to every reader, what exactly defines and differentiates the two terms may be less clear.

The word cancer refers to a disease, in which cells begin to divide uncontrollably, while tumor refers to a sample of cancer cells within a patient.

More precisely, according to Hanahan and Weinberg the development of a cancer cell requires an accumulation of mutations in a number of different genes [2]. Importantly, a cell with mutations is not necessarily cancerous. The authors established that for a cell to be classified as cancerous, it must acquire following set of functional capabilities: (1) the cell must be self-sufficient in growth signals, (2) insensitive to anti-growth signals, (3) able to avoid apoptosis, and (4) able to divide in an unlimited manner. Finally, it must be able to shape its microenvironment (process of angiogenesis), and posses the ability to invade and survive in foreign tissue.

Tumors and cancers originating from different cell types can, in general, be understood as different diseases. For instance, while basal-cell carcinomas of the skin rarely metastasize, melanoma often form metastases and tend to become more malignant[1]. This becomes apparent when taking a glance at the large body of research directed to specific cancer types. While the alteration of a gene may be playing a major role in one cancer, the same effect may not be visible in another cancer type.

The realization that not only different cancer types express different characteristics, but that drastic differences can be observed within the same tumor, has lead to the study of intratumor heterogeneity and paved the way for personalised treatment of cancers. In our work, we took into account transcriptional and genetic heterogeneity within tumors. Healthy cells (which do not posses the mentioned hallmark characteristics) in a tumor microenvironment (TME) offer a further aspect to study heterogeneity, however such cells were excluded from our analysis.

1.1.1 Transcriptional Heterogeneity

Gene transcription refers to the process by which a gene is copied to a new messenger RNA (mRNA) molecule. Transcription is integral for cellular behaviour, since mRNA is a key ingredient of protein synthesis, which in turn drives a cells expressed characteristics (phenotype). Transcriptional heterogeneity can be defined as the variation of gene expression within a population of cells, measured by identifying mRNA molecules present within the cell bodies.

Epigenetics refer to inherited changes in gene expression that don't involve changes to the underlying DNA sequence. In general, genetic and epigenetic changes can lead to the gain and loss of function for genes. Gains and losses of function can be observed

[1] The tendency of a medical condition to become progressively worse. In terms of tumors: a malignant (also: cancerous) tumor differs from a benign (also: noncancerous) tumor by invading neighboring tissue and metastasizing (forming secondary tumors).

through gene expression profiles. In this work, gene expression profiles linked with the phenotype of cells within a population will be referred to as transcriptional states.

To annotate transcriptional states, researchers supply marker genes for existing and novel states they have discovered. Marker genes typically show an increased expression in their identifying state compared to other states. State signatures, comprising of sets of marker genes, are discovered by statistical analysis methods, such as gene-set enrichment analysis and differential gene expression analysis.

In their work, Rambow et al. supply signatures for four transcriptional states of melanoma cells: *invasive, pigmented, neural-crest stem-cell like (NCSC) and starved-like melanoma (SMC)* [3]. While it is generally believed that both the pigmented (in which cells are faster growing) and invasive state (in which cells are slower growing, but more motile) are reproducible in melanoma [7], both the NCSC (in which cells exhibit stem cell features, possibly enabling them to escape treatment [8]) and SMC state (in which cells expressing genes typically upregulated in nutrient-deprived cells) are not commonly found in literature.

	Invasive	**Pigmented**	**NCSC**	**SMC**
1	PDRX1	APOE	ANXA1	CD36
2	SOX4	PMEL	A2M	RNF121
3	TMSB4X	EDNRB	CNN3	FRAT2
4	ADM	TYR	AQP1	PAX3
5	NES	TYRP1	IGF1	DLX5
6	GPR143	KIT	GFRA1	ARMC7
⋮	⋮	⋮	⋮	⋮

Table 1.1: Subset of representative genes for transcriptional states in melanoma.

1.1.2 Genetic Heterogeneity

Different models have been devised for the understanding of clonal evolution in tumors [9]. By clonal evolution we refer to genetic changes of cancer cells within a tumor, such as mutations and somatic copy number aberrations, consequently occurring in a subset of cells that are descendants of a shared relative, which first expressed this genetic change.

The term copy number aberrations[2] describes gains and losses of regions within a cells genome, compared to another cells genome. For copy number analysis of cancer cells it is common to use a set of reference cells, which have been identified as healthy (or normal).

Genome instability refers to an increased frequency of mutations within a cells genome over time. It is characteristic for cancer cells and can cause chromosomal instability,

[2]Also referred to as copy number alterations or copy number variations.

which is associated with altered chromosome number or structure[3]. Both promote gene copy number abberations (CNAs), which offer a potentially high adaptive advantage to evolving cancers, through which they can adjust to changes in the tumor microenvironment [9, 11].

It is likely, that understanding genetic differences between subclones, identifying the driver mutations that caused such changes and the selective pressures that lead to the development of subclones, will support the identification of treatment targets [10].

[3]Gain or deletion of chromosome fragments, translocations, inversions.

1.2 Methods and Data

In this work, we used the programming language R [29] and associated bioinformatics packages. Importantly, we used two publicised semi-supervised classifiers to annotate transcriptional states and an R package to infer CNAs from gene expression profiles. In the following sections, we provide a short description of the methods and data used for our analysis.

1.2.1 Data

In this work, we used data on patient derived melanoma samples, collected by the Tumor Profiler Study [13]. Specifically, we used gene expression data obtained through single cell RNA-sequencing and cell-type annotations[4] generated by NEXUS Personalized Health Technologies at ETH Zürich.

Single-cell RNA-sequencing (scRNA-seq) counts the mRNA expression in single cells, resulting in gene expression profiles of cells within a tissue. It enables researchers to analyze the transcriptome of cells and is specifically useful for studying heterogeneity within a tissue, because the ability to distinguish between cells within a tissue presumes the ability to measure genes that set them apart [12].

To make sure that only cancer cells were object of our study, we used the mentioned NEXUS cell-type annotations paired with copy-number aberrations inferred from the gene expression data using the `infercnv-package`. We provide a description of the package in the methods section. Under the assumption that cancer cells and subclonal structures of cancer cells within the tumor can be identified through CNA profiles [14], we excluded NEXUS-annotated melanocytic and mesenchymal melanoma cells, that showed CNA-profile more similar to that of healthy cells.

Prior to classification, we pre-processed the data with several known methods. For each of the two classifiers used, we decided to perform cell cycle regression implemented in the `seurat-package` [15]. This method generated a cell cycle score using a list of known cell cycle related genes. Based on this score, gene counts are regressed to remove cell cycle effects.

To normalize gene expression values, we utilized the methods recommended for the two cell-type classifiers. More information in this regard can be found below, in the respective introductions of the classifiers. Lastly, we removed cells that expressed no genes and genes that were not expressed in any cells of the sample.

1.2.2 Visualization

For visualization of cell assignments, we used the `ggplot2-package` [28]. For dimensionality reduction of the raw count matrix we chose the method of uniform manifold

[4] *Remark:* Cell-type annotations do not give any information about the transcriptional state of melanoma cells.

approximation (UMAP) [26], which claims to preserve more of the global structure than other standard dimensionality-reduction methods, such as t-SNE.

More precisely, we decided to perform UMAP only on a subset of features: before running UMAP, we reduced the feature-set to only the set of marker genes. Intuitively, assuming that marker genes contain the most information about the transcriptional state of a cell, this prior reduction should not have a negative influence on the visualization of assignment results. Visually, this assumption was confirmed.

1.2.3 Transcriptional Assignments

For assignments of transcriptional states, we decided to use two different classifiers and compare their results on our dataset. A more detailed description of both models is provided in Chapter 2.

Cellassign

The `cellassign-package` [19] implements a semi-supervised method for the probabilistic assignment of cells to cell types. Using prior knowledge given in the form of expert-derived type-specific marker genes, `cellassign` fits a model using maximum a posteori (MAP) estimation and an expectation-maximization (EM) algorithm.

To run `cellassign`, one must provide size factors to the model, which are consequently used by the model to normalize the input data. To calculate size factors, as proposed by Lun et al. [16], cells from a cluster (in our case sample) are pooled together according to their library size. Pool-based size factors (the median ratio between the summed counts of all cells in the pool and the average counts across all genes) are calculated for various pooling configurations, through which the cell specific size factors are then inferred[5].

Apart from annotations on single sample count data, cellassign offers the possibility to take into account batch-effects for annotations in multi-sample data using the covariate matrix X. When calling cellassign, different parameters can be defined, such as the learning (convergence) rate of the optimizer and the maximum number of iterations the EM-algorithm shall perform.

SCINA

The `SCINA-package` [20] offers a similar semi-supervised classification algorithm for scRNA-seq data as cellassign. While `cellassign` assumes a general over-expression of marker-genes in their identifying cell type, `SCINA` assumes a bi-modal distribution of the gene expression. The two main building blocks of this method are the random initialisation of model parameters and the EM-algorithm to optimize the model parameters. Further, the model can be tweaked using the `convergence_rate` parameter, which dictates the convergence of the EM-algorithm.

[5] More information on the calculation of size factors can be found in the documentation of the R-package `scran` at https://rdrr.io/bioc/scran/man/

1.2.4 Subclonal Analysis

InferCNV

Using the `InferCNV package` [21] we were able to infer large scale chromosomal copy number aberrations using tumor scRNA-seq data. Copy number profiles of cells are inferred by comparing the expression intensity of genes across positions of the tumor genome in comparison to a set of reference healthy cells, which are supplied by the user.

Analysis of Phylogenetics and Evolution

To analyse and adapt the subclonal structure within a sample (given as a phylogenetic tree in the output of InferCNV), we used the `ape-package` [27].

The term phylogenetics refers to the study of evolutionary relatedness among groups of organisms through molecular sequencing data (such as DNA sequencing data or CNA data). Phylogenetic trees are a form of dendrograms, that contain evolutionary information (for subjects within a studied group), such as evolutionary distance and common ancestors.

Hierarchical Clustering

For hierarchical clustering of CNA-data we employed the `hclust` function from the `stats R-package` [29]. More precisely, because `InferCNV` performed hierarchical clustering using the Ward method (`Ward.D2`) at the time of the project, we decided to work with the Ward method as implemented by `hclust`.

1.2.5 Data & Code Availability

Code used for the analysis is available on the Boeva Lab shared drive at `./data/projects/Philip_Tumor_heterogeneity`. Data that has been generated as part of this work is also available on the Boeva shared drive at the previously mentioned location. For access to the TuPro scRNA-seq raw counts and NEXUS cell-type annotations, please contact the Tumor Profiler board.

Any data or code referenced in this thesis can be found at the mentioned locations with the respective file name. In case a directory under the name `./sample_output_dir` is mentioned, the term `sample` must be replaced by the sample name of interest.

Chapter 2

Semi-Supervised Classification of Melanoma Cells

As described by Rambow et al., transcriptional states of melanoma cancer cells seem to be identifiable by a reduced set of genes, more generally referred to as marker genes [3]. Further, general identification of cell types in healthy tissue is frequently performed based on sets of marker genes. In the past, manual annotation of cell types has been a costly issue in studies. To enable fast and accurate labeling, automated cell type assignment algorithms have been devised, which greatly reduce the cost of labeling.

Due to the absence of transcriptional state labels in our dataset, using supervised classifiers such as `CIPR` and `ClustifyR`, which rely on a labeled reference set for cell type assignments, was not an option. Instead, two methods from the class of semi-supervised classifiers were employed to assign transcriptional states to cells from the TuPro dataset.

In the context of this thesis, the term *semi-supervised* will refer to classifiers that use a specified set of marker genes as prior knowledge. Using this prior knowledge, the classifier can probabilistically assigns a cell to a type based on transcriptomic data.

Questions regarding the labelling of cells by transcriptional state, which accompanied this project, included:

- Can semi-supervised classifiers be applied to annotate the transcriptional state of melanoma cancer cells?
- Is it possible to validate the labelling in absence of a gold-standard reference?
- How well do semi-supervised classifiers perform on cancer cells?

The following chapter will first provide an introduction into, and comparison between the classifiers employed, followed by a description of annotations that both classifiers delivered. Evaluation of the results can then be found in the next chapter.

2.1 Introduction of the Models

As previously mentioned, both SCINA and Cellassign use marker gene signatures to probabilistically annotate cells by their type. Fundamental for both classifiers are expectation maximization (EM) algorithms, which find maximum a posteori (MAP) estimates of model parameters. It is important to note, that the EM algorithm will converge to a local optimum, however it does not guarantee the convergence to a global optimum. For this reason, a 'smart choice' of model parameters is necessary, to ensure that a model performs as well as possible for a given task.

Apart from both classifiers employing the EM algorithm for parameter tuning, there are subtle differences, which will become apparent in the model descriptions below.

2.1.1 Cellassign

In general, cellassign uses a Bayesian statistical model, which is optimized through use of the EM algorithm [31]. Inputs to the model are the size factors s (per cell), which are used for normalization purpose. Furthermore, a binary signature matrix ρ (containing information on the marker gene to cell type mapping), a raw gene expression matrix Y and (optionally) a covariate / design matrix X to account for batch effects are inputs to the model.

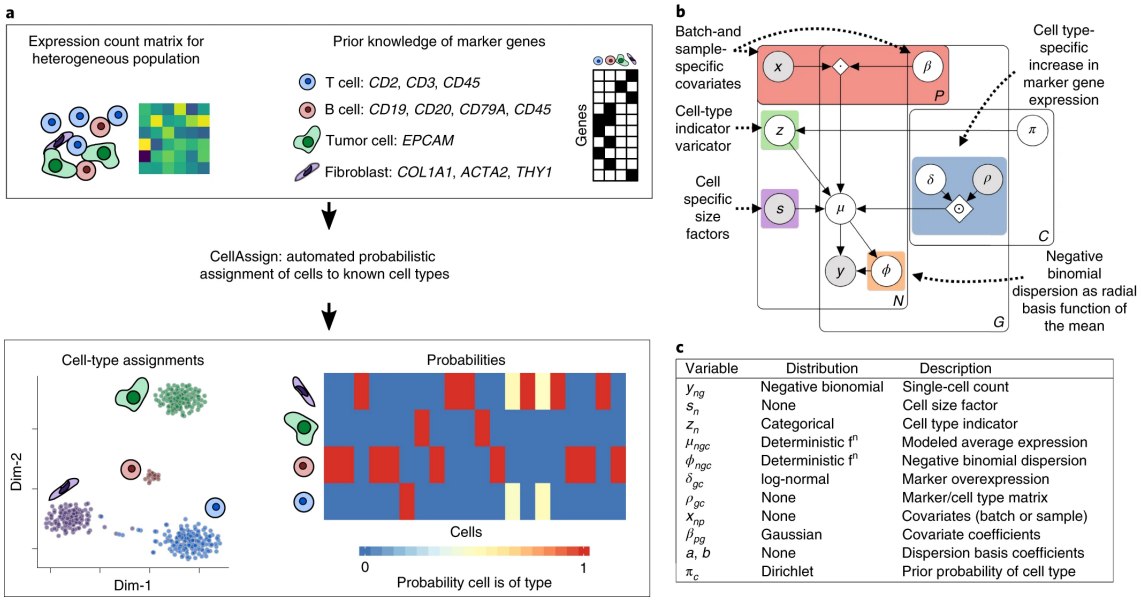

Figure 2.1: Illustrated summary of the cellassign model. **a,** General overview of the functionality that cellassign offers: given a marker gene matrix, for each cell and type the model calculates the probability that the given cell is of the given type. **b,** The random variables of the model and their relationships. **c,** List of the model's random variables and their prior distributions.

With the model inputs Y, ρ and X it is then possible to infer the probabilities that a cell z_n is of a given cell type c. Formally this is the probability $P(z_n = c|Y, \hat{\Theta})$ where $\hat{\Theta}$ are MAP estimates inferred by the EM algorithm. To be able to predict this probability, the authors of cellassign decided to use the expected gene expression $\mu_{ngc} = \mathbb{E}\left[y_{ng}|z_n = c\right]$, which takes into account the three cellassign inputs:

$$\overbrace{\log \mu_{ngc}}^{\text{Log mean expression}} = \underbrace{\log s_n}_{\text{Cell size factor}} + \overbrace{\delta_{gc}\rho_{gc}}^{\text{Cell type specific}} + \underbrace{\beta_{g0}}_{\text{Base expression}} + \overbrace{\sum_{p=1}^{P} \beta_{gp} x_{pn}}^{\text{Other covariates (incl.batch)}} .$$

Using the prior distributions outlined in 2.1, it is then possible to optimize the choice of parameters using the EM algorithm (a formal description can be found in the methods section of [19]). Code to run the cellassign classifier is available at `./scripts/CELLASSIGN.R`.

2.1.2 SCINA

In contrast to the Bayesian model of cellassign, SCINA operates by a bimodal distribution of marker genes in the sequenced cells [31], where the second peak (cells with high expression of genes in the specific signature) identifies a cells type.

Using the EM algorithm, SCINA estimates the probabilities of a cell being of type c, the two distribution centers $\mu_{1,r}, \mu_{2,r}$ (due to the assumption of bimodal distribution) for each marker gene r and the covariance matrices Σ_1^r, Σ_2^r (which are assumed to be equal) corresponding to each marker gene r. The EM algorithm terminates, once the cell type labels have stabilized, i.e. their change between EM iterations is sufficiently small. A more detailed description of the EM algorithm can be found in the supplements of [20].

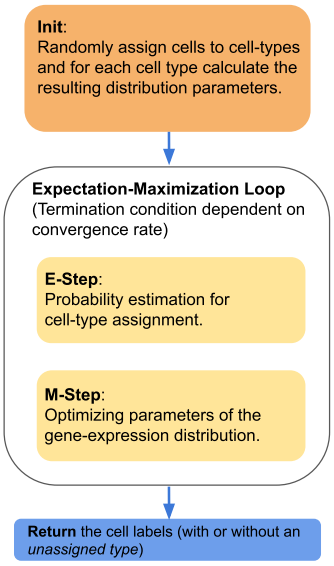

Figure 2.2: Schematic overview of the SCINA model.

Having computed the probability of a cell being of type c for each cell, the cell type with the highest probability is chosen. As with cellassign, SCINA allows the annotation of an 'unassigned' state. This state, similarly to cellassign, assumes missing over-expression of any marker gene supplied to the model.

It is noteworthy that labelling with the SCINA model runs much faster than that of cellassign. In our work environment, running SCINA could relate to a drastic runtime decrease of 10-100 times that of cellassign. Code to run the SCINA classifier is available at `./scripts/SCINA.R`.

2.2 Expression Analysis of Marker Genes

Before using the two classifiers to generate transcriptional state labels, it was important to make sure that, using the supplied marker genes, it is theoretically possible to distinguishing between transcriptional states of cells present in the sample.

Through visual confirmation it became apparent, that a differentiation between cells of the four transcriptional states, based on marker gene expression values, should be possible. The reasoning behind this hypothesis is, that firstly the quality of the Rambow signatures is sufficient: for each signature there are cells that express some gene from the signature. Secondly, separation should be possible due to the observation, that cells show different gene expression profiles within a sample.

As shown in Figure 2.3, where we reduced the set of displayed marker genes per transcriptional state to the the five that showed the highest overall gene expression across the sample, differences in gene expression between cells become well visible.

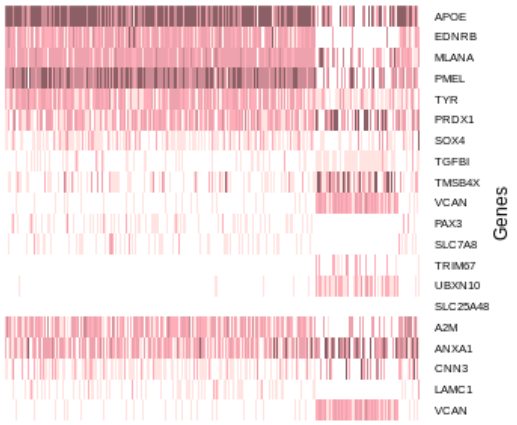

Figure 2.3: Heatmap of log-normalized gene counts generated by SCINA. For each signature, the 5 genes with the highest overall expression have been kept.

2.3 Classification Results

In the following two sections, we will briefly analyse the results of both cellassign and SCINA, why we believe that some samples had an unusually large fraction of `'unassigned'` cells and what the general takeaway of the results is. In the next chapter, these results will then be used to for evaluation of both methods.

The results of both models have also been uploaded to the shared drive. For each sample they can be found in the sub-directory `./sample_output_dir/results`, where they are stored in the form of a table.

2.3.1 Results of Cellassign

We called cellassign with the EM convergence parameter Θ set to $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and found the convergence rate of `1e-4` to work best for our classification task, according to the distribution of cell states that we expected a priori. Due to prior studies of transcriptional states in melanoma, we expected the pigmented state to be discovered most frequently, with cells of the invasive state being the second most likely. Regarding the SMC and NCSC signatures, we expected the cell state to occur with the least frequency.

In general, as expected, the cellassign model annotated a majority of cells as pigmented. This effect could be seen across samples, with the pigmented labels assigned to, on average, 53.4% of cells within a sample. The second most occurring label of cellassign was unassigned (on average 40.6%), followed by invasive (4.6%) and SMC (1.2%). Annotations of NCSC by cellassign were negligible, with close to no cells being annotated with this transcriptional state.

As visible in Figure 2.5, cells in the samples MOVAZYQ and MECYGYR were predominantly labelled as invasive. Due to the unexpected composition of labels, we further investigated the gene expression in both samples. First, we calculated the five most expressed marker genes for each signature. We then normalized counts by library size. The resulting heatmaps in Figure 2.4 show that gene expression, especially the abundant overexpression of invasive state marker gene MGP, offers an explanation for the large number of invasive annotations in the MECYGYR sample. The same applies for the MOVAZYQ sample, where TMSB4X showed a strong increase in expression intensity.

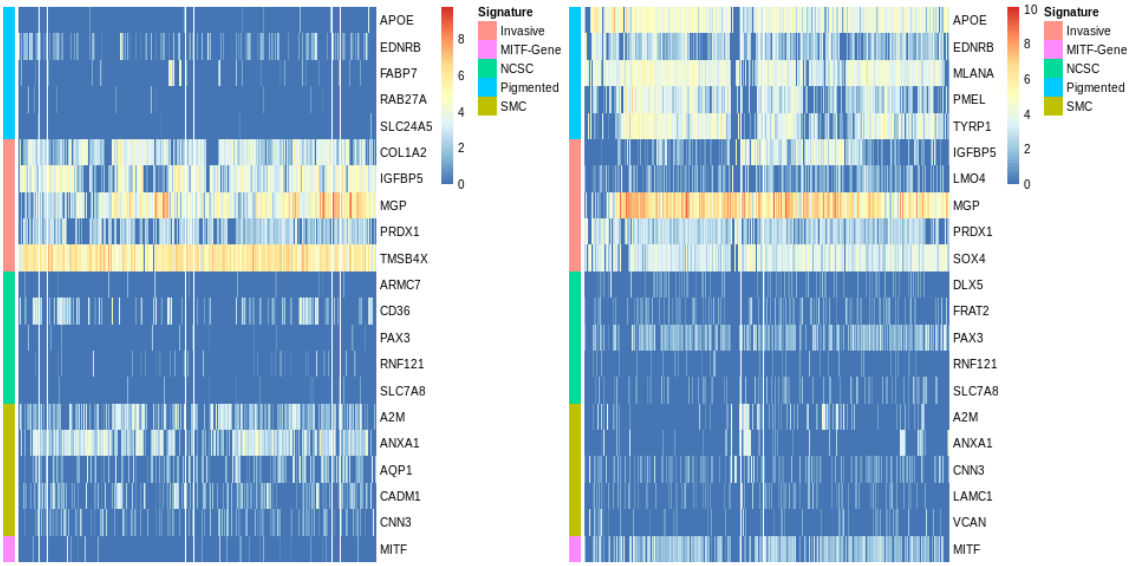

Figure 2.4: Gene expression heatmaps of the most expressed marker genes. Expression values were normalized by library size and log-transformed. Depicted are gene expressions for the samples MOVAZYQ (left) and MECYGYR (right). The signatures corresponding to the genes have been annotated on the left of both heatmaps.

Interestingly, while the pigmented signature in MECYGYR was still present, no trace of the pigmented signature was observable in MOVAZYQ. Analysing the MITF-expression in both samples, we found that there was no trace of MITF-expression intensity in MOVAZYQ, while MITF showed expression in MECYGYR. Given the central role of MITF in melanoma biology and the assumption, that MITF will be up-regulated in pigmented and down-regulated in invasive state cells [3, 7], it is possible that cells in the MECYGYR sample are in transition between the two states.

2.3.2 Results of SCINA

For SCINA, we tried a more fine grained grid-search on the convergence rate of the EM algorithm, iterating Θ over $\{0.01, 0.02, \ldots, 0.98, 0.99\}$. We then used the clustering evaluation metric of intracluster compactness (introduced in Chapter 3), to choose the best convergence rate. This optimization method leads SCINA to use model parameters, for which cells with the same label are most correlated on average.

As visible in Figure 2.5, SCINA returns a more evened out set of state assignments. Against our prior assumptions, both the NCSC and SMC state were assigned frequently, with annotations of the two states even exceeding those of invasive and pigmented in some samples, such as `MEFOCUR` and `MECYGYR`. The likelihood of such annotations, as is visible in the gene expression of the `MECYGYR` sample (Figure 2.4), is not high.

As we will establish in Chapter 3, the separation between clusters (measured by the inter-cluster complexity) is not of high quality in SCINA annotations. This does not come as a surprise due to our assumption, that the pigmented state should be the most occurring annotation for melanoma cells.

2.3.3 Comparison

To compare assignments of cellassign and SCINA, we looked at three different factors: how often are cells annotated to be unknown and to which degree are transcriptional states assigned, that we expect to be less common a priori.

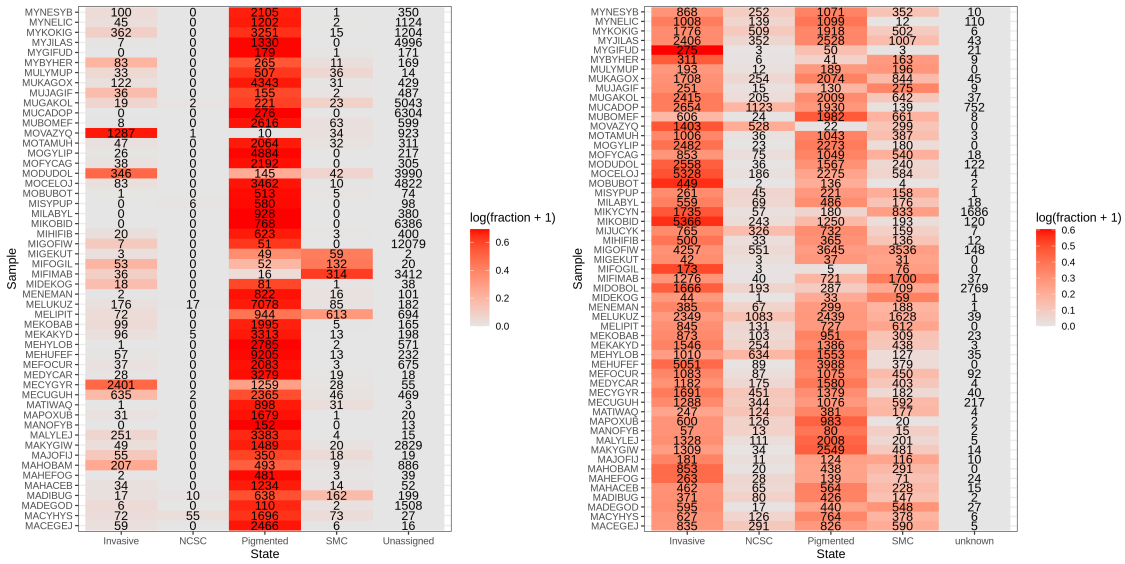

Figure 2.5: Heatmaps to visualize annotations by cellassign (left) and SCINA (right). In both plots we have coloured by the fraction of cells from each state in respect to all cells not annotated as unknown / unassigned.

As visible in Figure 2.5, the annotation profiles of the two models are notably different. In sections 2.3.1 and 2.3.2 we established that cellassign annotates a majority of cells to the pigmented state, with the second most prominent label being the invasive state. On the other hand, SCINA did not discover a state that clearly dominated labels across samples.

Furthermore, the frequency with which cells were labeled as unknown, and unassigned respectively, deviated between SCINA and cellassign. While cellassign made use of the unassigned label frequently, SCINA was much more confident in the assignment of transcriptional states, assigning a high number of unknown labels only in the samples `MIDIBOL` and `MIKICYN`. Code to generate the comparison table is available at `./scripts/color_by_percentages.R`.

Chapter 3

Evaluation of State Assignments

To determine how accurate cell assignments are and which model performs better, we evaluated the cell assignments of both. Due to the absence of a gold-standard reference (labels annotating cells by their true transcriptional state), it was not possible to use metrics such as the specificity and accuracy to evaluate cell assignments.

Instead we decided to adapt evaluation techniques for unsupervised classifiers, such as clustering algorithms. The evaluation was therefore based on metrics that evaluate the shape and separation of clusters. During evaluation we assumed, that marker gene expression profiles identify the transcriptional state of a cell. Therefore cells assigned to the same state should be similar based on the expression of respective signature marker genes, while cells not assigned to this state should be less similar based on their expression profile of marker genes. This meant that we could discard expression values of genes which were not present in any signature.

For similarity measurement two measures were applied:

- *Pearson Correlation* between cells
- *Bayesian Correlation* between cells.

Questions that a validation metrics should aim to answer in absence of a *gold standard reference* include:

- Are cells with a similar transcriptome annotated to the same cell type?
- Are cells with a different transcriptome annotated with a different label?

3.1 Clustering Evaluation

As previously pointed out, we heavily relied on similarity of cells that had been annotated with the same label. For the development of our evaluation metrics, we were heavily influenced by [22] and [18].

3.1.1 Definitions

For the following definitions, let Y_q denote the gene expression matrix of marker genes for cells annotated with state q.

We define the term of ***intra-cluster compactness*** for a transcriptional state q as:

$$Compactness = \frac{1}{n} \sum_{i=1}^{n} corr(c, Y_{q,i})$$

where n := *#Cells in state k* and c denotes the mean gene expression values of cells labelled with state q.

We define the term of ***inter-cluster complexity*** for a transcriptional state k as:

$$Complexity = \max_{r \in R} \frac{1}{n_k} \sum_{i=1}^{n_k} corr(c_r, Y_{k,i})$$

where R denotes the set of transcriptional states without k.

Using complexity as a measure of separation, a clustering could be considered to be desirable if the values are low, i.e. close to 0.

3.2 Evaluation of Assignment Results

We ran our evaluation using two different correlations: Bayesian correlation and Pearson correlation. We tested the Bayesian correlation because a previous publication by the University Hospital Bern had suggested that it delivers values more suitable for comparison of scRNA-seq data [23]. However, in our case we found the Bayesian correlation not to work. We therefore decided to perform the evaluation with the Pearson correlation.

We further included a baseline complexity and compactness as a reference, against which we could compare our assignments. We generated the baseline compactness by calculating the average cell of all malignant cells and calculating its correlations with all malignant cells. The baseline complexity was calculated by sampling two sets of malignant cells and then calculating the complexity between the two.

After computing the two metrics for cellassign and SCINA annotations, we computed their differences to the baseline values for each state. The results have been visualized as boxplots in Figure 3.1.

Through our evaluation we were able to confirm a set of prior assumptions. It became apparent that cellassign outperforms SCINA with regard to both the similarity of cells annotated with the same label. This becomes even more clear, when looking at the separation of cells from different transcriptional states, which we quantified with the complexity measure. This is in line with our expectations, as SCINA assigned to existing

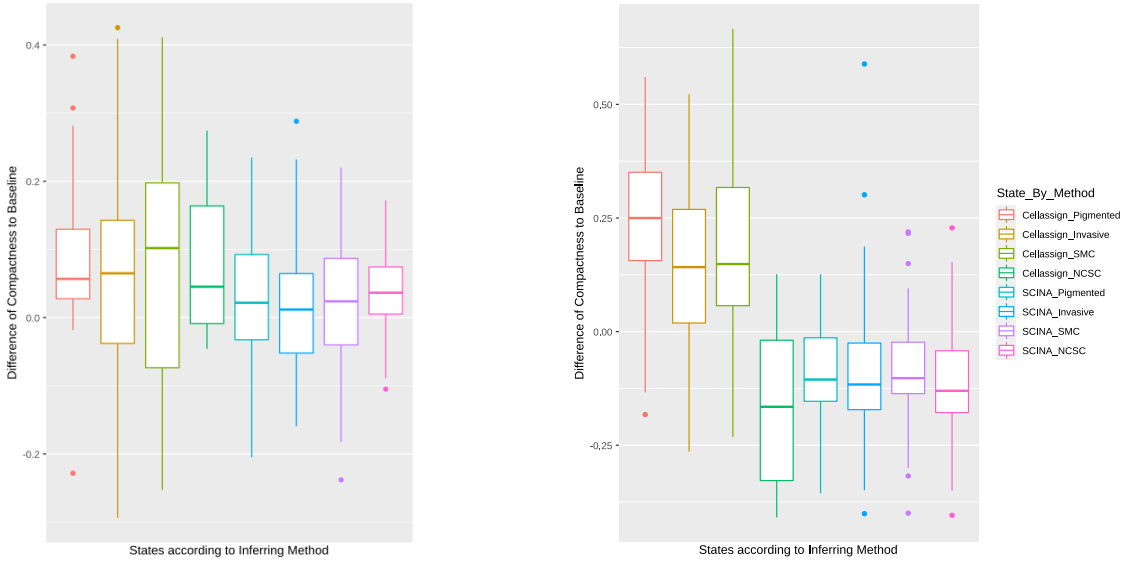

Figure 3.1: Boxplot comparing compactness (`left`) and complexity (`right`) to the baseline values for both classifiers. It is clearly visible that the complexity of annotations in SCINA is worse than in cellassign. In both figures, negative values correspond to the situation, that the baseline was better than annotations by a classifier. In both plots, the four leftmost boxes correspond to states annotated by cellassign and the four rightmost boxes correspond to states annotated by SCINA.

transcriptional states quite evenly, increasing the likelihood of cells from the same true state to be annotated with different states, therefore decreasing cluster separation.

Furthermore the inconsistent quality of annotations, especially by cellassign, for specific states became clear. This inconsistency was driven by two factors: the transcriptional state on the one hand and the sample on the other. While annotations for samples such as `MALYLEJ` and `MYNESYB` showed exceptionally good values of both compactness and complexity, annotations for samples like `MAHEFOG` and `MEHUFEF` showed values that did not even beat the baseline (see Suppl. Table A.1 and Suppl. Table A.2).

We conclude that when annotating the transcriptional state of melanoma cells by use of semi-supervised classifiers, one should resort to the use of cellassign - regardless of the quite drastic increase in runtime. When using the classifier, we recommend using the inter-cluster complexity to measure the plausibility of annotations.

The full evaluation tables for cellassign (and SCINA) have been made available on the shared drive at `./cellassign_eval_inter.txt` and `./cellassign_eval_intra.txt`. Replace 'cellassign' by 'scina' to retrieve SCINA evaluation results. The evaluation script has also been added to the shared drive at `./scripts/evaluation.R`.

Chapter 4

Subclonal Structures in Melanoma

Research has shown, that CNAs play a key role in the development of subclonal structures in a variety of cancer types. Influenced mainly by genomic instability, CNA profiles vary between subclonal structures. By utilizing this variance, it is possible to characterize subclonal structures in tumors based on CNA profiles.

In our work, we inferred copy number profiles from scRNA-seq data for melanoma samples. Using predicted copy number profiles and transcriptional state annotations of cellassign, we studied the association of transcriptional states with subclonal structures.

Motivating this approach were two findings. Firstly, collaborators at the Cochin Institute discovered that it is possible to link certain transcriptional states in neuroblastoma with subclonal structures based on copy number profiles. In their unpublished work they found, that cells expressing the invasive signature associated with subclonal structures that were different from those associated with the NCSC transcriptional state signature. Further, results from a pan cancer study in 2019 suggest a correlation between CNAs and differential gene expression [14], which in turn plays a key role in transcriptional heterogeneity.

By employing the Ward method for hierarchical clustering, InferCNV generates a phylogenetic tree. We inferred subclonal structures, based on the branching structure of the phylogenetic tree. Cancer cells within a subclonal structure are assumed to express a similar copy number profile, i.e. copy number gains and copy number losses occur at the same genomic regions.

4.1 Inferring Subclonal Structures

To study the genetic heterogeneity in melanoma, we inferred copy number profiles and predicted subclonal structures based on scRNA-seq data. In the following section, we give a short description how copy numbers were inferred and go into detail on the method we implemented, to predict subclonal structures. Lastly, we discuss the design

of our stopping condition for the tree traversal on the InferCNV-derived phylogenetic tree, which was fundamental for a working subclonal structure prediction.

4.1.1 Copy Number Prediction

Using the R-package InferCNV, we inferred the copy number profiles for each sample[1]. To predict copy numbers, it is necessary to supply a set of reference cells, typically consisting of healthy cells. On sample-to-sample basis, we used cells annotated by NEXUS as T lymphocytes, endothelial cells, resting NK cells and B lymphocytes as a healthy reference to estimate CNAs in malignant cells. We used the gene ordering file available on the shared drive at `./new_gene_annot_file.txt`, which was created using the human GRCh37 assembly.

To confirm the validity of copy number profiles predicted by InferCNV, we visually compared them to copy number profiles inferred using scDNA-seq data for the same sample. We assumed that an inferred copy number profile is valid, if significant copy number gains or losses occur at the same genomic regions for both methods. Importantly, cells were from the same tumor but not present in both datasets. This is due to the limitation, that both sequencing methods destroy the cell during their analysis. As visible in Suppl. Figure A.3, even regions of high detail were similar in predicted copy number profiles of both methods. Interestingly, predicted subclonal structures for scDNA-seq data also showed a high similarity to our predictions.

4.1.2 Phylogenetic Analysis

Using the phylogenetic tree generated by InferCNV, the question arose, at which depth of the branching structure we could declare a cell-cluster to be a subclonal structure. This presented an important design decision, as setting a high threshold could lead to artefacts in our results, while setting a low threshold could mean a loss of information.

Instead of using an absolute threshold on node depth, we decided to use a relative threshold between cell-clusters as a stopping condition. More precisely, let S_A, S_B be the child-trees of the root node of tree S_0. Because the ward method minimizes variances within clusters, we decided to analyse the difference between the variances of clusters S_A and S_B, relative to that of S_0. Intuitively, we defined the intra-cluster variance as the mean of all vector variances, where for each gene a vector contains the CNAs for cells within the cluster (see Suppl. Figure A.4).

This resulted in the stopping-condition:

$$|\text{variance}(S_A) - \text{variance}(S_B)| \leq \varepsilon \cdot \text{variance}(S_0).$$

To find a suitable ε, we performed Wards hierarchical clustering on a per-sample basis for reference cells and other healthy cells not included in the reference set. We then

[1] The results of this work have been made available on the shared drive at `./inferCNV_test/sample_output_dir`

calculated the intra-cluster variances on the top level and computed ε. Assuming the non-existence of subclonal structures in populations of healthy cells, any significant subclonal structure should be identifiable with an ε larger than that for healthy cells.

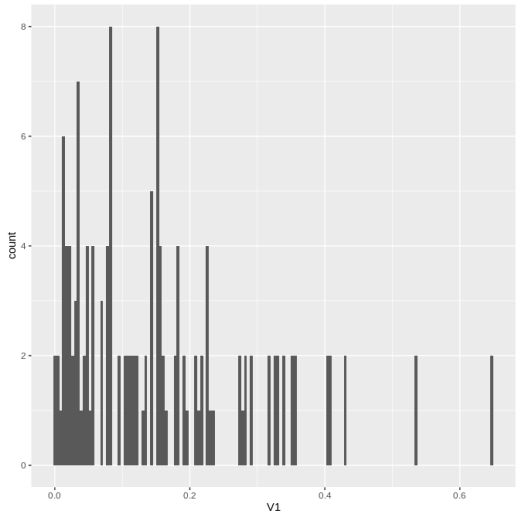

Figure 4.1: Distribution of the relative threshold ε in reference and healthy cells for all samples.

We identified the highest 5% values, resulting in $\varepsilon = 0.4$. Testing against other (larger & smaller) values, we found that while suppressing noise in the choice of subclonal structure, it preserved a good amount of information about structure within in the cell populations. This can be visually confirmed in the infercnv plots, which have been generated for each sample and contain annotations for subclonal structures (see Suppl. Figure A.3).

Subclonal structure annotations with $\varepsilon = 0.4$ were then generated using the script `infer_subclonal.R`, available on the shared drive at `./scripts`.

Furthermore, we implemented an automated split to differentiate between healthy and malignant cells. Goal of the method is to remove cells annotated by NEXUS as malignant, whose copy number profile is highly similar to that of non-malignant cells.

To achieve this, we used the phylogenetic tree computed by InferCNV. We excluded cells (more specifically a subtree) from the malignant set, if at least 50% of the cells in that subtree have been annotated as healthy by NEXUS. If this condition is met, we check if healthy NEXUS-annotations exceed 50% in both subtrees. If this condition is met, all cells on the leaf of the tree are annotated as healthy. Else, we continue traversing the branching structure.

Barcodes of the inferred malignant cells have been made available per sample on the shared drive in the file `./sample_output_dir/inferred_mel.txt`. Code to generate subclonal annotations is available at `./scripts/subclonal_annotation.R`. To generate healthy-cancer splits, run `./scripts/subclonal_remove_healthy.R`.

4.2 Association of Transcriptional State and Subclonal Structure

With the question of association between subclonal structures driven by CNAs and transcriptional state identified by the transcriptome in mind, we quantified the similarities between the two. Having generated both transcriptional state and subclonal structure for our samples, we calculated the overlap of both annotations for all samples.

To calculate the overlap between transcriptional state and subclonal structure, we calculated the fraction of cells from a transcriptional state T, that have been assigned to a subclonal structure S. Using this information we generated a heatmap, to visualize if and how the two cell annotations relate.

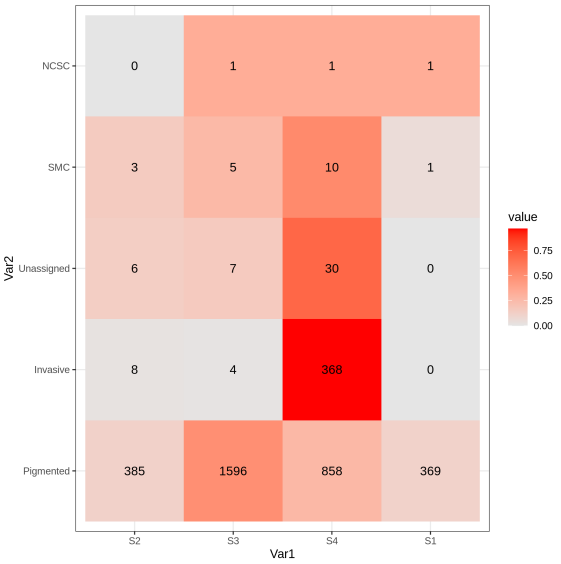

Figure 4.2: Overlap between transcriptional state and subclonal structure in the MALYLEJ sample. Heatmap tiles have been labeled with the cell-count present in the corresponding configuration.

Using the rand index [33], we filtered out samples where cells annotated as invasive and pigmented showed a high degree of overlap in subclonal structure. In short, the rand index looks at cell-pairs and how they relate in our two annotation-methods. Because our goal is to have cells from the same transcriptional state to be in the same subclonal structure and cells from different transcriptional states to be annotated to different subclonal structures, we prioritized the analysis of samples with a high rand index.

Overlap heatmaps have been made available for each sample in the shared drive at `./sample_output_dir/conf_matrix.png`. A visualization of the rand indices calculated for each sample is available in the supplementary material and on the shared drive at `./rand_indices.png`. Associations were calculated using code from `./scripts/overlap.R`.

4.3 Genomic Regions of Interest

Having extracted the overlap between transcriptional state and subclonal structure, we set out to analyse genomic regions of interest, that may be involved in cell differentiation to the invasive state. Such regions could be derived by looking at the differences between subclonal structures associated with invasive and pigmented melanoma cells.

We decided to determine driving CNAs on a chromosomal and gene level. Because InferCNV returns CNAs per gene, we first computed the genes, that are likely to be driving

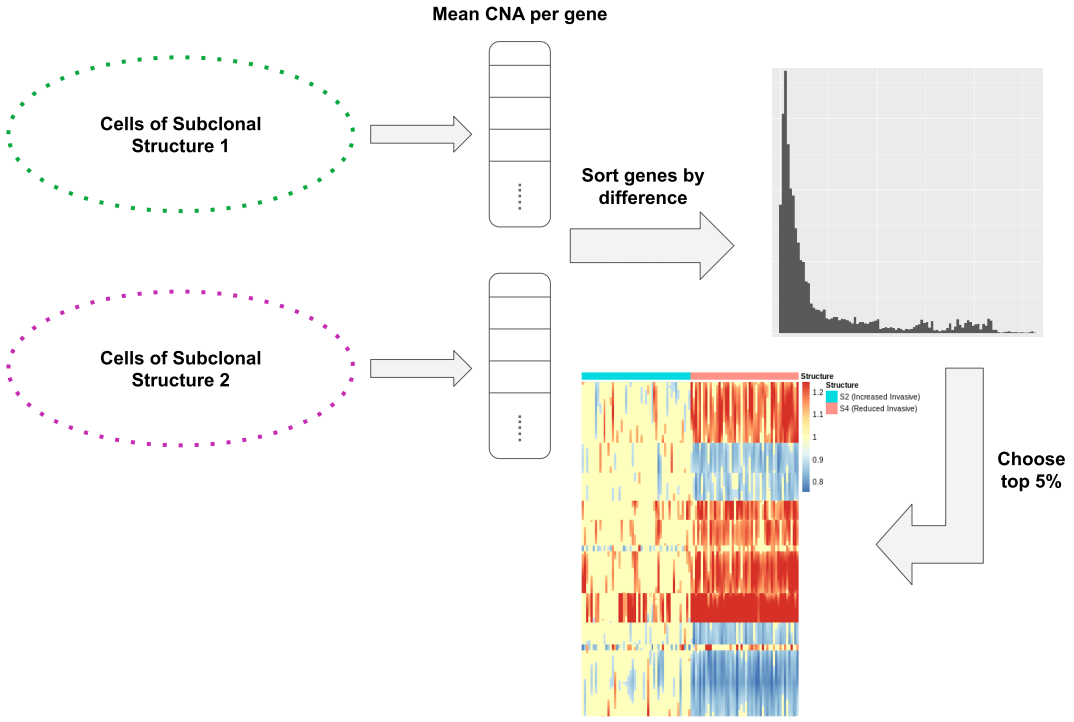

Figure 4.3: Overview of discovery pipeline for highly different genes between subclonal structures with regard to CNAs. After computing differences between per-gene average CNAs, we choose the 5% genes that are most different between subclonal structures.

the differentiation of cells to the studied subclonal structures. On an abstract level, we viewed subclonal structures as clusters and computed the clusters center (equivalent to computing the mean CNA per gene in a cluster). Then, to determine driving genes of subclonal differentiation, we calculated the distance between cluster-centers and extracted the top 5% distant genes (see Figure 4.3). Heatmaps generated have been made available for each sample on the shared drive at `./sample_output_dir`.

The computed genes were then used to generate heatmaps of CNAs for cells from the two subclonal structures (see Suppl. Figure A.1). Further, for each gene we counted its occurrence in the top 5% different genes across samples. While an analysis for all 67 samples was not possible[2], we were able to analyse differentiating genes for approximately 50 samples.

We found that there was not a single gene, that was present in the intersection over all samples, with the highest count over samples being 26. Interesting candidates for further study include the genes `S100A13`, `JTB`[3] and the genes `EPN1` and `S100A6`.

[2]Reason being that some samples did not have sufficient malignant cells to utilize cellassign and others lacking the numbers of invasive annotations.

[3]Both genes were among the top 5% genes in 24 samples, therefore being among the top 5 genes.

For S100A13, Tirosh et al. suggest a positive correlation between expression levels and metastasis [35], while the same study suggests a negative correlation of its expression values and invasiveness, which seems contradictory and therefore makes this gene an interesting target of study in the TuPro data. In addition, gene expression values of S100A6 showed a high correlation with metastasis, and in a study by Gerber et al. correlated with their EMT signature[4] for melanoma [36].

For future studies, the full list of occurrences has been made available on the shared drive at `./gene_occurrence_top_diff_subclonal.txt` and in reduced form at `./top_subclonal_genes.xlsx`.

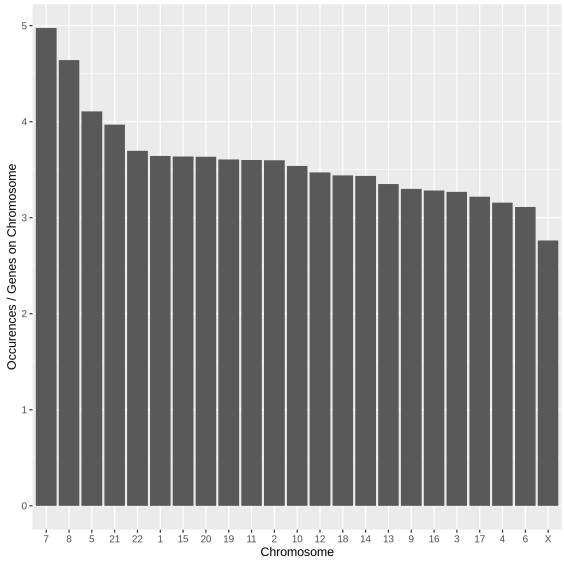

Figure 4.4: Chromosomes containing genes that are highly different between subclonal structures associated with the invasive and pigmented state. Counts were normalised by the number of genes located on the chromosome.

To study the influence of CNAs on a large scale, we converted the most distant genes to the chromosome that they are associated with (according to our gene ordering file). We plotted the sum of chromosome occurrences across samples (normalised by the number of genes located at the chromosome) and found CNAs to occur with the highest frequency at chromosomes 7 and 8. This discovery motivates further study of the proposed genes of interest, especially considering a possible impact of chromosome 7 aneusomy[5] on metastasis of melanoma [34].

The implementation of 4.3 can be found at `./scripts/subclonal_drivers.R`. To count occurrences of genes in the top 5%, we ran `./scripts/region_analysis.R`.

[4]Cells in the epithelial–mesenchymal transition (EMT) show a tendency to gain migratory and invasive properties.

[5]Condition where different nuclei in an organism contain different numbers of chromosomes.

Chapter 5

Conclusion

In this work we have analysed the association of subclonal structures inferred from predicted copy-number profiles and transcriptional states inferred using marker gene signatures. To conduct this analysis, it was necessary to infer transcriptional state labels using the existing semi-supervised methods *SCINA* and *cellassign*. Transcriptional state predictions were performed with and without regression of cell cycle effects. We found that while predictions with prior removal of cell cycle effects were more stable and showed marginally better results in the validation, performing cell-cycle effect removal is not absolutely necessary. A plausible reason for the relatively small effect on prediction quality is, that both classifiers only take into account the small number of marker genes for state-predictions.

In general, based on the evaluation metrics employed, cellassign showed much better classification quality. When employing cellassign one should keep an eye on the proportion of cells classified as unassigned / unknown. This proportion is highly dependent on the model parameters `min_delta` and `learning_rate` and a high proportion of unassigned cells *can* be a sign of bad parameter choice. Analysing samples, where annotations diverged from the expected distribution of transcriptional states, we found that indeed cellassign had annotated correctly - based on marker gene expression profiles.

Using copy-number profiles inferred by InferCNV, we were able to utilise the phylogenetic tree to predict subclonal structures based on cells' copy-number profiles. To infer a plausible depth for splitting the phylogenetic tree, we defined a stopping condition based on a relative threshold. Based on the difference between intra-cluster variances, which we have defined as part of this work, subclonal structures of different granularity levels can be predicted. We advise the use of our proposed relative change `EPS` for the stopping condition.

We visually confirmed the validity of our results, by comparing predicted subclonal structures and inferred copy numbers generated using scRNA-seq data to with results computed from scDNA-seq data. Quantifying the overlap of transcriptional states and CNA-subclones, we identified subclonal structures that indicate an association of tran-

scriptional and genetic heterogeneity. Finally, we identified a set of candidate genes for further study regarding the influence of CNAs on cell plasticity in melanoma.

5.1 Future Work

This work has given an insight into the applicability of semi-supervised methods for the transcriptional state labeling in scRNA-seq data sets of melanoma samples. Our evaluation metrics showed that the classification goals of both models are conserved in melanoma data. Further, we showed that the profiles of cells annotated with a transcriptional state showed an increased expression of the respective gene signature. For a full validation of classification results in melanoma data, it is however necessary to be able to use data with known transcriptional labels. To achieve this, two methods seem plausible. On the one hand, annotation of transcriptional states may be performed using biological methods. Due to the high expected costs of biological annotation, we propose to instead generate a generic scRNA-seq data set, for which transcriptional states will be known per definition.

Generating such a data set *in silico* will enable a full analysis of prediction quality and accuracy. As for feasibility, because the employed semi-supervised models only took into account the small set of marker genes, the daunting task of generating scRNA-seq data is reduced only to the question, how RNA-expression can be modeled for the roughly 100 marker genes. While in general systems for simulation of scRNA-seq data have been proposed [24, 25], a model for type-specific scRNA-seq generation employing prior knowledge in form of a marker gene signatures has not yet been developed.

An additional step to consider for future works around semi-supervised classification of transcriptional states, is using sampling prior to classification. Especially when classifying rarely expressed transcriptional states, trading the number of cells analyzed with transcriptome coverage can be of benefit when answering a given biological question, for instance that of transcriptional states within a sample [17].

Furthermore, in this work we have shown, that an overlap between transcriptional and genetic heterogeneity is likely in malignant skin melanoma. We showed that transcriptional states associate with subclonal structures inferred from copy-number profiles. As a consequence, the next step in studying the association of the two factors of heterogeneity is an in-depth analysis of genomic regions driving differentiation.

Genes showing contrary copy number profiles across samples have been identified for future study. A correlation between copy number aberrations at those genes and the gene expression of MITF, or marker genes of the transcriptional states, remains to be shown. Showing such correlations would, in effect, present an indication for an association of states and subclones. If such an association is reproducible, the next logic step would be establishing CNAs at the respective driving genes as further bio-markers for transcriptional states in melanoma.

Appendix A

Supplements

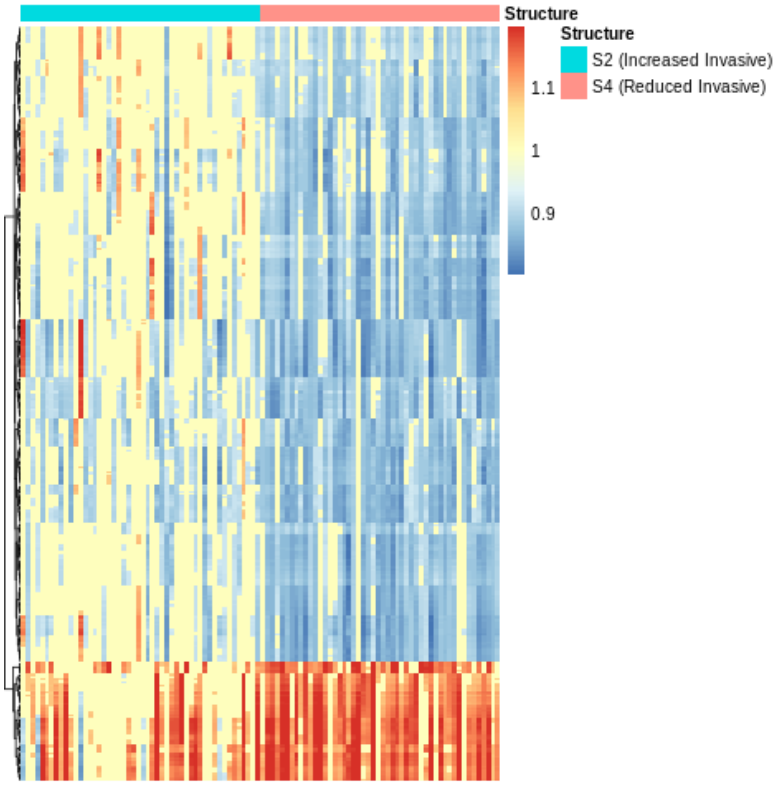

Figure A.1: Heatmap of CNAs potentially driving differentiation between subclonal structures associated with invasive and pigmented transcriptional state. Subclonal structure 2 (left) was associated with the invasive transcriptional state, while subclonal structure 4 (right) was associated with the pigmented state. It is clearly visible that the degree of copy number deletions and copy number gains for the depicted genes in subclonal structure 4 is much higher than in subclonal structure 2. Sample: `MALYLEJ`.

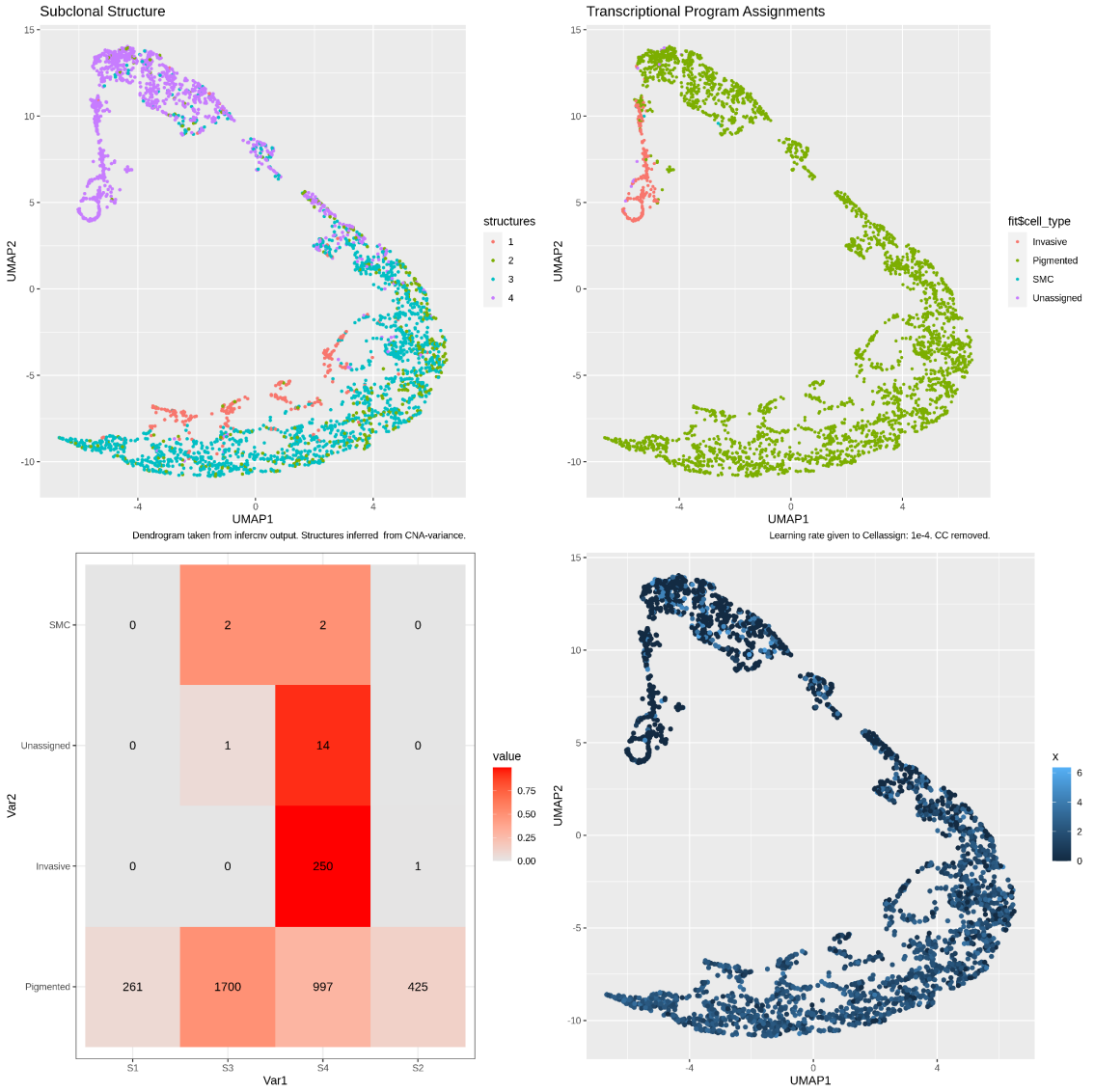

Figure A.2: Merged file containing UMAP plots for malignant cells annotated by subclonal structure (top-left), transcriptional state (top-right) and coloured by MITF-expression (bottom-right). Overlaps between transcriptional state and subclonal structure have been visualized in the heatmap (bottom-left). Plots correspond to the `MALYLEJ` sample. Merged files for other samples are available on the shared drive at `./sample_output_dir/merged_file.txt`.

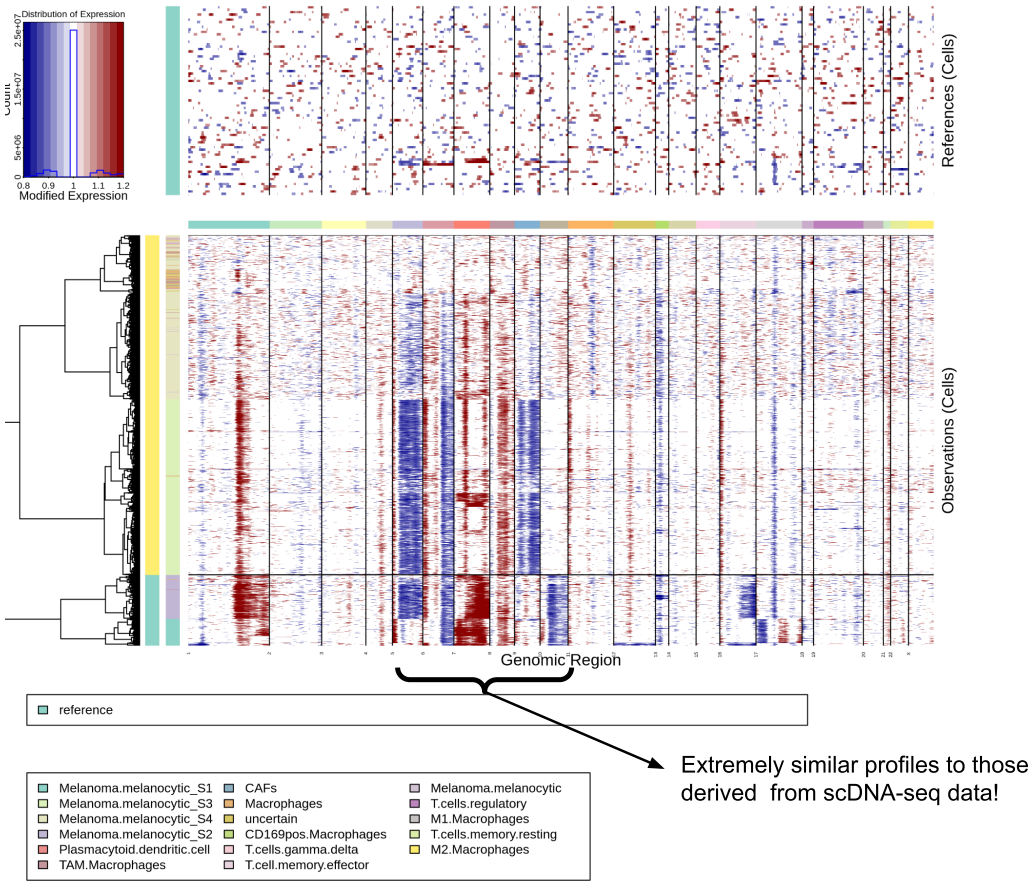

Figure A.3: Copy number profiles as inferred by InferCNV from scRNA-seq data. Subclonal structures and cell types have been annotated (left of heatmap). Blue entries correspond to copy number losses, red entries signalise copy number gains in the region. For comparison with CNA-profiles derived from scDNA-seq data, find the file `MALYLEJ-T_scD_Ar1v1.12__cluster_tree_sorted_cnvs_bins.png`, which is available in the TuPro study dataset.

	Pigmented	Invasive	SMC	NCSC	Baseline
MACEGEJ_Pearson	0.9075	0.6100	0.6380	NA	0.8908
MACYHYS_Pearson	0.8488	0.6096	0.8193	0.5053	0.7605
MADEGOD_Pearson	0.8672	0.6245	0.9837	NA	0.5862
MADIBUG_Pearson	0.8883	0.8079	0.7073	0.6115	0.5050
MAHACEB_Pearson	0.7871	0.8590	0.6969	NA	0.7596
MAHEFOG_Pearson	0.7568	0.8325	0.9714	NA	0.7209
MAHOBAM_Pearson	0.8411	0.6319	0.7617	NA	0.6262
MAJOFIJ_Pearson	0.8385	0.9029	0.8135	NA	0.8060
MAKYGIW_Pearson	0.7952	0.5806	0.7823	NA	0.6531
MALYLEJ_Pearson	0.8332	0.9353	0.7043	NA	0.7764

Table A.1: Per-sample compactness values for a subset of cellassign annotations.

	Pigmented	Invasive	SMC	NCSC	Baseline
MACEGEJ	0.3793	0.6615	0.7985	NA	0.4451
MACYHYS	0.3385	0.7389	0.5770	0.4504	0.5017
MADEGOD	0.1586	0.1257	0.4133	NA	0.1587
MADIBUG	0.3093	0.4680	0.6028	0.5993	0.6020
MAHACEB	0.3773	0.4711	0.7896	NA	0.7383
MAHEFOG	0.5797	0.5399	0.7063	NA	0.3480
MAHOBAM	0.4684	0.6738	0.6277	NA	0.4128
MAJOFIJ	0.6358	0.6875	0.8801	NA	0.8174
MAKYGIW	0.4065	0.6880	0.6421	NA	0.5342
MALYLEJ	0.0635	0.3574	0.5913	NA	0.2502

Table A.2: Per-sample complexity values for a subset of cellassign annotations. Lower values imply a better separation between clusters.

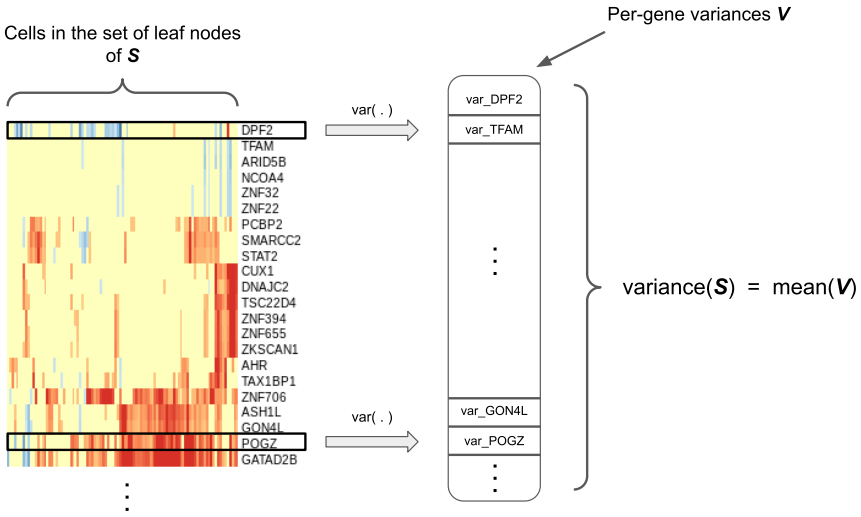

Figure A.4: Visualisation of the calculation of the intra-cluster variance for subclonal structure discovery. For each gene (rows of the CNA-Matrix), the variance over all cells within the cluster is calculated. Then the mean is calculated, yielding an average of per-gene variances.

Bibliography

[1] Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. 2020. "Cancer Statistics, 2020." CA: A Cancer Journal for Clinicians 70 (1): 7–30.

[2] Hanahan, D., and R. A. Weinberg. 2000. "The Hallmarks of Cancer." Cell 100 (1): 57–70.

[3] Rambow, Florian, Aljosja Rogiers, Oskar Marin-Bejar, Sara Aibar, Julia Femel, Michael Dewaele, Panagiotis Karras, et al. 2018. "Toward Minimal Residual Disease-Directed Therapy in Melanoma." Cell 174 (4): 843–855.

[4] Wouters, Jasper, Zeynep Kalender-Atak, Liesbeth Minnoye, Katina I. Spanier, Maxime De Waegeneer, Carmen Bravo González-Blas, David Mauduit, et al. 2020. "Robust Gene Expression Programs Underlie Recurrent Cell States and Phenotype Switching in Melanoma." Nature Cell Biology 22 (8): 986–998.

[5] Marusyk, Andriy, Vanessa Almendro, and Kornelia Polyak. 2012. "Intra-Tumour Heterogeneity: A Looking Glass for Cancer?" Nature Reviews. Cancer 12 (5): 323–334.

[6] Gaggioli, Cedric, and Erik Sahai. 2007. "Melanoma Invasion - Current Knowledge and Future Directions." Pigment Cell Research 20 (3): 161–172.

[7] Hoek, Keith S., Ossia M. Eichhoff, Natalie C. Schlegel, Udo Döbbeling, Nikita Kobert, Leo Schaerer, Silvio Hemmi, and Reinhard Dummer. 2008. "In Vivo Switching of Human Melanoma Cells between Proliferative and Invasive States." Cancer Research 68 (3): 650–656.

[8] Marin-Bejar, Oskar, Aljosja Rogiers, Michael Dewaele, Julia Femel, Panagiotis Karras, Joanna Pozniak, Greet Bervoets, et al. 2020. "A Neural Crest Stem Cell-like State Drives Nongenetic Resistance to Targeted Therapy in Melanoma." bioRxiv. https://doi.org/10.1101/2020.12.15.422929.

[9] McGranahan, Nicholas, and Charles Swanton. 2015. "Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution." Cancer Cell 27 (1): 15–26.

[10] Rübben, Albert, and Arturo Araujo. 2017. "Cancer Heterogeneity: Converting a Limitation into a Source of Biologic Information." Journal of Translational Medicine 15 (1): 190.

[11] Rampias, Theodoros. 2020. "Exploring the Eco-Evolutionary Dynamics of Tumor Subclones." Cancers 12 (11).

[12] Trapnell, Cole. 2015. "Defining Cell Types and States with Single-Cell Genomics." Genome Research 25 (10): 1491–1498.

[13] Irmisch, Anja, Ximena Bonilla, Stéphane Chevrier, Kjong-Van Lehmann, Franziska Singer, Nora C. Toussaint, Cinzia Esposito, et al. 2021. "The Tumor Profiler Study: Integrated, Multi-Omic, Functional Tumor Profiling for Clinical Decision Support." Cancer Cell 39 (3): 288–293.

[14] Shao, Xin, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu, and Xiaohui Fan. 2019. "Copy Number Variation Is Highly Correlated with Differential Gene Expression: A Pan-Cancer Study." BMC Medical Genetics 20 (1): 175.

[15] Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." Cell 177 (7): 1888–1902.

[16] L. Lun, Aaron T., Bach, Karsten, and John C. Marioni. 2016. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts." Genome Biol 17: 75-89.

[17] Torre, Eduardo, Hannah Dueck, Sydney Shaffer, Janko Gospocic, Rohit Gupte, Roberto Bonasio, Junhyong Kim, John Murray, and Arjun Raj. 2018. "Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH." Cell Systems 6 (2): 171–179.

[18] Abdelaal, Tamim, Lieke Michielsen, Davy Cats et al. 2019. "A comparison of automatic cell identification methods for single-cell RNA sequencing data." Genome Biol 20 (1): 194-219.

[19] Zhang, Allen W., Ciara O'Flanagan, Elizabeth A. Chavez, Jamie L. P. Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, et al. 2019. "Probabilistic Cell-Type Assignment of Single-Cell RNA-Seq for Tumor Microenvironment Profiling." Nature Methods 16 (10): 1007–1015.

[20] Zhang, Ze, Danni Luo, Xue Zhong, Jin Huk Choi, Yuanqing Ma, Stacy Wang, Elena Mahrt, et al. 2019. "SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples." Genes 10 (7): 531-548.

[21] Tickle T, Tirosh I, Georgescu C, Brown M, Haas B (2019). inferCNV of the Trinity CTAT Project. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. https://github.com/broadinstitute/inferCNV.

[22] Duan, Bin, Chenyu Zhu, Guohui Chuai, Chen Tang, Xiaohan Chen, Shaoqi Chen, Shaliu Fu, Gaoyang Li, and Qi Liu. 2020. "Learning for Single-Cell Assignment." Science Advances 6 (44).

[23] Sanchez-Taltavull, Daniel, Theodore J. Perkins, Noelle Dommann, Nicolas Melin, Adrian Keogh, Daniel Candinas, Deborah Stroka, and Guido Beldi. 2020. "Bayesian Correlation Is a Robust Gene Similarity Measure for Single-Cell RNA-Seq Data." NAR Genomics and Bioinformatics 2 (1).

[24] Dibaeinia, Payam, and Saurabh Sinha. 2020. "SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks." Cell Systems 11 (3): 252–271.

[25] Zappia, Luke, Phipson, Belinda, and Alicia Oshlack. 2017. "Splatter: simulation of single-cell RNA sequencing data." Genome Biol 18, 174.

[26] McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv. http://arxiv.org/abs/1802.03426.

[27] Paradis, Emmanuel, and Klaus Schliep. 2019. "Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." Bioinformatics 35 (3): 526–528.

[28] Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4.

[29] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

[30] Murtagh, Fionn, and Pierre Legendre. 2014. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" Journal of Classification 31 (3): 274–295.

[31] Pasquini, Giovanni, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp. 2021. "Automated Methods for Cell Type Annotation on scRNA-Seq Data." Computational and Structural Biotechnology Journal 19 (January): 961–969.

[32] Ross-Adams, H., A. D. Lamb, M. J. Dunning, S. Halim, J. Lindberg, C. M. Massie, L. A. Egevad, et al. 2015. "Integration of Copy Number and Transcriptomics Provides Risk Stratification in Prostate Cancer: A Discovery and Validation Cohort Study." EBioMedicine 2 (9): 1133–1144.

[33] Rand, William M. 1971. "Objective Criteria for the Evaluation of Clustering Methods." Journal of the American Statistical Association 66 (336): 846–850.

[34] Udart, Martin, Utikal, Jochen, Krähn, Gertraud M. and Ralf U. Peter. 2001. "Chromosome 7 Aneusomy. A Marker for Metastatic Melanoma? Expression of the Epidermal Growth Factor Receptor Gene and Chromosome 7 Aneusomy in Nevi, Primary Malignant Melanomas and Metastases." Neoplasia 3 (3): 245–254.

[35] Tirosh, Itay, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth 2nd, Daniel Treacy, John J. Trombetta, Asaf Rotem, et al. 2016. "Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq." Science 352 (6282): 189–196.

[36] Gerber, Tobias, Edith Willscher, Henry Loeffler-Wirth, Lydia Hopp, Dirk Schadendorf, Manfred Schartl, Ulf Anderegg, et al. 2017. "Mapping Heterogeneity in Patient-Derived Melanoma Cultures by Single-Cell RNA-Seq." Oncotarget 8 (1): 846–862.

ETH

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):

Supervised and Unsupervised Discovery of Intratumor Heterogeneity from Single cell RNA-seq Data

Authored by (in block letters):

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

Name(s):	**First name(s):**
Toma	Philip Erik David

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)**
Zürich, 19.04.2021	Philip Toma

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