NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

DAVID MIGUEL PRATAS PAIS

Bachelor in Computer Science

ASSESS THE EFFECT OF ANGIOGENESIS INHIBITION IN INTRA-TUMOR HETEROGENEITY

MASTER IN COMPUTER SCIENCE

NOVA University Lisbon
March, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

ASSESS THE EFFECT OF ANGIOGENESIS INHIBITION IN INTRA-TUMOR HETEROGENEITY

DAVID MIGUEL PRATAS PAIS

Bachelor in Computer Science

Adviser: Ludwig Krippahl
Assistant Professor, NOVA School of Science and Technology

Co-adviser: Daniel Sobral
Researcher, NOVA School of Science and Technology

Examination Committee:

Chair: Bernardo Toninho
Assistant Professor, NOVA School of Science and Technology

Rapporteur: Marta Lopes
Assistant Researcher, NOVA School of Science and Technology

Adviser: Daniel Sobral
Researcher, NOVA School of Science and Technology

MASTER IN COMPUTER SCIENCE

NOVA University Lisbon
March, 2022

Assess the effect of angiogenesis inhibition in intra-tumor heterogeneity

Acknowledgements

First, I would like to thank my advisers, professor Ludwig and researcher Daniel, for the support, opinions, reviews, suggestions, and dedication throughout the making of this dissertation.

Then, I would like to acknowledge the Computational Multi-Omics research lab into which I was integrated during this master thesis development. Thank you for making me feel part of the team. A special mention to the Fundação para a Ciência e a Tecnologia for providing me with research fellowship funding.

Finally, I would like to express the most gratitude to my parents, my best friends, that have never let me down. In particular, honor my two grandmothers, who are still with me, and my two grandfathers, whom I did not get the chance to meet but who I know are protecting me from above. My appreciation to the rest of my family. To my university godfathers, Tiago and Inês, thank you for the unforgettable moments of friendship. To my university colleagues, especially João and Carlos, thank you for the endless fun hours working together. A kind word to all my friends that were always rooting for me.

Abstract

The genetic diversity of the populations that arise within a tumor following cancer progression is known as intra-tumor heterogeneity. Angiogenesis is the formation of new blood vessels from the existing vascular network. It is involved in different physiological processes, including wound healing. Tumors induce the excessive production of pro-angiogenic factors that promote the proliferation and variety of cancer cells. The inhibitors of tumor angiogenesis were initially designed to destroy the tumor blood vessels, causing the death of cancer cells. However, they have been associated with selecting therapy-resistant cells, thus leading to treatment failure. This thesis aims to study the effect of angiogenesis inhibition in intra-tumor heterogeneity based on an experiment in which a breast cancer cell line, designated as MDA-MB-231, was injected into four mice. Two of them were administered with sunitinib, an anti-angiogenic drug. In particular, this thesis investigates whether the two groups of mice, control and treatment, are distinct. Two variables were analyzed to distinguish between the mice: the intra-tumor heterogeneity and the mutational profiles of the genes. Three different heterogeneity estimation methods were chosen: the tumor heterogeneity index, PyClone-VI, and Canopy. These worked with the sequencing data of the mice tumor biopsies, specifically with the somatic mutations and copy number alterations. Dimensionality reduction techniques were applied to extract information from several genes. These relied not only on the mice samples but also on the tumor data of patients stored in The Cancer Genome Atlas, which allowed access to more examples. None of the methods could identify a clear difference between the two groups of mice. Their intra-tumor heterogeneity values were similar, and the mutational profiles of their genes appeared to follow the same pattern. Considering these results, we can assume that destroying the tumor blood vessels of the mice from the treatment group did not drive the diversification of cancer cells. Nonetheless, further research should be conducted to confirm this conclusion. For example, test the latest heterogeneity estimation methods and explore the capabilities of neural networks.

Keywords: intra-tumor heterogeneity, angiogenesis, anti-angiogenic therapy, clonal reconstruction, machine learning

Resumo

A diversidade genética das populações que surgem dentro de um tumor após a progressão do cancro é conhecida como heterogeneidade intra-tumoral. A angiogénese é a formação de novos vasos sanguíneos a partir da rede vascular existente. Está envolvida em diferentes processos fisiológicos, incluindo a cicatrização de feridas. Os tumores induzem a produção excessiva de factores pró-angiogénicos que promovem a proliferação e variedade de células cancerígenas. Os inibidores da angiogénese tumoral foram inicialmente concebidos para destruir os vasos sanguíneos tumorais, causando a morte de células cancerígenas. No entanto, têm sido associados à selecção de células resistentes à terapia, levando assim ao fracasso do tratamento. Esta tese visa estudar o efeito da inibição da angiogénese na heterogeneidade intra-tumoral, com base numa experiência em que uma linha celular de cancro da mama, designada por MDA-MB-231, foi injectada em quatro ratos. Dois deles foram administrados com sunitinib, um medicamento anti-angiogénico. Em particular, esta tese investiga se os dois grupos de ratos, controlo e tratamento, são distintos. Duas variáveis foram analisadas para distinguir entre os ratos: a heterogeneidade intra-tumoral e os perfis mutacionais dos genes. Foram escolhidos três métodos diferentes de estimativa da heterogeneidade: o índice de heterogeneidade tumoral, o PyClone-VI, e o Canopy. Estes trabalharam com os dados de sequenciação das biópsias tumorais dos ratos, especificamente com as mutações somáticas e as alterações do número de cópias. Técnicas de redução da dimensionalidade foram aplicadas para extrair informação de vários genes. Estas basearam-se não só nas amostras dos ratos mas também nos dados tumorais de pacientes armazenados no Atlas do Genoma do Cancro, o que permitiu o acesso a mais exemplos. Nenhum dos métodos conseguiu identificar uma diferença clara entre os dois grupos de ratos. Os seus valores de heterogeneidade intra-tumoral eram semelhantes, e os perfis mutacionais dos seus genes pareciam seguir o mesmo padrão. Considerando estes resultados, podemos assumir que a destruição dos vasos sanguíneos tumorais dos ratos do grupo de tratamento não impulsionou a diversificação das células cancerígenas. No entanto, devem ser realizadas mais pesquisas para confirmar esta conclusão. Por exemplo, testar os métodos de estimativa da heterogeneidade mais recentes e explorar as capacidades de redes neuronais.

Palavras-chave: heterogeneidade intra-tumoral, angiogénese, terapia anti-angiogénica, reconstrução clonal, aprendizagem automática

Contents

List of Figures

List of Tables

Acronyms

BAF	B-allele frequency
CCF	cancer cell fraction
CNA	copy number alteration
CP	cellular prevalence
Isomap	isometric feature mapping
ITH	intra-tumor heterogeneity
MDS	multidimensional scaling
PCA	principal components analysis
SNV	single nucleotide variant
SVD	singular value decomposition
t-SNE	t-distributed stochastic neighbor embedding
TCGA	The Cancer Genome Atlas
VAF	variant allele frequency
WES	whole-exome sequencing

Chapter 1

Introduction

Cancer is one of the leading causes of death worldwide, with millions of deaths each year [1]. Furthermore, a rise of 70% in the number of new cases is expected over the coming two decades [2]. To stand up to this threat, the investigation within the oncology field has been an ongoing global priority with substantial investment [3].

Currently, up to one in two people are diagnosed with a tumor during their lifetime [3], which is related to an increasing prevalence of risk factors for cancer, such as smoking, lack of exercise, and nutrient imbalance [4].

Decades of research have significantly expanded our understanding of the disease, with improved outcomes for many patients [5]. However, despite all the progress, the battle against cancer still faces several challenges, especially the low efficacy of the traditionally used procedures [6].

Cancer treatment success depends on an early diagnosis and evidence-based choices among the available methods. These include surgical intervention, radiation therapy, and chemotherapeutic drugs, frequently limited by the toxic side effects caused to healthy cells [7, 8].

For this reason, treatment alternatives with fewer harmful effects were developed. In particular, knowing the biological mechanisms that drive cancer progression led to new anti-cancer therapies. These interfere with molecules that promote tumor growth [9], targeting cancer cells but sparing normal cells [10].

On top of this, there is growing evidence that they play an essential role in influencing immune system responses. By acting on different cell types within the tumor microenvironment, they contribute to anti-tumor immunity [11]. Because of this, they can be repurposed to treat other diseases [12].

One limitation affecting the targeted and standard methods has been the development of drug resistance [9]. Moreover, recent studies show that exposure to targeted therapies

often induces genomic alterations that explain cancer progression and the appearance of recurrences post-treatment [13].

A focal point of the resistance to therapy is the dynamic genetic diversity within cancer [14], known as intra-tumor heterogeneity (ITH). Therefore, it is fundamental to estimate it with high accuracy to improve cancer treatment [15]. In the past few years, many computational tools have been proposed to do so [16].

Technological developments have been crucial in cancer research progress, namely *DNA sequencing*. This technology timely determines the order of the nucleotide bases in biological samples, allowing the recognition of genomic modifications responsible for tumor formation [1].

Such advances prompted an unprecedented influx of data [5, 17], giving rise to an explosion of new genomic datasets, which allowed the creation of an ever-expanding catalog of genomic variants in the human population [18, 19].

For instance, the public-funded project The Cancer Genome Atlas (TCGA) has made available genomic data on over 11000 individuals, mapping the genomic changes that occur in more than 30 types of human cancer [19, 20]. This large-scale cancer genomics program has generated petabytes of data [20].

This exponential growth has urged the application of a range of *machine learning* techniques to analyze complex biological data efficiently. Machine learning is a discipline that studies how computers can simulate the human cognition process. This reproduction works by identifying patterns in the data and error-based experience, adjusting the performance of the learning tasks [21].

In unsupervised machine learning, the algorithms detect similarities and differences between data points. The system makes decisions without being trained by a dataset, searching for naturally occurring patterns or groups within the data. As a result, it is possible to understand hidden relationships or detect abnormalities that humans overlook when observing multiple data [22].

Unsupervised learning is becoming an indispensable tool for exploring increasingly large amounts of data [23]. The application of *dimensionality reduction* methods enables the extraction of valuable information from these data [22]. It can help identify significant variables in high-dimensional genomic datasets of cancer [24].

Dimensionality reduction algorithms can capture the structure of the data manifold, which is embedded in a higher dimension. These can preserve the meaningful information of the original dataset without structural content loss, using a lower-dimensional representation of its features that is easier to interpret [23].

Motivated by the contribution of estimating ITH for cancer treatment and the efficiency of dimensionality reduction for deconstructing large datasets, this thesis comprised a work divided into two parts.

Based on the case study described in section 1.2, tumor genome sequencing data of mice were used as input for both stages. For the second part, the tumor genome sequencing data of TCGA patients supplemented the datasets fed to the algorithms.

Three types of ITH estimation methods were tested in the first part. The second phase covered the execution of a set of dimensionality reduction algorithms. These two tasks aimed to distinguish between the two groups of mice from the case study by assessing heterogeneity values and other relevant features.

The objectives of the work of this dissertation, including the outline of the following chapters, are listed in section 1.3. Next, section 1.1 presents the background knowledge of the thesis subject, separated into four subsections.

Subsection 1.1.1 points out the progressive role of the next-generation sequencing technology in understanding cancer evolution. Afterward, subsection 1.1.2 addresses the characterization of ITH and the variables that influence it. Then, subsection 1.1.3 focuses on angiogenesis and its appropriation by tumors. Lastly, subsection 1.1.4 refers to angiogenesis inhibition therapies.

1.1 Background knowledge

Stored in chromosomes and encoded by DNA nucleotide bases chunked in *genes*, the *genome* defines the instructions for the function of the cells in our body. Usually, we have 23 chromosome pairs, including the sex chromosomes [25].

Human cells are *diploid* organisms since each chromosome pair has two copies of a gene, denominated by *alleles*, one inherited from each parent [25]. The location of a gene on a chromosome is designated as its *locus*. The condition of having two different gene versions at a locus, a reference and mutant alleles, is defined as *heterozygosity*.

Mutations are alterations to the genome. A *heterozygous* mutation affects only one copy of a chromosome pair, whereas a *homozygous* variant is present in both. Typically, most of the new mutations that arise are heterozygous [26].

The ones inherited from our parents are identified as *germline* mutations and are present in all cells. Variants that occur during the lifetime of an individual are denoted by *somatic* mutations and only exist in a subset of cells. Generally, they are responsible for tumor development, although germline mutations can also promote it [26].

The abnormal cell multiplication and survival, together with new mutations occurring, are the foundations of cancer. In each cell division, the DNA is copied and passed on to its children. However, this process is not error-free, given that variants arise during DNA replication [27].

Each new mutation gained by every cell in a tumor, if not repaired, will be forwarded to its descendants, thereby recording the ancestor-descendant relationships between cells [27]. Essentially, each cancer cell genome is an imperfect copy of the genome of another tumor cell that existed in the past [27].

This way, because of the successive accumulation of mutational events, the differences between the genetic material of individual cancer cells contain a historical record of the tumor evolution [27].

It is commonly assumed that tumor growth proceeds according to the *clonal evolution* theory. This assumption argues that most cancers evolve from a normal (non-cancerous) single founder cell through the continuous acquisition of somatic mutations. These allow the tumor to grow separately from its surrounding normal tissue [28, 29].

Tumor expansion is not triggered by all modifications. Some have a neutral effect, being labeled as *passenger* mutations [30]. The ones that provide the cell with a selective advantage that makes it more fitted to its micro-environment are referred to as *driver* variants [30]. They promote tumor maturation, resulting in higher cell proliferation and lower apoptosis (cell death) [31].

The spreading of tumor cells that culminates in the seeding of a new population is called *clonal expansion* [15]. The generated population is also known as a *clone* [32]. A clone is a group of cells that descend from a common ancestor. These have a similar genetic constitution and share a unique set of mutations that distinguishes them from other populations [28].

The random incidence of mutational events in individual tumor cells has become a platform for adapting and selecting the tumor fittest clones [11, 14]. Clones gaining advantage will expand, whereas the less fitted ones will be superseded and may eventually become extinct [11].

These clonal advantages may differ in time and space, as different tumor regions may meet the requirements for expansion in different periods. The tumor micro-environment conditions influence the strategy used to select the clones. While specific zones may opt for the clones that better tolerate the lack of oxygen and nutrients, other more nutrient-dense areas may choose the clonal populations that grow faster [11].

This natural selection process carries on throughout the cancer lifetime, inducing a set of clonal expansions that yield genetically distinct subclones inside the tumor [15, 29]. ITH refers to this diversity of the cell subpopulations within a tumor, raised by clonal evolution [33].

1.1.1 Next-generation sequencing-based cancer evolution study

Tumors are composed of evolving heterogeneous groups of cells. The permanent diversification of these subclonal populations, allowing them to adapt to therapy, is an obstacle to the positive outcome of cancer treatment. Hence, it is critical reconstructing the clonal history to characterize the heterogeneity level of tumors [15].

Even though tumor evolution can not be directly observed, we can recreate it [31]. The genetic variations inherited by tumor cells during cell division encode their ancestries. By detecting the mutations present in the cells, we can reproduce the evolutionary paths that led to the different clones and infer the tumor evolution history [27].

Two types of genomic events have been considered when performing tumor subclonal reconstruction and assessing heterogeneity: single nucleotide variant (SNV) and copy number alteration (CNA) [28, 34].

A SNV, also designated as a small somatic mutation, is an abnormality of a nucleotide in a specific genome position or a short genomic insertion or deletion [30, 34]. A CNA is characterized by the rearrangement of chromosomes, considering gains or losses of large segments of the genome, specifically duplication or deletion [30].

The latest research in ITH involves computational methods that use SNVs and CNAs measured through *next-generation sequencing*, also referred to as high-throughput sequencing or massively parallel sequencing [16]. This technology generates sequences of DNA nucleotide base pairs with a short fixed-length from the tumor genome, termed *reads* [30], based on a specific tumor and genome sampling approaches [35].

Tumor sampling can be carried out via a single or multi-region *biopsy* (removing a piece of the tumor mass to be examined under a microscope). This fragment comprises populations with not only cancer cells but also normal cells [16]. Single and multiple samples can be pooled as a bulk specimen, sampling hundreds of thousands to millions of tumor cells, or disaggregated into single cells, typically hundreds [27, 35].

A single tumor sample composed of bulk cells is limited, representing only a particular spatiotemporal snapshot of the tumor content. It can be used to depict the tumor clonal evolution, but it may fail to identify early clonal events [14]. As a result, single-region sequencing may underestimate the number of subclonal populations [28].

Multi-region sequencing of spatially separated tumor zones can help mitigate this to some extent, providing a more precise reconstruction of the tumor subclonal architecture [14, 31]. From a practical perspective, it would be challenging to scale up the application of multiple sampling to hundreds of tumors [15, 31].

Subclonal reconstruction from single-cell profiling has revealed an incomparable *resolution* (measurement accuracy) of the genetic diversity among cancer cells [27]. Unlike bulk sequencing, it can profile the genome of every single cell in a tumor [26]. But, despite that, single-cell sequencing still has a high cost and technical limitations that come from the low amount of DNA extracted from single cells [15, 27].

Selecting the quantity of genome material from the sampled tumor to be sequenced relies on three methodologies, including sequencing targeted panels of a few hundred genes. This option has some disadvantages, such as depending on the genes selected for the panel and being unable to recognize smaller scale CNAs, which exist only in a subset of all tumor cells [30, 35].

The other two alternatives are to sequence the entire genome or to perform whole-exome sequencing (WES). Both have been extensively used to explore ITH from multi-region tumor biopsies [35].

For subclonal inference, whole-genome sequencing presents clear advantages over WES, as it identifies nearly two orders of magnitude more mutations than the latter [30]. However, its low average *sequencing depth* is a limiting factor of the subclonal reconstruction data resolution [27, 30].

Given a certain number of reads with a specific length and assuming that they are distributed evenly across an idealized genome, the theoretical sequencing depth (also

cited as coverage or depth of coverage) is the average number of times each locus in the target region is expected to be sampled [25, 36]. For instance, a 5× sequencing depth denotes that each locus is expected to be sampled an average of 5 times [28].

Sequencing at a higher coverage is recommended in cancer research, given that some relevant mutations are only present in a small fraction of tumor cells, being undetectable at a lower depth [25].

The exons, or collectively the exome, carry the protocol for protein production. The exome has been the focus of the latest cancer studies, given that it is much more intuitive to assign functional relevance to mutations identified in protein-coding regions. These represent 1% of the genome [25].

On this account, by applying WES, with an equal cost, we can sequence at a much greater depth than targeting the entire genome. Sequencing the whole genome area with the same coverage used in WES requires many more reads, and thus the sequencing cost is much more expensive [25].

On the other hand, the coverage from WES is much noisier than the coverage from sequencing the entire genome [25]. This noise comes from the bias in a step of the experiment where the exons to be sequenced get selected [25]. A superior sample sequencing depth increases the confidence and resolution of the clonal populations identified and possibly overcomes the technical noise limitations found at lower coverage [27].

Again, higher sequencing coverage implies higher sequencing costs. The quality of the tumor evolution reconstitution can be improved considering this trade-off. Sequencing more samples with sufficient coverage for precise mutation detection is better than sequencing fewer samples at a greater sequencing depth [28].

1.1.2 Intra-tumor heterogeneity estimation

Next-generation sequencing has offered new insights into the evolution and internal heterogeneity of cancers [15]. The subclonal composition of a tumor can be inferred by the analysis of highly parallel sequencing data [30].

The subclonal reconstruction of tumors to estimate ITH involves three key aspects. First, it differentiates the dominant tumor cell populations by detecting the different sets of SNVs that each one shares. Second, it quantifies the proportion of cells that belong to each clone. Third, it rebuilds the phylogenetic paths by which separate clones evolved from their common ancestor [28].

To identify somatic SNVs and distinguish them from germline, normal (with only healthy cells) and cancer samples need to be sequenced and compared [26]. Low-quality reads are removed, as they may contain sequencing errors [26].

After variant filtering, the sequencing reads of both samples are mapped, or aligned, to a reference genome, an idealized template featuring the current knowledge about human genomics, assembled from the genome of multiple people [25, 26]. Only the variants that are unique to the tumor sample are kept [26].

Based on the mapping, a bioinformatic tool for variant calling (to detect mutations from sequencing data) can derive the number of reads that support the reference and alternative alleles of a SNV i (respectively, $r_{ref,i}$ and $r_{mut,i}$) [26, 30].

The frequency of a mutation in sequencing data [35] or the fraction of reads identified with a variant among all the reads mapped to a specific location of the genome is known as variant allele frequency (VAF) [16]. The fraction of reads mutated by a given SNV i (VAF_i) can be calculated as [30]

$$\text{VAF}_i = \frac{r_{mut,i}}{r_{mut,i} + r_{ref,i}}. \tag{1.1}$$

For example, the VAF of a heterozygous SNV in a diploid copy region is 0.5. Since it affects only one chromosome copy, it will only be identified in half of the sequencing reads [30]. Mutations that occur in all cancer cells should be detected in more sequencing reads than variants that only appear in specific subclonal cell groups [15].

Different methods have been developed for variant calling from tumor-normal sequencing pairs. These tools look for greater sensitivity to mutations with a lower frequency in cancer cell populations [31].

Besides the sequencing depth, the VAF varies according to another three variables. These are the copy number variation of the gene alleles, the ratio of cancerous cells in a tumor sample, and the proportion of tumor cells with a specific mutation [30].

CNAs may transform normal diploid copy regions into non-diploid [28]. Variants from the same population can have different VAFs because of copy alterations [30]. CNAs arising in a tumor can affect one or both the inherited copies. The accurate estimate of the number of copies of each allele, typically called the *allele-specific copy number*, represents a complex problem in the reconstitution of tumor evolution [25].

Not all CNA calling methods can calculate the *minor* and *major* copy numbers, respectively, the number of copies of the least and most frequent alleles [31]. Alternatively, they provide a genome-wide evaluation of the copy numbers sum of the two inherited alleles, also known as *total copy number* [25].

A copy number gain duplicates a genome segment, increasing both its allele-specific and total copy numbers [25, 26]. A deletion can remove either a portion of or a complete chromosome, decreasing both copy number types [25, 26].

The time order in which SNVs and CNAs appear can also change the values of the allele-specific copy numbers [30]. When a CNA gain event strikes an allele previously mutated by a heterozygous SNV, the number of chromosome copies with the variant is duplicated. In case a heterozygous SNV impacts an allele after a CNA gain, the mutation will only occur in a single chromosome copy [30].

Allele-specific copy number estimation is even more relevant in recognizing *loss of heterozygosity* phenomena, which defines the complete deletion of a parental copy. These may, or not, change the total copy number [25].

Some CNAs cause a copy-neutral loss of heterozygosity. For instance, the simultaneous loss of a parental copy and a gain on the other copy of the chromosome pair changes the allele-specific copy numbers, not affecting the total copy number [25, 26].

The identification of CNAs can be based on imbalances in the number of maternal and paternal alleles in germline heterozygous SNVs [26, 30]. Without knowing which allele is the parental and which is the maternal, the reference and alternative alleles are, respectively, denoted by *allele A* and *allele B* [26].

The allelic ratio in the tumor sample relative to the normal sample is known as B-allele frequency (BAF). Considering that the total sequencing reads associated to allele A and allele B are, respectively, $r_{A,j}$ and $r_{B,j}$, the mutant allele frequency of a germline heterozygous SNV j (BAF_j) can be quantified as [30]

$$\text{BAF}_j = \frac{r_{B,j}}{r_{A,j} + r_{B,j}}. \tag{1.2}$$

A germline heterozygous SNV has a BAF of approximately 0.5 in the absence of any CNAs, as only one allele gets mutated [30]. Hence, CNAs can be spotted by absolute value shifts relatively to the expected heterozygous BAF of 0.5 [30].

CNAs can also be inferred from sequencing data by comparing the local read depths of a genome region in a tumor sample, represented by log_2 [28], and in a normal sample bearing germline heterozygous SNVs [37]. These metrics, including the BAF, are used by CNA calling tools and subclonal reconstruction algorithms that infer CNAs [28] (even though only a limited number does so [26]).

Most subclonal reconstruction methods use the VAF of SNVs to approximate the fraction of sampled cells mutated by each variant, accounting for the influence of CNAs. The cellular prevalence (CP) of a SNV is the ratio of all cells, normal and cancerous, from a sequenced tissue carrying the mutation [28].

The algorithms then cluster the SNVs with similar CP values. To distinguish each clone, they consider that a specific set of mutations is only shared within a population. The frequency or proportion of cells from each population at the time a sample is taken is called *clonal frequency*, population frequency or clonal composition [16, 28]. The sum of all population frequencies is equal to 1 per sample [26].

The VAF is also influenced by the *tumor purity*, also named purity, sample purity, or cellularity, which is the fraction of cancer cells in a tumor sample [30]. In turn, the percentage of non-cancer cells is equal to $1 - purity$ [28]. Some SNV and CNA calling methods also produce an estimate of the tumor purity [15].

A lower purity corresponds to a lower number of mutated cells, which leads to sequencing fewer variant reads [30, 33]. The purity decreases when the analyzed specimen has non-cancer cells that come from the normal tissues surrounding the tumor [33].

The third attribute that affects the VAF is the cancer cell fraction (CCF) [30]. Based on the fraction of sampled cells carrying a SNV k (CP_k) and the purity, the ratio of cancer

cells with the mutation (CCF_k) can be measured as [28]

$$CCF_k = \frac{CP_k}{\text{purity}}. \tag{1.3}$$

SNVs can also be clustered by their CCF values. A group of variants present in all tumor cells of a sample has a CCF equal to 1, whereas a mutation set found just in some clones has a CCF value of less than 1 [28].

Figure 1.1 represents an overview of a tumor subclonal reconstruction process to characterize the ITH. Through single-cell sequencing, it is possible to analyze the content of and the association between all single cells whose mutations have been sequenced [16]. With the VAFs estimated from bulk sequencing reads, we can infer the clonal composition and the evolutionary relationships between the tumor cell populations.

Some computational methods aim to infer the evolutionary relationships among clones, their phylogeny, also called *phylogenetic tree*, based on the clusters CP or CCF values and shared sets of mutations [28]. The phylogenetic trees are often built upon rules that constrain the hierarchical ordering of the populations in a tumor [30].

Namely, the *infinite sites assumption*. This rule asserts that each mutation occurs only once during the tumor lifetime [28, 30]. In other words, the same chromosome genome position is not mutated twice in the course of the tumor evolution, and once mutated, it can never revert to a normal state [16, 28]. Without this assumption, the subclonal reconstruction problem would become computationally intractable [16].

Another important criterion is the *pigeonhole principle*, establishing that the sum of the CP or CCF of the descendant populations can not exceed the values of the ones of their ancestor [28, 30]. The inferred CCFs of the tumor populations from bulk sequencing in figure 1.1 exemplify this.

There are 2 normal cells (empty circles) and 14 tumor cells with mutations. Three populations arose throughout tumor evolution: red, yellow, and blue. Each set of mutations is part of, respectively, all, half, and 4 of the tumor cells, with inferred CCF values of, in percentage, 100%, 50%, and approximately 28.5%.

The red mutation set is the first ancestral population, as it is inherited by all cells and has the higher CCF value. As $100\% + 50\% > 100\%$, the yellow clone can only descend from the first population. The blue population is also a successor of the red clone, and sibling of the yellow, as $100\% + 28.5\% > 100\%$ and $50\% + 28.5\% < 100\%$.

Alternatively, considering that the second and third clones had CCF values of 60% and 50%, the second population would be a direct descendant of the first, and the third population a successor of the second, as $60\% + 50\% > 100\%$ and $60\% > 50\%$.

The clonal composition and evolution can also be annotated with the frequency of the populations, in percentage, as illustrated by figure 1.1. Normal cells grew into the tumor population in red, which further developed into the yellow and blue clones.

These expanded through selective advantages and became, respectively, 50% and 29%

of all tumor cells. In this example, the normal population frequency is not included. As a consequence of natural selection and the supremacy of its descendants, only 21% of the tumor cells from the ancestor population remained at the time of tumor sampling.

Not all the populations that ever existed in the tumor need to be present when sampling. A *vestigial* population has a clonal frequency of 0 after being extinguished by other clones, with none of the cells with this specific group of shared mutations remaining at the time of tumor sampling [26]. However, we can infer that it existed based on the presence of this subset of mutations in all its observable descendants [26].

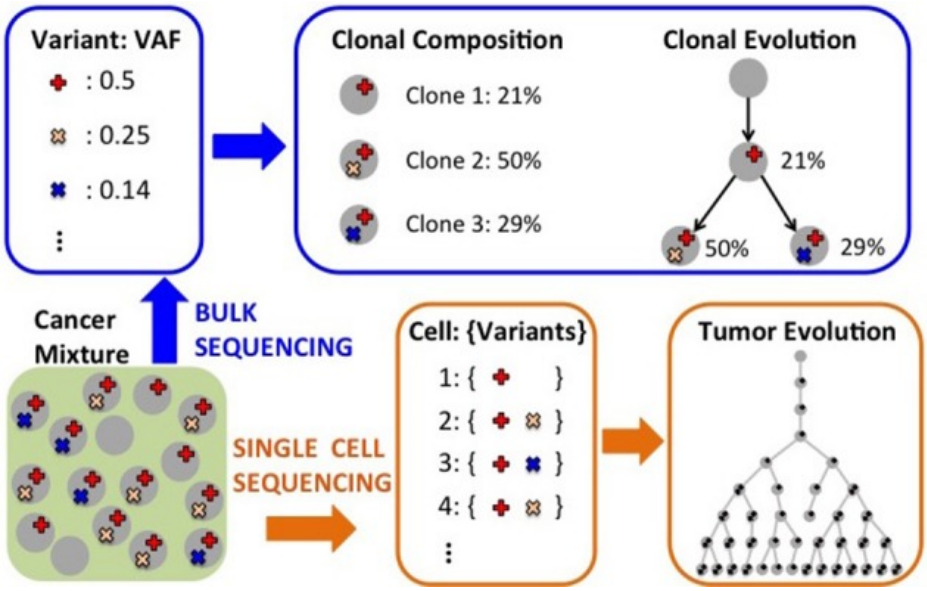

Figure 1.1: Subclonal reconstruction process based on next-generation sequencing [16].

1.1.3 Tumor angiogenesis

Cancer is a frightfully adaptable and misleading disease, which, by exploiting different biological mechanisms, disrupts normal physiologic and cellular routines, evading the immune system [5]. One process appropriated by tumors is *angiogenesis* [38].

Through blood circulation in vessels, the cardiovascular, vascular, or circulatory system assures the transporting of nutrients and gases, removing metabolic waste products to or from tissue cells of every part of our body. It also aids immune system surveillance and maintenance [38–40].

Blood vessels are typically ordered tubular networks, branched in detail [41]. They can be hierarchically divided into three types: arteries, veins, and capillaries. These have different morphologies reflecting their functions. Specifically, capillaries participate in the maturation and consolidation of vessels [40].

The formation of blood vessels is mainly achieved via *vasculogenesis* and angiogenesis [38]. Vasculogenesis refers to the primitive creation of blood vessels and the subsequent systematization of the different vascular architecture types [38, 40]. Although both

processes play an indispensable role in the early stages of vascular network development, new vessels essentially arise from angiogenesis during adulthood [42].

Angiogenesis refers to the sprouting of capillaries from pre-existing vessels, which are progressively renovated [40, 42]. It is involved in several pathological conditions, including bone repair and tissue regeneration for wound healing [38, 42].

This mechanism is controlled by the balance between positive and negative activity regulation molecules. These pro- and anti-angiogenic factors are, respectively, inducers and inhibitors of angiogenesis [42]. The *angiogenic switch* denotes the increased production of one or more positive factors that intensifies its action [42].

Given that blood vessels nurture most tissues, a vasculature in good conditions is necessary to guarantee their correct functioning [42]. In adults, angiogenesis acts in response to the metabolic requirements of tissues. In particular, it is triggered by the lack of cell oxygen [42].

Moreover, considering the influence of the vascular system on every organ of our body, discrepancies in natural vessel growth, either insufficient or abnormal, contribute to many diseases [39].

Deficient vessel formation can give rise to myocardial infarction and stroke. On the other hand, excessive maturation promotes inflammatory disorders and a faster expansion of malignant tissues, that is, tumor progression [39, 42].

The shape of tumor blood vessels is highly anomalous. The tumor vascular network is tortuous, dilated, and disorganized, in contrast to the consistent structure of the normal vasculature [39, 40]. This vascular immaturity leads to a higher tissue permeability, which reduces not only the oxygen in the tumor micro-environment but also the diffusion of drugs, facilitating the entry of cancer cells into the bloodstream [38, 40].

The changing conditions in the tumor micro-environment constantly put the cells under pressure [11]. The absence of oxygen prompts stress responses in cancer cells, and only the fittest ones survive. This deprivation favors the selection of more invasive and aggressive clones, enriching ITH [40].

Tumors are restricted to microscopic size and can not expand without a vascular supply. They need to rapidly develop new vessels to support the elevated proliferation rate of cancer cells [40].

Angiogenesis is mainly initiated by the tumor itself [38]. Stimulated by the shortage of oxygen, tumor cells secrete high levels of angiogenesis positive regulation molecules [40]. These excessively induce the angiogenic switch, culminating in the creation of an abnormal vascular network that provides the means to the tumor spreading [40].

In normal angiogenesis, the negative regulation molecules stabilize the positive ones after the creation of vessels is completed [41]. In *tumor angiogenesis*, cancer cells persistently produce pro-angiogenic factors to respond to the continuous tumor metabolic requirements resulting from its inconsistent vessel structure [38, 41].

Tumor angiogenesis is the process by which blood vessels infiltrate and grow inside the tumor micro-environment [41]. They supply the oxygen and nutrients necessary for

the malignant cells to proliferate. Hence, it is widely acknowledged as one hallmark of cancer, being significantly responsible for tumor growth [41].

The tumor outgrowth and dissemination of malignant cells form metastases. Similar to cancer, *metastasis* is an evolutionary process in which one or more clonal populations acquire selective advantages that make them detach from the *primary tumor*, which arose first, and seed a new tumor at a secondary site [33].

Most cancer studies have been focused on primary tumors, whereas fewer have examined tumor metastases. Such can be explained by the challenge of collecting biopsies from distant metastatic sites, which may sometimes be inaccessible [31].

1.1.4 Anti-angiogenic cancer therapies

The discovery of tumor angiogenesis ushered in a novel alternative to fight cancer. After observing that tumors need a vascular supply to grow, it was first proposed that vessel formation inhibition could suppress their development. A treatment approach identified as *anti-angiogenic therapy* materialized from this theory [43].

This type of treatment relied on the concept of blocking the creation of new vessels and eradicating the existing ones, depriving cancer cells of oxygen and nutrients. This supply cut would drive the tumor cells to starve to death or become quiescent [38, 40, 43]. Not only would it affect the primary tumor but also prevent cancer cells from escaping through the circulatory system, thus avoiding the seeding of metastases [38].

Anti-angiogenic agents emerged as a class of drugs that aimed to disrupt tumor vascularization, directed to regulating specific molecules responsible for tumor progression and promotion of metastases [38]. Furthermore, they were expected to target multiple angiogenesis triggers and have limited toxic side effects [38].

Currently, there are some drugs for angiogenesis inhibition that have been tested in clinical trials and approved to treat different cancers [43]. Most of these are small molecules and developed antibodies that are focused on angiogenesis activation factors, blocking their activity [44].

Among these, *sunitinib* is a small-molecule inhibitor that reduces tumor vascularization and leads to the apoptosis of cancer cells. It has been applied to renal cell carcinoma, a type of kidney cancer, and gastrointestinal stromal tumor [38, 44].

These clinically approved anti-angiogenic drugs showed great potential and considerable practical effectiveness in reducing tumor growth. However, they could not erase it as a stand-alone strategy nor act on different pro-angiogenic factors at once [38]. Their application had few noticeable benefits, with only moderate responses [43].

Along with that, they are associated with toxic side effects. Some of the usual adverse reactions include hypertension, thrombosis, and skin toxicities [44]. Sunitinib itself has been appointed as the cause of fatigue and thyroid issues [38].

The use of anti-angiogenic drugs was only effective when combined with conventional treatment options, for example, chemotherapy. Their co-adjuvant application showed

significant improvements compared to their separate administration [43].

Knowing that the efficacy of chemotherapy depends on the delivery of its agents to cancer cells through efficient vascular structures, these were counter-intuitive results [43]. According to the theory, anti-angiogenic therapy should eradicate blood vessels, which would prevent drug delivery and inactivate chemotherapy [43].

The notion of *vessel normalization* was then introduced in opposition to the tumor vessel abolition that exacerbates the production of pro-angiogenic factors and neutralizes other treatments [43]. It seeks to control and normalize the tumor vasculature, being compatible with the combination of anti-angiogenic therapies with other treatment modalities. This way, these can target multiple angiogenesis enhancers at once [45].

Tumor vessel normalization benefits the delivery of drugs to the tumor and their distribution, affecting a higher fraction of tumor cells. Besides, it can supply more oxygen and nutrients to cancer, which will ease the harsh conditions of the tumor micro-environment and allow the anticancer immune cells to work better [45].

Recently, there have been advancements in nanotechnology for cancer treatment. Nanomaterials have unique size and shape properties [38]. The use of nanoparticles provides a rigorous drug diffusion into specific locations of the tumor micro-environment, having a higher therapeutic efficacy by reducing toxic side effects [38].

Despite all the progress in cancer research and therapy, and considering that some of the available treatments can eliminate most cells of a tumor, many advanced cancers remain incurable [46]. Cancer is a clonal evolutionary process characterized by heterogeneity, which difficulties its effective control. Two primary strategies may explain the resistance of tumor cells to therapy [31].

The chance of cancer cells surviving increases if, previous to the treatment, there is ITH [46]. In addition, eliminating drug-sensitive populations speeds up the selection of therapy-resistant clones that can grow without competition [31].

Alternatively, during the treatment, the ongoing diversification of tumor cells may allow them to adapt to the selective pressures of therapy, leading to *de novo*/acquired resistance, or tumor recurrences [31, 46].

Tumor relapse takes place when the tumor, once thought of as inactive or erased, regenerates. Usually, this happens when low-frequency clonal populations that survived the treatment expand [46].

The effect of therapy is usually transient. Given that the tumor micro-environment keeps on transforming, proceeding with the selection of clones, it is hard to modulate the anti-angiogenic action to keep up with these alterations [45].

The ultimate goal would be to predict and permanently monitor tumor evolution to improve the treatment and achieve durable responses. For instance, the earlier the first clonal expansion is detected, the less likely the tumor cell population fits the environment and the better the prognosis. Some mathematical models can infer this information. Nonetheless, they are still hard to implement in practice [31, 46].

1.2 Overview of the case study for the research

Even though it noticeably exacerbates the development of resistance and severe proliferation of tumors [38], different studies show that the application of anti-angiogenic drugs has improved the prognosis of patients [13, 39]. Given the contradictory nature of these observations, the actual effect of angiogenesis inhibition on tumor shrinking and overall survival remains to be determined [45].

An experiment was carried out to assess how the inhibition of angiogenesis, using targeted therapy, affects ITH. This case study was the basis of the thesis work. The experiment-related concepts are presented next.

A *mouse xenograft model* defines the implementation of human tumor cells into an immunocompromised mouse, a condition needed to avoid the rejection or destruction of the human tumor cells by the mice organism [33]. The similarities of the xenograft models to the human tumors make them well-suited to study ITH [33].

A *cell line* results from the technique of cell culture, in which cells can be grown [47]. A human cancer cell line comprises human tumor cells grown in culture. Multiple reasons justify using cancer cell lines, such as being cost-effective and immortalized. Plus, they possess characteristics similar to those of primary human cancers. They confidently recapitulate SNVs, CNAs, and gene expression identified in tumors [47].

The experiment involved the use of mouse xenograft models, in which a human breast tumor cell line known as *MDA-MB-231* was injected into four mice. Two of them, *Mouse 49* and *Mouse 55*, represented the *control group*. *Mouse 61* and *Mouse 62*, that went under sunitinib treatment, became the *experimental group*. The two mice groups were compared to evaluate the impact of the anti-angiogenic drug on ITH.

The genomic profiles were obtained by WES of bulk tumor cells from four distinct regions of the primary tumor of each mouse, resulting in a total of 16 samples; 8 are from the control group and 8 from the sunitinib-treated mice. The sampling of each mouse tumor is depicted in figure 1.2.

The SNVs and CNAs got from the sequencing data of the tumor samples were called using *Mutect2* [48] and *CNVkit* [37], respectively. It is possible to over-estimate ITH when calling variants from multiple samples that have different coverage or purity values [31]. Thus, it was sought to adjust the average sequencing depth of all samples to 20×.

The same human tumor cell line was injected into each mouse, allowing the comparison of the different samples. Since the mice xenograft models only contain cancer cells, we do not need to estimate the purity value of the samples, which will be 100% for each one of them.

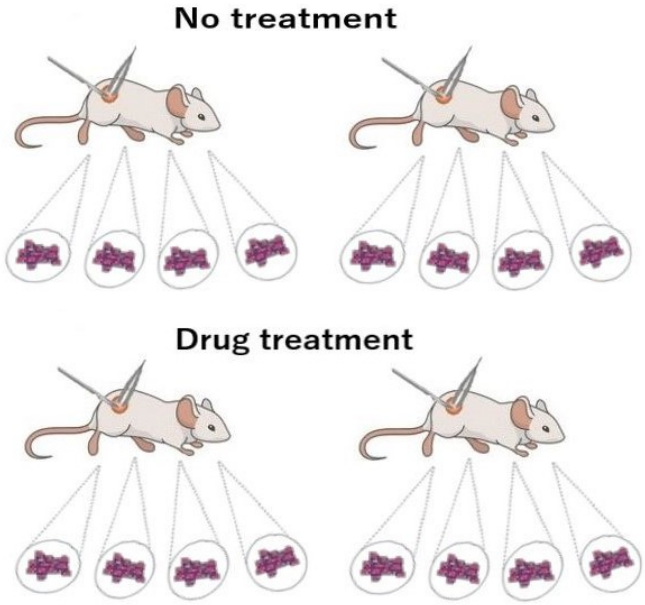

Figure 1.2: Mice tumor sampling scheme, adapted from [49] and inspired by [50].

1.3 Objectives of the dissertation

The mice case study raises two key questions. First, are there differences in the heterogeneity of the tumor genome sequencing data of the control and treated groups? In other words, how did the ITH progress in each of the groups? Did the anti-angiogenic treatment drive ITH?

Second, instead of focusing the comparative analysis of the control and treated groups exclusively on heterogeneity, can the mutational profiles of the genes of the samples provide relevant information about what differentiates them? For instance, can we distinguish two samples, regardless of whether they have the same ITH values, based on the genes data?

The work of this dissertation had two main objectives, including applying computational methods to estimate and compare ITH, assessing the effect of angiogenesis inhibition in the context of the case study.

The other goal was to explore the genetic information features to compare the groups. Different dimensionality reduction techniques were run. Together with the mice sequencing data, the sequencing data of TCGA patients, comprising multiple tumor types, were input into the algorithms. Such allowed the observation of how the mice data points of the two groups were positioned along the data manifold when using more examples.

The content of this thesis is divided as follows: chapter 2 overviews the methods for heterogeneity estimation and dimensionality reduction techniques; then, chapter 3 describes the work that was done to achieve the two mentioned goals, which methods were applied and their results; at the end, chapter 4 discusses the conclusions obtained, with future directions for investigation.

Chapter 2

Related work

Cancer research has been driven by technological progress, especially by informatics tools that process, analyze, and efficiently display large amounts of biological data. These were fundamental to increase the knowledge about the disease and, in particular, about the heterogeneity within tumors [5].

The investigation on ITH has also been supported by other fields. Section 2.1 refers to mathematical statistics measures that can be used to quantify ITH based on data distribution analysis, namely, *population diversity indices*.

The computational characterization of ITH has been the research topic of many publications over the last few years [16]. Considering that not all of them can be covered here, only some of the most popular are reviewed.

Several computational methods have been developed to estimate ITH and perform subclonal reconstruction from next-generation sequencing data [16]. They can be divided into two classes according to the level of information provided.

Section 2.2 discusses a category of tools that have been implemented specifically for ITH estimation. These computational models approximate the number, mutational compositions, and frequencies of the tumor populations.

Section 2.3 addresses a group of techniques that not only infer the composition of the clones but also recreate the ancestral-descendant relationships between these. They generate different phylogenetic trees according to the configurations that likely can represent the true phylogeny of the tumor populations.

Machine learning has a widespread impact on society, including on the investigation of cancer. Unsupervised learning, namely dimensionality reduction algorithms, has been involved in detecting relevant features hidden in very high-dimensional genomic datasets. In Section 2.4 are outlined a set of dimensionality reduction techniques that can also be applied to cancer data.

2.1 Population diversity metrics

Defining the different clonal populations within a tumor is challenging, requiring expensive and complex methods [51]. Tools that combine information of SNVs or/and CNAs are computationally intensive and typically imply the comprehension of their underlying theoretical models [34].

For this reason, some studies have chosen other strategies to estimate ITH, for instance, conventional mathematical statistics approaches to the analysis of the data distribution of populations [51].

Most of these are population diversity measures that were reoriented to the characterization of the genetic diversity within a tumor [52]. The standard *population average metrics* assume a normal distribution of the data to save time and simplify the analysis at the expense of the precision of the estimate [52].

They produce a single value that can be easily understood, even though it is not fully representative of the ITH of a sample [52]. Some are distribution statistics parametrized with a single attribute, as the mean and the standard deviation [52]; others are measures that work with distributions defined by two or more parameters [52].

Gaussian statistics like the z-score, skew, and kurtosis of the data distribution are also alternatives. Histograms and Q-Q plots are suitable for the characterization of the shape or modality of a distribution (its number of different peaks) [52].

Mutant-allele tumor heterogeneity or *MATH* is one of the most adopted indices to describe ITH from WES [15, 31, 34, 51]. It is based on interpreting a histogram or density plot of the VAFs in a tumor sample, ignoring the effect of CNAs [15, 34].

MATH is a metric of the distribution width, normalized by the median VAF to account for normal cell contamination (with tumor purity < 100%). This scoring method is given by the *mean absolute deviation*, the difference between the VAF of each mutation and the mean VAF, divided by the median VAF [34]. The wider the distribution, the more dispersed the frequencies are and the more diverse the population is [51].

Nonparametric statistics do not make assumptions about the parameters of a population distribution. These include the *interquartile range* or IQR, percentage of outliers (points that are far away from or are inconsistent with the data pattern), *Kolmogorov-Smirnov distance* or KS, *Simpson diversity index*, *Shannon index*, *quadratic entropy*, which have been applied to characterize the tumor population diversity [52].

The percentage of outliers is also a means by which we can assess the normality of the distribution. The higher the number of outliers, the more heavy-tailed the distribution is and diverges from a Gaussian shape, as the values are more spread [52].

KS distance is a well-known statistic that can function as a normality test to check whether a distribution differs from a Gaussian. The *Anderson-Darling* statistical test, which evaluates how well the sample data fits the normal shape, is another option to carry out this verification. The distributions of tumor populations that are highly heterogeneous present deviations from a bell shape, having multiple peaks [52].

The Simpson diversity index can be adapted from ecology to estimate the heterogeneity within the populations of a tumor. In this setting, it is the probability that two mutations randomly picked from a sample have different VAFs, belonging to distinct sets of variants [31]. A higher probability corresponds to a higher ITH.

The Shannon index is another of the measures that have been widely applied to diversity estimates of species in ecological sciences [52]. It can also be adjusted to quantify ITH, calculating the proportion of each subpopulation in a tumor [31].

Both the Shannon and Simpson indices are *entropy* measures. These describe how much information is provided by an attribute of a population. In ITH estimation, they represent the entropy associated with the VAF of mutations [52].

These indices have the disadvantage of not accounting for the difference in magnitude between the frequencies of the species. Such implies that low-frequency populations will not be detected since their contribution to the population diversity is almost negligible. Thus, these measures may underestimate ITH [52].

The quadratic entropy incorporates a distance matrix that considers the magnitude of the differences [52]. This addition allows a more accurate measurement of the ITH and the identification of low-frequency clones, with a wider range of detected values smoothing high data distribution peaks.

These strategies need to be supported by an informatics tool having libraries with mathematical statistics functions and allowing the computation and visualization of data distributions. Open-source software, for instance, *R* and *Python*, have been the preferences for these tasks [52].

Population average measures aim for simplicity and obtain fast results. On the other hand, compared to the analysis of single cells, the generalization of conclusions to the population level excludes much information [52].

However, cell-by-cell scoring is costly and thus becomes restricted to examining fewer genome regions. Nonetheless, the newest developments in single-cell sequencing have promoted the use of single-cell metrics [52].

2.2 Clonal composition inference methods

Recently, several computational methods have been designed to characterize ITH [16]. This first class of tools infers the clonal composition of a tumor sample, detecting the different populations and estimating the frequency of each clone [16, 31]. Understanding the population structure of a tumor is crucial for prognosis and treatment [16].

Some techniques perform this reconstruction based only on SNVs, also known as *SNV-based reconstruction methods* [30]. Others solely identify regions that are non-diploid by inferring CNAs, named *CNA-based reconstruction methods* [30, 53].

Some tools take both SNVs and CNAs as input [54]. The CNA-based reconstructions are more demanding, given having to simultaneously estimate the population frequencies and the new copy numbers of each segment affected by CNAs [53].

Most cancer studies have profiled tumor bulk samples [31]. Determining the population structure and estimating ITH from bulk sequencing data is a complex computational task [55]. This *clonal deconvolution* is a challenging problem, as neither the genetic constitution of the clones nor the number of populations is known [55, 56].

Among the many methods that have been developed to automate clonal deconvolution from tumor bulk sequencing data [16, 31, 54], some work only with CNAs, *THetA* [57] and *TITAN* [58], or only with SNVs, *Clomial* [59] and *QuantumClone* [60], or with both, *PyClone* [61] and its new/improved version *PyClone-VI* [55].

The low sequencing depth used in many large-scale cancer investigations (between 30× and 40×) displays a high variance in the number of SNV reads covering a genome position [16, 31]. In turn, each CNA affects many reads, providing more reliability for clonal inference in tumors with populations presenting different copy number profiles [16].

THetA takes as input raw copy number segment estimates from sequencing of the entire genome [62] to perform subclonal reconstruction on tumor populations that also contain normal cells [16, 57].

By applying a *Bayesian information criteria* (BIC) it defines and optimizes an explicit probabilistic model for estimating the proportion of each subpopulation in a tumor [16]. The use of the BIC allows it to choose among the different models that may represent the underlying data while balancing its likelihood and the model complexity [16].

TITAN adopts a different strategy [16], comprising a two-factor *Hidden Markov Model* or HMM [58]. A HMM is a statistical framework representing the evolution of observable events that depend on internal factors that are not directly visible [63].

The HMM is fit to the data via *expectation-maximization* (EM) [58]. The EM is an iterative algorithm to assess the model latent/hidden parameters that maximize the data likelihood [58]. In each iteration, it takes in the parameter values got in the previous step to try to find new ones that further increase the likelihood [59].

Similar to THetA, TITAN is based on whole-genome sequencing. While it estimates at the same time CNAs and loss of heterozygosity events from sequencing read depths and the BAF of germline heterozygous SNVs, THetA solely takes the former into account to assess ITH [16, 57, 58].

THetA only estimates the total copy number of each genome segment, whereas TITAN also calculates the allele-specific copy numbers [57, 58]. Since both techniques rely just on CNAs, they can not detect tumor populations that do not contain any copy number variations [57]. Besides, THetA and TITAN can not extract information for analysis from multi-region tumor sequencing data simultaneously [16].

CNA-based reconstruction methods from multi-region tumor sequencing data are currently lacking [31]. These have the potential to improve the identification of CNAs [31], as shown by *CNT-MD* [64]. This technique exclusively considers CNAs for clonal composition inference but can pool information from multiple samples [54, 62].

Clomial is a program designed to work merely with SNVs in diploid, copy number-neutral regions, which is a limiting factor of the precision of its reconstruction. It has the

advantage of being able to pool information from multiple samples [62].

It takes raw SNV calls in the format of two matrices containing, respectively, the total and alternative number of reads ($R_{N\times M}$ and $X_{N\times M}$, where N is the number of mutated segments and $M > 2$ is the number of samples, with $M - 1$ tumor samples), and the assumed number of clones ($C > 2$, with $C - 1$ tumor clones) as input [59].

Clomial infers the underlying tumor populations through a matrix deconvolution framework [62]. By dividing the R and X matrices, it calculates the matrix of VAFs ($VAF_{N\times M}$), decomposing it into a product of two matrices that will be inferred [59, 62].

The first is the *genotype/clone composition matrix* ($Z_{N\times C}$), a matrix of hidden variables indicating to which unknown populations each mutated locus belongs to [59, 62]. The other is the *clone frequency matrix* ($P_{C\times M}$), representing the proportion of each population in each sample [59]. At the start, it is filled with any value combination that sums to 1 in all its columns [59].

To find the optimal clone composition matrix parameters that maximize the likelihood of the data, Clomial trains its underlying model using the C value [59] while it simultaneously changes the clone frequency matrix values accordingly [59]. To do this, it uses a probabilistic binomial model with an EM algorithm [54, 56, 59].

The VAF estimates are affected by the assumed sequencing noise, decreasing the subclonal reconstruction accuracy. Choosing a distribution unable to handle this noise can lead to under or overestimating the number of clones. Binomial models can capture the read depth influence, while fixed-variance Gaussian noise models can not [28].

Notably, Clomial fulfills some phylogenetic rules that allow us to manually build a phylogenetic tree on its results [59]. However, it does not automatically output an approximation of the real phylogeny of the tumor clones, thus belonging to the class of methods that only automate the inference of the tumor populations and their frequencies [56].

Another technique that relies only on SNVs and can analyze multi-region data is QuantumClone [54]. This method applies a *Bayesian mixture model* that can deal with multimodal data, which in this setting are the different tumor populations [55, 60, 65].

QuantumClone performs the clustering of the SNVs by their CP values ($\hat{\theta}$) that depend on the *a priori* number of copies bearing each variant (NC), which values are unknown [60]. Each VAF therefore corresponds to several possible values of CP, in which each solution is associated with a value of NC [60].

To solve the problem of the non-uniqueness of a solution [60], the Bayesian mixture model is fit to the data using an EM algorithm that approximates the *maximum a posteriori* (MAP) probability estimate [55]. It finds the model parameters of a binomial distribution that maximize the probability of observing a specific number of reads with a variant, given that the latter belongs to a clone with a CP value of θ [55, 65].

The MAP estimation for mixture models is prone to overfitting since it tends to use the maximum number of clusters possible [55]. QuantumClone addresses the model selection problem using the BIC to determine the number of clusters/clones [55, 60].

The application of the BIC demands that multiple runs of the method are performed, varying the number of clusters to test which one is the most appropriate. This model search is computationally expensive [55].

Based on SNVs, allele-specific copy numbers, and tumor purity data, PyClone identifies the populations in a tumor and their proportions. This method uses a *hierarchical Bayes statistical model* [15, 16, 61].

It defines a beta-binomial distribution for VAFs that accounts for the various sources of noise introduced by sequencing more effectively than a binomial model does, for example, for datasets with overdispersed variance in coverage [27, 61].

PyClone simultaneously identifies groups of mutations and their frequencies in a tumor sample with *Bayesian nonparametric clustering*, deducing the existing number of clusters and their composition [16, 61].

Such avoids the need to fix the number of groups *a priori*, allowing for the uncertainty of the CP estimates in this parameter [61]. Similar to QuantumClone, PyClone carries out a complex model search that tests all the numbers of clusters/clones possible, making its use computationally prohibitive [55].

The runtime of PyClone increases quickly with the number of SNVs in a tumor sample since each model selection iteration becomes much slower. This variable can be a limitation for the analysis of heavily mutated tumors [55].

Such led to the development of an improved version of the algorithm, *PyClone-VI* [55]. By changing the approach to inferring the number of clones, PyClone-VI becomes orders of magnitude faster than PyClone, providing a similar quality of subclonal reconstruction and improved clustering performance [55].

Instead of using a nonparametric model, PyClone-VI relies on a *finite mixture model* for clustering of mutations, assuming that there is a limited number of populations in a tumor sample [55, 65].

PyClone-VI uses a *variational inference* (VI) procedure to perform model selection, considering only the reasonable number of clusters supported by the data. The cutting in the number of restarts of the iterative model selection process that are used to test different numbers of clones allows a much cheaper search than with PyClone [55].

Still, PyClone-VI performs worse than PyClone concerning the rigor of the estimate. The VI delivers only posterior approximations, updated probabilities of the unknown variables considering newly observed data, of unknown precision [55, 65].

On the other hand, PyClone guarantees to estimate the posteriors to arbitrary accuracy. In most cases, this increase in the accuracy is not mandatory, nor is it worth the much higher runtime of PyClone compared to PyClone-VI [55].

Table 2.1 summarizes some characteristics of the different methods discussed, namely if they require as input SNV or CNA calls, or a purity estimate, and also if they can pool information from multiple tumor samples. Since both share all the listed features, PyClone and PyClone-VI are represented together. Nonetheless, as mentioned, PyClone-VI is much more efficient than PyClone.

Table 2.1: Comparison of clonal composition inference methods [26, 57–60].

Method	SNVs as input	CNAs as input	Purity as input	Multiple samples
THetA	×	✓	×	×
TITAN	×	✓	×	×
Clomial	✓	×	×	✓
QuantumClone	✓	×	✓	✓
PyClone(-VI)	✓	✓	✓	✓

2.3 Phylogenetic reconstruction tools

The fast advancements in high-throughput sequencing and the growing awareness of the role of the evolutionary theory in cancer research led to phylogenetic studies of the tumor progression [56]. While tools to infer the clonal composition of a tumor provide clinically relevant insights into ITH, they do not approach the evolutionary relations between the different clones [16].

Such knowledge can have a fundamental impact on clinical intervention, as the identification of the evolutionary paths followed by tumors is crucial to predicting the progress of the disease [16]. Computational modeling can support the analysis of the mechanisms of evolution and enable the inference of the tumor dynamics [31].

To date, various methods have been developed to reconstruct phylogenies from bulk tumor sequencing data [31]. These combine clonal deconvolution with the automated inference of phylogenetic trees [16].

The computational reconstruction of the phylogenetic relationships among the different clones in a tumor is also a demanding task. In general, multiple evolutionary tree configurations can consistently represent the true phylogeny of a tumor [16].

These techniques often require prior clustering of CP or CCF values to define groups of mutations because of the demanding nature of enumerating a limitless tree space with multiple possibilities of configurations [31].

Similar to the clonal composition techniques, among the different methods that can output clonal evolutionary trees, some work simply with SNVs, or CNAs, or both. *PhyloWGS* [53], *SPRUCE* [66], and *Canopy* [62] are all methods that receive as input SNVs and CNAs. Only the referred complete clonal evolution reconstruction methods are approached here, but there are many others [16, 31, 54, 56].

PhyloWGS defines clones by grouping mutations with identical or similar VAFs, requiring as input the SNVs, allele-specific copy numbers, and tumor purity [67]. It was the

first fully automated method to use SNVs and CNAs to reconstruct tumor composition and evolution, generating phylogenetic trees as output [26].

PhyloWGS is based on a very similar model to the one of PyClone but substitutes the Bayesian nonparametric clustering with a *tree-structured stick-breaking process* [55], which is characterized by an *infinite mixture model* [68].

The components of the model have dependencies corresponding to the evolutionary relationships between tumor populations [68]. Identical to PyClone, it is not suitable to be applied to highly mutated tumors, quickly increasing the algorithm runtime [15].

The output of PhyloWGS provides many trees that may represent the clonal evolution, each scored according to its complete-data log likelihood [67]. PhyloWGS was developed from whole-genome sequencing data, which is restricted to a lower sequencing depth. Still, some studies argue that the higher number of mutations sequenced across the entire genome may compensate for this limitation [53, 67].

For the overlapping of SNVs and CNAs, PhyloWGS infers their temporal ordering and resolves their *phasing*. Such consists of determining if two heterozygous genomic events within a chromosome belong to the same or different allele copies [26, 62].

However, PhyloWGS does not infer CNAs. This way, it requires absolute and not relative allele copy numbers as input and assumes that the ratios of the cancer cells with each CNA are known [26, 62]. The CNAs are previously processed by other algorithms, for example, THetA [62].

SPRUCE is another phylogeny reconstruction method that does not infer CNAs [26]. On the other hand, it relies on a very different structure from the one that is used by PhyloWGS [54, 56].

PhyloWGS adopts a probabilistic model, using a Bayesian sampling algorithm [56]. Instead, SPRUCE relies on a specialized model that jointly provides clonal deconvolution and tree generation, applying a *combinatorial enumeration* algorithm [54, 56].

Combinatorial methods optimize over a discrete set of possible topologies. Generally, they are the most efficient methods, being suitable for simpler models only [56]. Examples of these strategies include the use of *integer linear programming*, which converts the topology selection to a mathematical optimization problem that can be performed by efficient solver programs available [56].

SPRUCE infers phylogenetic trees jointly from SNVs and CNAs, pooling information across multiple tumor samples [26, 66]. It represents each tumor population mixture, called *taxa*, as a sequence of characters (genomic events that can have one of several distinct states) [66]. For instance, CNA events are modeled as multi-state [16].

The algorithm receives a matrix with rows as vectors of the different possible taxa SNV and CNA states [66]. It outputs a tree whose leaves are the taxa mixture, the tumor clones, labeled by their different states.

The internal vertices of this tree are the *ancestral*, the previous states of each mutational group [66]. The goal of SPRUCE is to infer the tree that best represents the tumor evolution, maximizing an objective function, the maximum data likelihood [66].

Canopy is a computational method that follows the same approach as PhyloWGS. It is established on a probabilistic model with a Bayesian algorithm to sample evolutionary trees and compute the posterior distributions of the different tree configurations [26]. At the same time, it receives input data in matrix format like SPRUCE [54, 56].

Different from PhyloWGS and SPRUCE, Canopy starts with raw allele copy number ratios estimated by CNA caller tools and not absolute values. It is the first subclonal reconstruction technique that outputs phylogenetic trees and infers CNAs [16, 26].

Canopy applies a *Markov Chain Monte Carlo* or MCMC procedure that samples from a distribution using Markov Chains. Each random sample is used to calculate the next one, iteratively approximating the sample distribution of the current step to the one that maximizes the likelihood of the data [26, 65]. A higher number of MCMC chains corresponds to a more accurate estimation but also a higher algorithm runtime [55].

Canopy builds binary trees and assigns mutations to the inner nodes and the leaves, except for the leftmost one, which is used to represent the normal population (not containing any variants) [26]. Not all nodes need to receive mutations [26].

Nodes with mutations represent tumor populations, and the nodes without variants, excluding the normal population node, can be collapsed to their parent node [26]. Shared groups of mutations present at the internal nodes but not in any leaf evolved into new populations and were eradicated by other clones. The number of tree leaves represents the different expected tumor populations [26].

The range of possible values of the assumed number of tumor clones is input into the algorithm. Based on this value, Canopy calculates the tree configurations for the different numbers of populations provided. It then compares all the reconstructions via BIC to assess which one is the most likely to represent the true tumor phylogeny [26].

Table 2.2 reviews the features of the different phylogenetic methods approached. Notably, Canopy also has the advantage of being able to phase CNAs with different genome endpoints, in opposition to PhyloWGS and SPRUCE [62].

Table 2.2: Properties of clonal phylogeny reconstruction tools [26, 56, 62].

Method	Raw CNAs as input	Infers CNAs	Phases CNAs	Model type	Algorithm type
PhyloWGS	×	×	×	Probabilistic	Bayesian sampling (MCMC)
SPRUCE	×	×	×	Specialized	Combinatorial enumeration
Canopy	✓	✓	✓	Probabilistic	Bayesian sampling (MCMC)

2.4 Dimensionality reduction techniques

The massive amount of genomic data produced by next-generation sequencing technologies has urged the need for more powerful tools. These should manage, analyze, and extract information that can support and speed up the scientific discovery [23, 69], being increasingly seen as fundamental elements in cancer research [5].

Genomic event datasets contain many instances, each described by several features. Namely, knowing which genes they affect may provide meaningful biologic insights that are hidden in very high-dimensional data [23].

The limitations of the classical statistics methods in processing these data, characterized by multiple variables, led to the emergence of machine learning in the oncology field [70]. In the last few years, machine learning algorithms have been applied to several tasks in cancer research, including tumor diagnosis and screening [21, 22, 70].

Machine Learning algorithms can be divided into three categories (including supervised and semi-supervised). Unsupervised learning aims to detect hidden patterns in the data and understand these without relying on feedback or prediction [23, 70].

Among the range of different unsupervised learning algorithms, the dimensionality reduction techniques are essential for the analysis of datasets with a very high number of attributes [69]. Datasets having many features are usually very sparse, with most points being likely to be farther away from each other. Thus, the identification of similarities in the data becomes a computationally expensive and complex task [71].

Unsupervised learning, especially dimensionality reduction, is well suited to face this problem, known as *curse of dimensionality*. Dimensionality reduction algorithms decrease the high-dimensional feature space to a much lower dimensional representation without significant information loss, finding the most relevant attributes of the data [72]. Dimensionality reduction allows an effective noise removal [69].

Specifically, these methods project high-dimensional input data to a low-dimensional space, keeping as many of the most significant features as possible while removing the redundant ones [72]. Once the data is in a lower dimension, we can detect its patterns more efficiently because much of its noise was reduced [72].

There are two major branches of dimensionality reduction, the first of which is denominated as *linear projection* of the data from a high-dimensional to a low-dimensional representation of the features. It includes algorithms such as principal components analysis (PCA), singular value decomposition (SVD) and its truncated version, and *Non-negative matrix factorization* (NMF) [71, 72].

The second class of dimensionality reduction methods is known as *manifold learning*, or *nonlinear dimensionality reduction*. When the data does not lie on a hyperplane, for instance, a two-dimensional line or a three-dimensional plane, linear methods can not effectively separate the different instances [23, 71].

A manifold is an object that can be bent and twisted in a higher dimensional space but which locally, in a lower dimension, resembles a Euclidean space [71]. The techniques for

manifold learning rely on the assumption that most datasets in high-dimensional spaces lie close to a much lower dimensional manifold representation [71].

There are multiple techniques that can perform non-linear projections [69], including t-distributed stochastic neighbor embedding (t-SNE), metric and non-metric multidimensional scaling (MDS), and isometric feature mapping (Isomap) [72].

Among the first class of dimensionality reduction algorithms, PCA is the most widely used method [73]. This algorithm has different versions, including the incremental and sparse PCA; the standard PCA is described here [72]. The goal of PCA is to find a set of orthogonal axes along which the variance in the data is the highest possible [23].

PCA does this by identifying which features are the most important in explaining the data variability [72]. An attribute that does not contribute much to the variance of a dataset does not provide relevant information about it [71]. PCA addresses the correlation of the highly variable characteristics and attempts to linearly combine them, identifying the dominant patterns in the data [72].

The reduction of the set of correlated attributes carries on, with the algorithm finding the directions of maximum variance in the original high-dimensional space and projecting them into a lower dimension [23, 72]. This process repeats until the newly found component contribution to explain the dataset variance is almost negligible [73].

These derived components are called *principal components* [73]. This method aims to represent the data using the smallest number of linearly uncorrelated features possible [72]. With these, it is easier to understand the underlying structure of the data and we can reconstruct the original dataset without losing much information [72].

Another approach that can be applied to learn the data hidden patterns is a matrix factorization technique named SVD. This method is based on the rank reduction of the original features matrix [72], which corresponds to decreasing the maximum number of its linearly independent vectors or using the minimum number as possible [74].

The original matrix can be recreated using a linear combination of some of the vectors in the smaller rank matrix [72]. To generate it, SVD keeps the vectors of the original matrix that provide the most information. In other words, it chooses the vectors that have the highest singular values, retaining the characteristics that better describe the original features space [72].

The original matrix is decomposed into three matrices: $Z = USV^T$. The column vectors of the matrices U and V are orthonormal. The diagonal of the matrix S contains the singular values of the matrix Z [75].

The columns of V^T (that is, the rows of V) define the new axes, with the rows of U representing the coordinates of the objects in the space spanned by these axes. Singular values are scaling factors that characterize the relative importance of each new axis. The output of this technique is always dense, even if the input data is sparse [75].

In *truncated* SVD, a faster version of the original full SVD, only the t largest singular values of the S matrix, corresponding to the t column vectors of matrix U and t row vectors of V^T, are kept [76].

This way, while the original full SVD keeps all the columns of U and all the rows of V^T, the truncated SVD drops all the features except for the t number of attributes provided as input [76]. The SVD produces an optimal low-rank approximation of the initial matrix Z with a minimal reconstruction error [76].

Non-negative matrix factorization yields additive non-negative basis vectors, creating a parts-based representation [75]; it is used for the dimensionality reduction of non-negative data. One of the differences between NMF and SVD is the non-negativity aspect, which allows only non-negative additive combinations of vectors, the hidden features, to reproduce the original features space [75, 77].

The whole becomes a combination of the parts, which are the different non-negative basis vectors. NMF factorizes an input matrix into two others. The columns of the first are the non-negative basis vectors; the other matrix contains the weights that, together with the basis vectors, approximate the columns of the original matrix [75].

The two matrices resulting from the decomposition have fewer entries than the original one. Not all entries of the original matrix are needed to perform a decomposition [75]; NMF should be able to handle missing entries in the target matrix [77]. NMF retains more localized patterns of the data. However, it consumes much memory when the input matrix is large [75].

NMF performs a blind decomposition, whose results may not be reliable; this can be a limitation for its application in areas where high accuracy is needed, as it happens in cancer research, especially prognosis prediction. Different versions of NMF seek to overcome these limitations [77].

Concerning nonlinear dimensionality reduction methods, t-SNE models each high-dimensional data point into a 2D/3D space [72]. It converts high-dimensional Euclidean distances between data points into conditional probabilities on similarity [73].

It estimates, based on the distances between instances in the high-dimensional space, the probability of each data point being neighbors with one another [23]. The similar and dissimilar points are, respectively, close and farther away from each other [72].

The goal of this technique is to obtain a set of projected coordinates in a lower dimensional representation that better approximates the probabilities measured in the original high-dimensional space [23].

Hence, it defines two probability distributions, one over the pairs of points in a high-dimensional space and the other over the pairs of points in a lower dimension. Distinct instances have a lower probability, whereas those alike have a higher value [72].

t-SNE is a powerful tool. However, when working with a huge number of dimensions, its complexity contributes to an enormous increase in the runtime, being difficult to apply it directly in practice [23].

In real-world applications of t-SNE, it is recommended to use another dimensionality reduction technique, such as PCA, to reduce the number of dimensions before applying the method. Consequently, the noise of the features that are fed into the algorithm is reduced, and the algorithm computation speeds up [72].

MDS is characterized by reducing the dimensionality of the features while preserving the similarity or dissimilarity between data points [71, 78]. It fits the data locally to capture its global structure [78]. By learning the similarity of the instances in the original high-dimensional space, it models them in lower dimensions [72].

It does so by minimizing a cost function that quantifies the difference between the similarities or dissimilarities of two points in different dimensions, one measured in the original high-dimensional space and the other calculated in the low-dimensional embedding, finding the d dimensional space that best preserves these [23].

There are two main variations of MDS, in particular, the metric and non-metric MDS [75, 78]. The difference lies in the criteria used by these. While the classical version of the method is a linear dimensionality reduction technique, the metric and non-metric MDS are nonlinear [78].

In the new MDS approaches after the classical version, the cost function was changed to preserve the distances between data points instead of the similarities [78]. Metric and non-metric MDS try to conserve the high-dimensional space distance between data points when reducing the dataset to a lower dimension [78].

The first version is called metric MDS given that it applies a distance measure in its optimization [52]. On the other hand, non-metric MDS, rather than applying a distance metric, uses a non-parametric monotonic function [78].

Isomap is a nonlinear generalization of the classical MDS. It uses a kernel constructed from the *geodesic distance* between points, which is the length of the shortest path between two data points on a manifold that may be curvy [78].

However, it is complex to calculate this distance since it requires traveling from one point to another point on the manifold [78]. Therefore, Isomap approximates the geodesic distance by piece-wise euclidean distances [78].

It comprises three steps. First, a graph of the local connectivities between the data instances is constructed; each point is linked to its gth nearest neighbors, with the edges weighted by the euclidean distances between them [23].

Next, the shortest paths on the graph are used as an approximation of the geodesic distance between all pairs of points, calculated using methods such as the *Dijkstra algorithm* or the *Floyd-Warshall algorithm* [23, 78].

MDS is then performed on the matrix containing the estimated pairwise geodesic point distances. In the end, the output of Isomap provides the low-dimensional Euclidean projection that best preserves the computed/approximated geodesic distances [23].

The dimensionality reduction performed by Isomap has limited quality. The precision of the representation depends on the accuracy of the pairwise geodesic distances calculated, which are complex to compute and are only approximately estimated [23].

The computation of the geodesic distances is also the main performance bottleneck when using Isomap with large datasets. Similar to t-SNE, it may be beneficial to apply another dimensionality reduction tool before performing Isomap [23].

Chapter 3

Methods and results

The practical work of this dissertation was divided into two parts, considering the two main goals defined. The first part focused on the estimation and comparison of the tumor heterogeneity values of the control and treatment groups of the mice case study.

Among the options available to assess the ITH, a popular technique from each of the existing three classes of methods was selected, including a population diversity measure, a clonal composition inference tool, and a phylogenetic reconstruction algorithm. This way, the entire spectrum of approaches to the computational modeling of the heterogeneity within tumors was covered.

The *tumor heterogeneity index* [79] was the population diversity metric chosen to be implemented. Its definition and application to the mice tumor input data are summarized in section 3.2.

From the second group of techniques, the computational method selected to be run was PyClone-VI. The processing of the mice tumor data before being input into PyClone-VI, the algorithm execution, its functions with the different experimented parameter values, and the analysis of the output are discussed in section 3.3.

Regarding the third class of methods, the tool for the clonal evolution reconstruction opted for was Canopy. The different input matrices needed for this algorithm and the decisions made towards the phasing of the CNAs are explained in section 3.4.

The second part involved the application of dimensionality reduction techniques to explore whether the SNVs and CNAs of the mice genes could distinguish between the two groups. How the data of the mice and TCGA patients were associated, the different techniques tested, and the interpretation of the results are addressed in section 3.5.

Next, the input used for the two parts, the tumor data of the mice and TCGA patients, are described in section 3.1. The code scripts, output images, and other complementary files produced are available at the following GitHub repository.

3.1 Input data description

The somatic mutations of the mice tumors identified by Mutect2 were output in *variant call format* (VCF) files. The VCF is a generic text file format that stores DNA-related data, including SNVs [80]. It consists of a header and data sections; the header contains meta-information of the tags and annotations used in the data section [80].

The header also has a field definition line, with columns corresponding to the data section columns, describing each of the detected SNVs. It includes the number of the chromosome affected, genome position, reference and alternative alleles, site filter information, and other genetic data, in which are encoded the read counts of the normal and mutant alleles [80].

The VCF files corresponding to the 16 samples (4 regions extracted from each mouse) got their header section removed since only the data section content, in a table format, was necessary. The VCF file was then converted into a *comma-separated values* (CSV) format file so it would be easier to input it into a software tool.

Some input processing decisions were applied for all the tasks in both parts of the thesis work. Only the variants with a *PASS* value in the filter column were selected (that is, a variant in a specific genome position had passed all filters of the caller method [80]). This way, there is higher confidence that the SNVs were identified correctly.

Moreover, the sex chromosomes (X and Y) of the SNVs and CNAs were excluded, as multiple instances of the latter had missing or wrong information about the allele-specific copy numbers, for example, both with a value of 0.

The CNVkit (copy number variation calling tool) outputs files in a tabular format similar to the *browser extensible data* (BED) format, with additional columns [37]. The BED is a text file format used to save genomic region information [37]. It has a similar structure to the data section of the VCF files.

The CNAs were characterized by columns such as the chromosome number, genome start and end positions, local read depth of the genome region, BAF, and total and allele-specific copy numbers. Germline CNAs were chosen over somatic copy number variations for all tasks because only these had the BAF values that were required by some algorithms, for example, Canopy.

An issue emerged after noticing that some of the CNAs were following an unrealistic pattern. Namely, loss of heterozygosity events (with a major copy number of 2 and a minor copy number of 0) were detected multiple times in narrow chromosome portions, which seemed very unlikely. Thus, these could potentially be considered sequencing artifacts/errors.

Different executions of the tumor heterogeneity index and the PyClone-VI were performed considering or not these CNAs in the input data, testing the effect that they could have on the results. The possible CNA sequencing errors were removed when processing the input of Canopy, given that, to reduce the high execution time of the algorithm, it would also be helpful to include fewer mutations.

To evaluate the impact of the genes affected by SNVs and CNAs in distinguishing between the two mice groups, files with the gene annotations of SNVs were required for the second stage of the thesis work.

These described the SNVs, including the chromosome number, genome position, reference and alternative alleles, the influence of the variant on protein production, and also the name of the gene affected. This file was then intersected with another one that contained the identification code and genome start and end positions of each mutated gene so that all the information related to the genes could be together in one table.

The mice sequencing data was supplemented with TCGA patients data for the application of dimensionality reduction methods. These datasets can not be directly compared, but the distribution of the control and treatment groups on the data manifold composed of all the different patient cancers from TCGA can be analyzed.

The SNVs of TCGA tumors were described, among others, by the chromosome number, genome position, reference and alternative alleles and their read depth counts, the name of the gene affected by the variant, and the identification code of the patient to whom the SNV sample belonged.

The CNAs of TCGA tumors were divided into two files. One had the genome start and end positions, the identifier of the sample of the copy number modifications, and the identification code of the patient from whom it was extracted. The other contained the total copy number of each CNA sample identifier in each gene identification code.

All the CNAs were kept in the second part. This decision was taken considering that TCGA data added to the input would be highly predominant over the mice tumor samples and that using or not the possible CNA sequencing errors would have a negligible impact on the output.

3.2 Part I - Tumor heterogeneity index

The tumor heterogeneity index is a population diversity metric that can be applied to measure the heterogeneity within tumors based on the Shannon index with the VAFs of mutated loci [79]. Its value is higher when there are more groups of SNVs with similar VAFs, that is, a higher number of tumor populations [79].

By obtaining the VAF of each mutation and assigning them to *i*-th of N bins, the tumor heterogeneity index (THI) can be calculated using the Shannon index, given the probabilities of each VAF belonging to each bin (p_i) [79]:

$$\mathrm{THI} = -\sum_{i=1}^{N} p_i \ln p_i. \tag{3.1}$$

Different input configurations of the mice data were considered to evaluate the consistency of the tumor heterogeneity metric. First, only the SNVs were used. No CNAs were

included, which means that all the SNVs were considered, whether they were in diploid regions or positions with CNA events, that is, non-diploid regions.

Then, both SNVs and CNAs were taken into account. As mentioned previously, because of the chance of having loss of heterozygosity events as sequencing errors, two different executions were performed: one considering all the CNAs and the other removing those with a major copy number of 2 and a minor copy number of 0. The SNVs and CNAs of all mice samples were processed at the same time.

In any of the cases, the SNVs were loaded first. When using CNAs, these were input after. Regardless of considering or not all CNAs, these influence the VAFs of SNVs and, consequently, the heterogeneity estimation.

The tumor heterogeneity index is not a computational model that can use both of the types of genomic events to estimate ITH, being unable to consider the impact of CNAs on the VAFs. The solution to this problem was to use the CNAs as filters to delete all the SNVs in non-diploid copy regions (with major and minor copy numbers different from 1), masking this limitation.

The SNVs kept were used to build the THI. The VAFs of each somatic mutation in each sample were determined via the equation 1.1, based on the allele read counts extracted from the genetic data column of the tables describing the variants.

These were assigned to 10 bins, which were a reasonable number to represent the distribution of the VAFs. Each bin had a 10% range interval given that the VAF values, in percentage, are between 0% and 100%. For each sample, the probability of a variant frequency belonging to a specific bin (p_i) resulted from dividing the number of VAFs in the bin i by the total number of mutation frequencies in all bins.

Finally, the THI was calculated using the formula 3.1. If the probability of a VAF belonging to a bin was 0 (meaning that the bin was empty), that parcel of the sum would be automatically 0 (since $\ln(0)$ is undefined), or p_i multiplied by $\ln p_i$ if otherwise. The heterogeneity value for each sample was thus output as a single value, resulting from the negation of the sum of all probabilities.

Parallel to computing the heterogeneity index values, two other variables were explored. First, because of the possible noise introduced by sample preparation, sequencing, and genomic alignment [36], and that the population average metrics simplify the heterogeneity estimation assuming that the data follows a normal distribution, the normality of the distribution of the VAFs was checked.

The other studied parameter was the coverage of sequencing. The theoretical/average sequencing depth used for all samples was 20×. However, since the reads are not distributed uniformly over the genome and in consequence of sequencing noise, the *actual empirical coverage* (the exact number of times each locus in the target region is sampled) is not equal for all the genome positions [36].

Therefore, some regions of the genome can lack coverage or have much higher coverage than the expected sequencing depth [36]. This way, it was important to assess the

uniformity/quality of the data coverage by calculating the variance in the sequencing depth across the genome [36].

The total depth was determined by the sum of the read counts of the reference and alternative alleles of each SNV. The distribution of the total coverage of the reads was plotted, showing how much the average sequencing coverage differed from the actual sequencing depth in each sample/genome region.

To improve the quality of the analysis, one aspect that could have been investigated was the existence of outliers, both in terms of the VAFs and the total coverage of the SNVs, looking to smooth the data distributions. In practice, the thesis work did not delve into this question.

Which SNVs were used to generate the VAFs and the total coverage (if CNAs were considered or not, if the copy number events with major copy number 2 and minor copy 0 were removed or not) did not impact the results produced, as the number of SNVs that were kept in each case was similar.

Such was verified by plotting the different distributions in all three situations. The images of the distributions presented in the appendix A refer to the case where CNAs were used, but the possible sequencing errors were removed.

The data distributions of the VAFs and coverage of the SNVs were plotted using all samples together (average distribution) and each of the samples separately (each in a different subplot and all in the same plot). These were plotted with kernel density estimation (KDE), outputting the probability density function of the data [65].

The distribution of the coverage using all the SNVs from all samples together is represented in figure A.1. Its shape is similar to a normal Gaussian, even though the coverage values above 50 introduce some irregularities. Notably, the probability density function mode is close to the applied theoretical sequencing depth of 20×.

The same behavior can be observed in figure A.2, with the distributions of the coverage of the SNVs for each of the different samples assuming similar shapes, with slight variances in the mode value of each curve. The average sequencing depth value used was thus a good approximation of the actual genome data coverage.

The distribution of the VAFs of the SNVs from all samples is depicted in figure A.3. The data follows a positively skewed distribution, with the probability density function curve slopping to the right, distinct from a normal Gaussian. The same happened when the different samples were plotted (figure A.4), with the curves mostly overlapping.

Such indicates there is noise associated with the sequencing data that could have affected the results of the THI, even though it functions only as a qualitative measure of the heterogeneity, with little accuracy.

The heterogeneity index values for each of the samples of the mice, compared between groups, are depicted in figure 3.1. It includes using CNAs, removing the ones that could be artifacts, although the results were similar in all three settings.

As expected of this simple measure type, the results obtained do not allow us to make a solid conclusion about whether the heterogeneity values are different between the two

experimental groups.

By just looking at the plot, we would assume that the inhibition of the angiogenesis in the experimental group did not induce ITH, as the THI values of all mice are similar. However, these are only representative values, reducing much of the information in a single value.

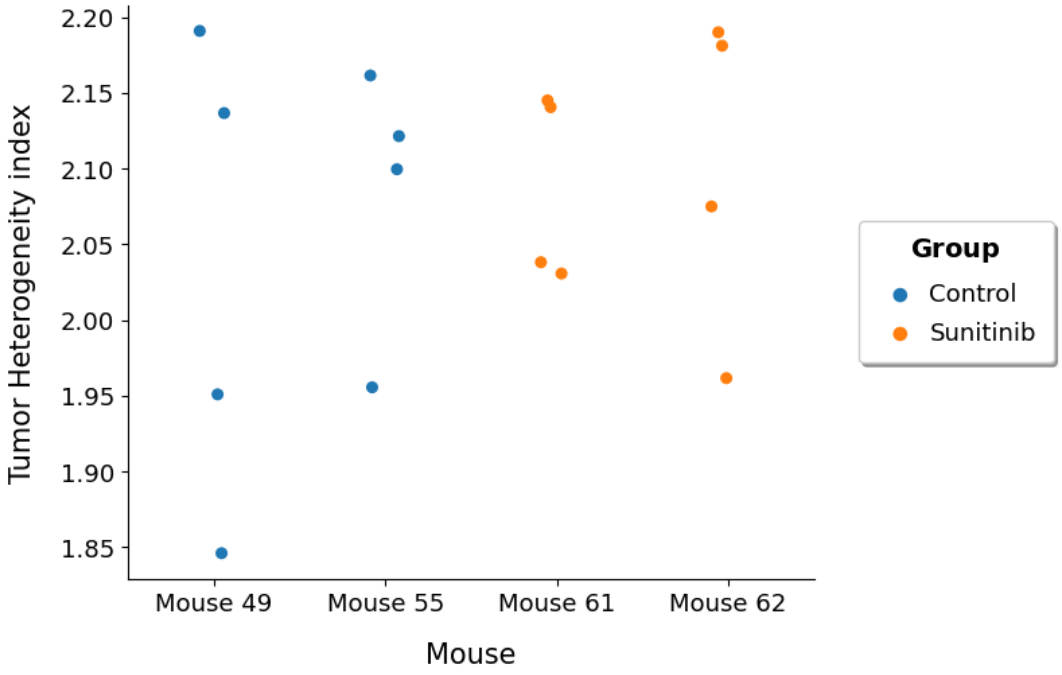

Figure 3.1: Tumor heterogeneity index comparison between mice groups discarding the CNAs that are possible sequencing errors.

3.3 Part I - PyClone-VI

Looking to improve the limited ITH analysis made via the tumor heterogeneity index, another more powerful tool was tested. The PyClone-VI is a computational model that can infer the clonal structure of a tumor.

It outputs the number of tumor clones, the mutations in each clonal population, and the ratio of tumor cells with each mutation in each sample. Since the tumor purity is 100% for all samples, the CCF will be equal to the CP according to the equation 1.3. The model implementation details and all its features can be accessed at the GitHub repository of PyClone-VI.

The initial step consisted in processing the input necessary to run PyClone-VI, a *tab-separated values* (TSV) format file with the SNVs of tumor samples described by eight columns: the identifier of the mutation and sample, the number of reference and alternative allele read counts, the minor and major allele copy numbers, the normal copy number and the tumor purity/content. The last two columns had fixed values of 2 and 1.0.

Different input configurations were experimented to get a more extended interpretation of the mice tumor data. The first one involved using all the mice samples together in the same file for the execution of the algorithm (names starting with “allMice”).

Considering that all the samples come from the same parent (the same tumor cell line was injected into the four mice), we can use them all together. Eventually, they could even share the same clones, which would make it easier to compare them.

At the same time, each mouse represents a separate evolution, so it makes sense to use only the samples of each of them separately in each input. Hence, four types of files were processed (names starting with “M49”, “M55”, “M61”, and “M62”).

After loading the data of all mice samples with the SNVs and CNAs, two different settings were once again explored. Two different versions of each of the five input files were created: one considering all the CNAs for the filtering of the SNVs (names ending with "_snvsInfo") and the other without the possible CNA sequencing errors (names ending with "_snvsInfoWithoutCNAsErrors"), to analyze their effect on the output.

The mutation identifier of each variant was built on the concatenation of the chromosome number, the genome position, and the alternative nucleotide base. The read counts of the reference and alternative alleles were extracted as in the previous method, with the sample identifier referring to each of the mice samples.

The processed SNVs were then intersected with the CNAs. Different from the THI input processing, in the PyClone-VI, the attributes of the CNAs used to filter the SNVs were not discarded. The mutation loci located between the start and end genome positions of copy number variations got their allele minor and major copy numbers information from these CNAs present in their region.

Before writing the tables to the files, one last aspect was considered. Since PyClone-VI removes variants not present in all samples, most of the SNVs would have been deleted as many of them were not shared by all samples. Each variant was thus inserted into all samples to prevent such from happening by assigning 0 to the reference and alternative allele read counts and 1 to both the minor and major allele copy numbers.

The algorithm was then applied to each of the input files. It consisted of two stages, corresponding to the two main commands of PyClone-VI, *fit* and *write-results-file*. First, the use of the fit performs variational inference based on the data, receiving as input the TSV files and outputting a *H5* data file with multidimensional arrays of data, saved in a *hierarchical data format* (HDF).

The fit command supports multiple executions, with the best one being output by the write-results-file command [55]. Two of the several optional arguments it can receive were considered: the number of clusters and the number of restarts of the variational inference step.

The number of clusters provided is used while fitting the model, setting up the maximum number of clusters used by the algorithm. The authors recommend a value between 10 and 40. The higher the number of samples input, the higher the number of clusters that should be used [55].

The number of executions of the algorithm is defined by the number of restarts. It will have as many runs as its value; usually, it refers to the restarts of the beta-binomial probability density function used, although this can also be parametrized [55].

There is a higher probability of finding an optimal variational inference approximation when using a higher number of restarts. However, the running time increases accordingly. The authors recommend a value between 0 and 100 [55].

Four different values of the number of clusters used for fitting were experimented with in the PyClone-VI execution for each file: 5, 10, 20, and 30 clusters. For each of these, three different numbers of restarts of the algorithm were applied: 10, 100, and 1000. This way, it was possible to assess the confidence in the results obtained and how variable these were.

For instance, to fit the data to a beta-binomial distribution, using a file with the SNVs of all the mice samples together and as parameters allowing up to 5 clusters and 10 random restarts, having defined an input and output name on the current path (-i and -o), the fit command can be written as it follows: pyclone-vi fit -i allMice_snvsInfo.tsv -o allMice_snvsInfo5c10r.h5 -c 5 -d beta-binomial -r 10.

To output the best random restart of the variational inference step, the output file of fit can be input to the write-results-file command, for example: pyclone-vi write-results-file -i allMice_snvsInfo5c10r.h5 -o allMice_snvsInfo5c10r.tsv.

It outputs a TSV file that can be imported into Python to be manipulated. Files were output for all mice samples together, and the samples of each mouse in each file, one with all CNAs used to filter and the other removing the CNAs with major copy number 2 and minor copy number 0, for each of the different combinations of parameters used.

The results file describes each mutation assigned to the found clusters, the tumor clones. It has six columns: the mutation and sample identifiers, the identifier of the cluster (starting from 0 to "found number of clusters" − 1), the CP/CCF and its standard error, and the posterior probability of a mutation belonging to a specific cluster [55].

The last column was used to remove mutations that had a posterior probability of belonging to a cluster of less or equal to 60% to have higher confidence in the composition of each of the tumor populations.

After processing the output files by creating columns with the execution identifier, the number of clusters used for fitting, and the number of restarts, among others, different approaches were taken on how to output/represent the results.

Even though the different samples of each mouse have the same identification, regions 1 through 4, these can not be compared, whether using all mice samples together or in different files. Region 1 in a mouse is not the same as region 1 in another.

The numbering of the regions does not have any physical meaning. The mouse tumors were cut in four, and each tumor was given the same identifiers for the samples; however, these could have been defined in sequential order (for instance, regions 1 through 16).

For both file types, using all mice samples together and each mouse samples in separate files, the mutations of each cluster were grouped by the average CCF, which is equal

to the CCF values of every mutation belonging to the same sample cluster.

However, when using the samples of each mouse in separate files for the PyClone-VI execution it is not guaranteed that the clusters are comparable, even though they could have the same identifiers in case the same number of clones was found. The cluster 0 of Mouse 49 is not the cluster 0 of Mouse 55.

On the other hand, such is possible when using all the mice samples together. The clusters are estimated by the algorithm using the frequencies of all mutations from all mice samples. Thus, the number of clones found in all samples will be the same, and the sets of mutations will be identical for clusters with the same identification, even though the CCFs of the same mutations in different samples and mice can vary.

Since the clusters can not be compared between the different samples when using each separate mouse data in each file, after joining the information of the four files in a table and calculating the average cluster CP for each sample, the total number of clusters/clones of each mouse for each configuration of parameters used was computed.

Given the restricted space to include the full plots of all executions with all the different parameters, only two of them are shown for each figure: 30 clusters and 10 restarts, and 30 clusters and 1000 restarts, discarding the possible CNA errors.

This way, we can see how the algorithm predictions change with increasing the number of restarts and selecting a higher number of clusters used for fitting, which allows PyClone-VI to obtain more robust results. The comparison between the total number of clusters found in the two groups of mice, using each mouse data in each file for the mentioned execution configurations, can be seen in figure A.5.

There are not many conclusions that we can take from these plots based just on comparing the number of clusters between the two mice groups. The ITH will be higher as greater the number of clusters found. However, there is not any pattern that can tell us that the heterogeneity is different between the groups.

To assess the influence that grouping the SNVs of each cluster would have on the quality of the information provided by the data for all the mice samples together used in a file (since it is possible to compare between all the clusters), it were output plots with the CCF of each cluster in each sample and the distribution of the CCFs of the variants in each cluster of the two mice groups.

Figure A.6 illustrates this for the clusters 3 and 5. Looking at the distribution of the CCFs of cluster 3 in the top left corner, we could argue that the control group has more mutations with a higher CCF, thus pointing to a possibly higher ITH.

While the average CCFs of cluster 3 in the two mice groups agree with the CCF distributions of the mutations, cluster 5 shows a different perspective. The average CCFs of cluster 5 in the two mice groups seem similar. However, by looking at the CCFs distribution plot, we can notice that there are more mutations with a higher CCF in the treatment group.

To be able to compare the different clusters when using each mouse data in different

files, the tumor heterogeneity index applied in the previous method was calculated according to the equation 3.1 using the CP of each cluster of each sample. Therefore, each sample of each mice got represented by a singular heterogeneity value.

The THI was also used for the file configurations with all mice samples together. A statistical measure was chosen to assess whether the populations of the two mice groups were different or not based on the heterogeneity indexes. Since it was more logical to compare the clusters when using all the mice together, this metric was only applied to the heterogeneity indexes in this case, not when using each mouse in separate files.

The *T-test* is a type of statistical index that can compare the means of two groups and evaluate whether they are similar or different. It is a parametric method for situations in which it is possible to define the probability distribution of the populations. This test can be used when the distribution of the samples follows a normal distribution [81].

For the normality check of the heterogeneity indexes distribution of the mice samples, a KDE plot and a histogram with the count of the samples in each of the automatically defined heterogeneity index value bins by the function were plotted (figure A.7). In both of the execution configurations, the distribution does not follow a normal shape, being skewed to the right and left, respectively.

Normal probability plots were output to confirm the non-normality of the heterogeneity indexes distributions. Also called the *quantile-quantile* plot, they determine whether a specific distribution deviates from normality, comparing the quantiles of a particular distribution against the quantiles of a standard normal distribution. If a distribution follows a normal shape, the points should fall on a straight line [82].

As illustrated by figure A.8, the points do not follow a normal distribution, which means we could not use the T-test to assess whether the THI distributions of the two groups of mice were similar or distinct.

The *Mann-Whitney test*, also known as the Mann-Whitney U test or Wilcoxon rank-sum test, is a non-parametric test, being an alternative to a T-test when the data is not normally distributed. Its null hypothesis assumes that two independent samples come from the same population, meaning they have the same distribution [83].

An alpha of 0.05 (95% of confidence) was fixed and compared with a p-value to accept or reject the null hypothesis. If the p-value was higher than the alpha, the null hypothesis would fail to be rejected, with the data of the two mice groups being assumed to have the same distribution.

On the other hand, if the p-value had a smaller value than or equal to the alpha, we would have rejected the null hypothesis and assumed that the heterogeneity indexes of the two mice groups had different distributions.

Figure 3.2 illustrates a point plot with the average THI value of the samples of each mouse, with the indication of the uncertainty around it using error bars. The first execution has a p-value higher than the alpha value and shows there is not much difference between the two mice groups. However, the plot below has a p-value of 0.012, much smaller than the alpha value.

By analyzing the second plot, we can see that the sunitinib group seems to have a higher heterogeneity index than the control group. Such tells us that the anti-angiogenic treatment might have increased ITH, even though there were more executions showing that the groups followed the same distribution rather than the opposite. This way, it is not easy to make a conclusion that supports either of the arguments.

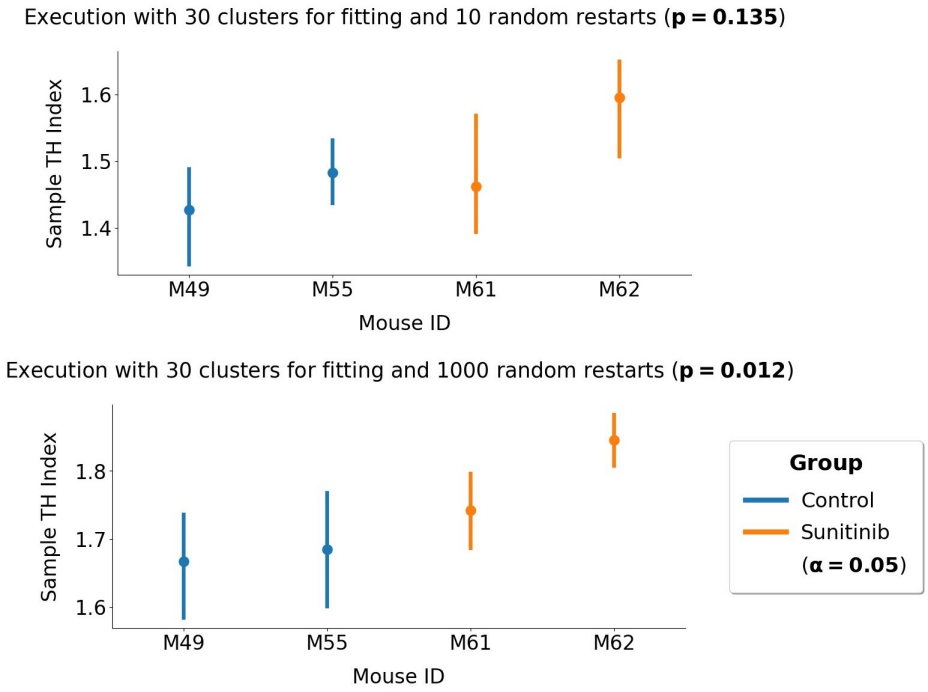

Figure 3.2: Comparison of all the samples of the two experimental groups using point plots with the mean THI values and error bars, including the results of applying the Mann-Whitney test.

3.4 Part I - Canopy

The third and final method applied to estimate the ITH of the mice case study data computes not only the structure of the clonal populations but also the evolutionary phylogeny of a tumor, using both the SNVs and CNAs as input [62]. Canopy is an open-source R package that can be found at this GitHub repository.

Canopy requires as input eight different matrices: the mutant allele read depths and the total coverage of SNVs, the major and minor copy number ratios of the CNAs with their corresponding standard errors, which CNA regions contain each copy number variation (when there are overlapping CNAs), and which SNVs are in CNA regions [62].

The input processing for this algorithm was complex, especially having to handle the overlapping of CNAs. The procedure for computing each of the necessary matrices, the strategy applied to define the CNA regions, how the algorithm was run, and the analysis of the output are described next.

Again, the input was processed using all the mice samples together and for each of the four mice separately. In opposition to PyClone-VI, the CNAs were not just used as filters to select the SNVs and define their allele-specific copy numbers but were also required to run Canopy.

The SNVs and copy number variations were loaded first, keeping only the non-diploid mutations (with major and minor copy numbers different from 1). Similar to PyClone-VI, the variants were inserted with default values for the reference and alternative allele read counts in the samples that they were not part of since the SNVs matrices required that each SNV was in all samples.

It was possible to create the first two matrices having all the mutations in all samples, $R_{S\times N}$ and $X_{S\times N}$, where S are the number of mutated loci and N the number of samples. These contained, respectively, the alternative allele read counts and the total number of reads covering each locus (summing the reference and alternative allele read counts).

Given that there were identified overlapping CNAs in the data, it was necessary to build the matrix $C_{T\times t}$, where T are the number of copy number variation regions and t the number of CNA events.

A manual inspection of the overlapped events, as recommended by the authors [62], led to the computation of an approach to building the CNA regions. If these existed, it would also be needed to identify and manage the nested/intermediary CNAs.

Overlapping CNAs refer to distinct copy number variation events that occur in separate samples but affect the same genomic region, the same chromosome, overlapping across samples [62]. An example can be observed in figure A.9.

Looking at the first diagram, we can see that there are three copy number variation events: E1, E2, and E3. The first and last ones are, respectively, overlapped with and nested in E2. The diagrams (a) and (b) display the identification of each corresponding copy number variation event across all samples.

Diagram (a) illustrates the first approach that was taken to the definition of CNA regions. R_i stands for the CNA region number *i*. Each region would be identified by unique sets of CNA events. R1 would contain only the CNA E2; R2 would have only the event E3; R3 would have two events, E1 and E3; R4 would intersect both E1 and E2; finally; R5 would consist only of E1.

Using this approach, the C matrix would have 5 rows and 3 columns, corresponding to the 5 CNA regions and 3 CNA events identified. The matrix is filled with either 0 or 1, indicating whether a CNA region contains a CNA event or not. For example, the rows of R3 and R4 would have a 1 in columns E1 and E3 and E1 and E2, respectively. This way, E1 would have a 1 in two different rows, or, equivalently, two 1 in its column.

After following the diagram (a) method to build the CNA regions, creating the matrix C, and inputting it to one method of Canopy, there was an interruption of the execution followed by a message. According to the source code of the referred Canopy function: "Matrix C should have one and only one 1 for each column" [62].

The author itself considered that diagram (a) was a valid interpretation of a CNA region [62]. However, at the same time, the algorithm had this limitation, just allowing a CNA event to be contained inside one CNA region. A different methodology was designed to solve this issue, as depicted in diagram (b).

Instead of considering each CNA region as a distinct set of CNAs/intersections of CNA events, it was determined that each region would be the union of all adjacent copy number variation intersections. Such would lead to each CNA event belonging to only a CNA region, with the C matrix only having a 1 in each column.

In particular, the following procedure was applied to build the CNA regions of the mice tumor data. First, the chromosomes that only had a unique copy number variation event that was present in all the samples were identified. This way, these would automatically consist of regions since they were not intersected by any other CNAs.

Then, the nested copy number variation events were detected based on multiple conditions, including having CNAs exactly between two other events, with all of them sharing the same copy number.

Other of the rules would select to join two CNA events if they were less than 10MB away (millions of nucleotide base pairs as a unit) from each other and had the same copy number. The union would result in defining a new copy number variation from the beginning of the event with the smallest genome start position to the ending of the event with the biggest genome end position.

After joining the nested CNAs and other events that verified the conditions to be selected, the copy number variation regions were identified by the union of all the consecutive intersected events in the same chromosome across different samples, thus allowing to fill the matrix C with the information of the events present in each region.

The following four matrices were then generated: the observed major and minor copy numbers in each CNA region (respectively, $\mathrm{WM}_{\mathrm{T\times N}}$ and $\mathrm{Wm}_{\mathrm{T\times N}}$), and their standard errors (accordingly, ${\epsilon^{\mathrm{M}}}_{\mathrm{T\times N}}$ and ${\epsilon^{\mathrm{m}}}_{\mathrm{T\times N}}$).

Each CNA region was intersected with the events that were inside its genome position range to compute the allele-specific copy number matrices. Independently of how many copy number variations there were, the BAF and the depth ratio ($2^{\log_2}$) of the CNA events could be used for the estimation of the region copy numbers.

Based on the formula of the depth ratio, the major copy number can be obtained by the equation 3.2. In turn, given the equation 3.3 of the BAF, by substituting the WM definition, we can then calculate the values of Wm and WM [62]:

$$\text{depth\_ratio} = \frac{WM + Wm}{2} \Leftrightarrow WM = 2 \times \text{depth\_ratio} - Wm \tag{3.2}$$

$$\begin{gathered} BAF = \frac{Wm}{WM + Wm} \Leftrightarrow BAF = \frac{Wm}{(2 \times \text{depth\_ratio} - Wm) + Wm} \Leftrightarrow \\ Wm = 2 \times BAF \times \text{depth\_ratio} \end{gathered} \tag{3.3}$$

One of the details that had to be accounted for when adding the allele-specific copy numbers to each of the matrices were the cases in which BAF > 0.5, with the major copy number having a smaller value than the minor copy number, and thus these would have to be switched before being inserted into the WM and Wm matrices, respectively.

To estimate the standard error matrices of both the major and minor copy numbers, the $\log_2$ upper and lower confidence intervals of each CNA segment were used, whether the region was intersected by only one copy number variation or more.

If a CNA region was intersected by or contained more than one event, a weighted average of the BAFs and depth ratios would be calculated based on the sum of the product of these with the fraction of the region covered by each copy number variation, with the result being divided by the total number of events that intersected a region.

The CNA regions with major and minor copy numbers equal to 0 in at least one of the samples were removed from the rows of WM, Wm, and C matrices, and subsequently all the events that, after removing specific regions, did no longer intercept any region were also deleted from the columns of the matrix C.

The last matrix that needed to be processed for the Canopy input indicated whether each SNV and CNA region overlapped or not: the $Y_{S \times (T+1)}$ matrix, where S is the number of variants and T + 1 is the number of CNA regions with an additional column, named as *non-cna_region*, that identified the SNVs that were not in any CNA region. Before proceeding to its creation, a heatmap of the VAFs of the SNVs was plotted.

Based on the heatmap, following the advice of the authors on how to select the best SNVs to be used, it was evaluated if there were non-informative/redundant mutations [62], with similar VAF values in most of the samples; each mutation also had to be present in at least 6 of the 16 mice samples. By removing some variants, the execution time of Canopy was reduced. The deleted SNVs were also cut from the R and X matrices.

Finally, using the kept SNVs, these were intersected with the CNA regions to fill the overlapping information in the Y matrix. If a mutation was contained inside a copy number variation region, a 1 was assigned to the corresponding SNV row in the CNA region column, otherwise being 0. If a variant was not in any region, a 1 was inserted into the non-cna_region column for the row referring to that SNV.

All the eight matrices, together with a user-defined project name, were input into the MCMC sampling of Canopy through the *canopy.sample* method. It is the major function of Canopy, sampling the posterior tree configurations [62].

The MCMC sampling is the most computationally expensive method in Canopy. This step can be highly time-consuming, requiring a day or more to obtain results depending on the size of the input [62].

It receives as input an attribute K, a list of the assumed possible number of populations (the smaller number of populations allowed is 3, with K - 1 tumor clones and one normal population) [62]. Since PyClone-VI determined between 3 and 6 clones when using the different input alternatives, the parameter was defined with this range of values.

Moreover, it demands the *numchain* argument, the number of MCMC chains with random initiations to be used, as well as the minimum and the maximum number of iterations for each chain (*min.simrun* and *max.simrun*). The numchain should have a value not too large to speed up the MCMC computation but also not too small to allow the chains to converge to a specific posterior probability distribution.

The method returns the found tree configurations, modeled as lists, with the mutation assignments to each clone and the relationships between the populations [26, 62], which may correctly model the true tumor phylogeny for each of the different numbers of clones provided in the range parameter.

It also outputs a *Portable Document Format* (PDF) file for check of convergence, with the plots of the posterior likelihoods and acceptance rates in each of the subtree spaces generated for each of the numbers of clones provided [62]. An example of the posterior likelihood and acceptance rate plots using 5 clones can be seen in figure A.10.

After more than 10000 iterations, the different number of chains converged to a *stationary distribution*, which means that all the simulated samples after it could be considered as a sample of the posterior distribution of the clonal populations [84].

The acceptance rate remained low after the initial rounds. Such indicates that the chains got stuck to just some points in the parameter space of the sampled posterior distributions, needing many iterations to make one jump [84].

To determine the optimal number of clones, the *canopy.BIC* method received the output of the MCMC sampling step. Based on the mean likelihood of the reconstruction configurations with the different populations, it compared and selected the number of clones with the highest BIC value [26, 62].

Two of the canopy.BIC parameters that are required to be specified are the *burn-in* and *thinning* of the MCMC chains. The posterior likelihood plot can be analyzed to wisely choose the value of these arguments, as one should guarantee that there are any posterior trees obtained left to be evaluated by the next method [62].

The burn-in, also known as warmup, are the iterations spent by the Markov chains to reach the stationary distribution, which should be removed since they are not considered samples of the posterior distribution [84]. The thinning interval specifies that only every *i*th sample is saved, accounting, for example, for not enough storage space or compression being required [62, 84].

The *canopy.post* function is then run, receiving as input the best number of tumor populations chosen, proceeding to the posterior evaluation of all sampled trees by MCMC for that number of clones [62]. It returns a table with the posterior probability in the entire tree space (comparing between the tree spaces of the different configurations) and the mean posterior likelihood of each configuration [62].

Canopy computes the mean posterior likelihood for the subtree space of each configuration. Even though the trees have the same structure/configuration (the evolutionary relationships between the populations are the same), the VAFs of the mutations of each clone can be distinct, resulting in slightly different likelihoods [62].

In the end, receiving as input the configuration with the highest mean posterior likelihood, the *canopy.output* method selected the tree within it with the highest posterior likelihood. By using the *canopy.plottree* function, the tree structure and a table with the population clonal frequencies in each of the samples of the two mice groups were plotted, as represented by figure 3.3.

Each of the mut_i denotes the mutational profile (SNV and CNA occurrences) of each tumor population that existed, which led to new clonal expansions. They were obtained from the posterior distributions of the tree space of the configuration chosen [62].

The list of the different mutations in each population throughout the tumor evolution, including the order in which the SNVs and CNAs appeared (for example, if there are relevant loss of heterozygosity events), can help to identify specific treatment targets [62]. However, that goes beyond the scope of this thesis.

Following the PyClone-VI analysis of the results, the THI was computed for each of the mice samples based on the clonal frequencies and it was plotted the mean THI of each mouse with confidence intervals of 95% (figure A.11). Once again, the values of the sunitinib group seem to be slightly higher, indicating that the anti-angiogenic treatment increased ITH. Nonetheless, the results are not unequivocal.

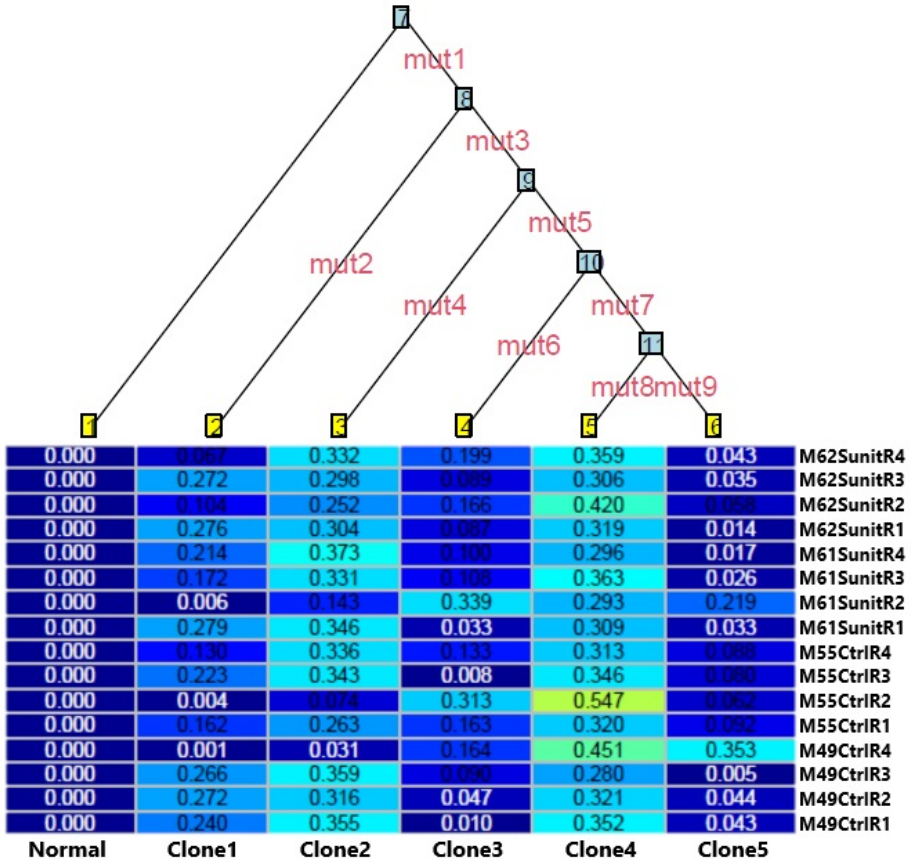

Figure 3.3: Tree with the highest posterior likelihood and a table with the clonal frequencies of each population in each sample, using all mice samples together as input. Adapted from the output of Canopy, with all the labels fitting the plot.

3.5 Part II - Genetic information feature analysis

In the first part, different heterogeneity estimation methods were used to calculate the ITH of the mice case study data, allowing us to compare the heterogeneity values of the two groups and assess whether the angiogenesis inhibitor drug induced the differentiation of the tumor cells in the treated specimen.

The results obtained by these techniques were not consistent enough to be able to declare that either of the outcomes was verified. There may also have been estimation errors within the methods. Some could not identify evident differences between the two groups, while other approaches inferred that the treatment had some effect.

Given these incoherent conclusions, it were explored better ways to try to distinguish between the mice. Instead of just focusing the analysis on the heterogeneity, other data features were investigated to find how the samples of the two groups were different from one another and, if they were, what those differences were.

Machine learning methods were used for this task. Compared to the ITH computational models, they have the advantage of abstracting the biological meaning, the details associated with the data. Based on these, one may reach more solid conclusions and information about other characteristics rather than the heterogeneity, for example, the different mutated genes.

The tools used in the first part of the thesis work can inform us about the relationship between the different clones and the frequencies of the different mutations, but do not tell us anything about the genes. What is the function of a gene is not relevant for the machine learning techniques since the goal is to study how the different genes affected in each mice group may distinguish these.

This way, the machine learning techniques can be used to complement the analysis done towards the heterogeneity to determine, in generic genetic terms, if the mice are different or not.

Considering that there is a huge number of genes, unsupervised learning, especially dimensionality reduction, can be used for this type of problem in which there are a lot of features. In the beginning, there was no algorithm that was known to be the best to be used, as it all depended on the input characteristics. Multiple alternatives were tested to choose which ones were better for the visualization of the data.

The small size of the mice dataset was one of the problems that needed to be addressed. Since there were only 16 samples, it was necessary to enrich it with data from the outside to include more examples, specifically, with the data of TCGA patients. Given that unsupervised learning methods are used to discover patterns in unlabelled datasets, we can, for example, input different types of data to assess if these have any relation.

The processing of the input for the dimensionality reduction algorithms started with the loading of the table with the information of all the genes. Those that had more than one identification symbol were removed, as each gene should have one and one only identifier. Both the mice and TCGA data had a common subset of these genes.

The CNAs from the germline files of the mice data were intersected with the genes to identify which copy number variations impacted each of these. One condition had to be verified by the genes: they needed to have a CNA in every sample, independently of having or not SNVs in all of them. Their total copy numbers were encoded by representative values: 0, -1, or 1, if it was equal to 2, less than 2, or more than 2, respectively.

The SNV files of the mice data with the genes annotations were then input. Similar to the rule that was imposed on the CNAs, the genes that did not have copy number variations in all samples, whether they were affected by SNVs or not, were deleted. The CNAs of the genes were joined with the somatic variants to assess the mutations in non-diploid copy number regions in each of the genes.

Having all the information about the SNVs and CNAs that affected the genes, it was created the table for the mice data to be input to the dimensionality reduction methods. The same format was used for both the mice and TCGA data. The genes were the features, the columns (names ending with "_cn" and "_sm"), filled with 0s and 1s, indicating, respectively, the absence or presence of a CNA or a SNV on a gene.

The instances, the rows, were the mice samples. Each point was represented by a very high dimensional tuple, with the information of the genes ordered equally for all, in which the coordinates of the instances were the presence or not of variants and copy number alternations in each of the genes. As an alternative, by doing the matrix transpose, it would also have been possible to project the genes instead of the samples.

Concerning TCGA data needed for the input, first were loaded the files with the representative values of the total copy numbers (0, -1, and 1) of the CNA sample identifiers from each patient in each of the gene symbols. Next, these tables were joined with the information of the genes to get the names associated with each of the symbols, cutting those genes with more than one entry, that is, multiple identifiers.

Afterward, the files with the information of the CNAs of each patient were loaded and intersected with the transpose of the matrices of the representative values of the CNA sample identifiers from each patient to have all the information in one table.

With all the copy number alteration features in one table for each of the patient samples, it was detected that some patients had more than one CNA sample. To aggregate the CNA samples of a patient, the representative values of all its CNA samples were summed, with the representative value transformation being again applied to the result so that the values used to codify the total copy number were only -1, 0, and 1.

Finally, the SNVs of TCGA patients were input. Based on these, the patients and the genes that had variants in a sample but did not have CNAs in it were removed. A table with the indexes of each gene was joined with the SNVs of TCGA patients to find which genes were mutated by which variants for each of the patients.

One task was left to do to finish the input processing. To use both the mice samples and TCGA patients together in a table with the genes as features these had to have the same genes to guarantee a consistent format for all the tuples. Thus, only the genes that were shared by both of them were used to generate the last matrix.

The three matrices were input to the dimensionality reduction techniques, testing the gene features by comparing the mice alone, TCGA patients alone, and both of them together. It was important to experiment with the different alternatives to see how these behaved when projected together.

Several methods were tested to represent the different high-dimensional arrays, including different values for the parameters of each. The procedure to run all the methods with the mice sample genes data was the same. The matrices were converted to arrays of values and fed to the algorithms. Then, a column with the experimental group of each data point was added to the output for visualization.

The dimensionality reduction techniques applied to the mice genes data alone were the PCA, the truncated SVD, the t-SNE, the PCA to 16 dimensions followed by the t-SNE, the MDS, the PCA to 16 dimensions followed by the MDS, the Isomap, and the PCA to 6 dimensions followed by the Isomap, in both two and three dimensions.

The use of PCA before the methods t-SNE, MDS and Isomap can be justified by the longer execution times of these used alone, especially the last one, to speed up the algorithms [72] and compare between the plots output. For the same reason, truncated SVD was preferred over the PCA with a *"svd_solver"* argument having a value of *"full"*, and reducing to 6 dimensions instead of 16 when using the PCA followed by the Isomap.

Given that it would not be possible to show the output of all the algorithms running for all the three input arrays used, these are discussed, in general, with some examples, and the conclusions obtained are compared.

When using the mice samples for all the dimensionality reduction methods, the two experimental groups did seem to have some separation, even though points from both classes could be seen near the instances of the other. The control instances were more concentrated in one spot, while the treated points were more dispersed.

The algorithm combination that showed a more clear division of the two mice groups was the application of the PCA to 16 dimensions and the result being used as input to the t-SNE down to 2 dimensions (figure A.12).

The control samples were closer to each other and more distant from the sunitinib group, and vice versa, even though one sample of each class was still found on the wrong side and there was a big dispersion inside the two clusters found.

Such slightly indicates that there may be a difference between the two groups. However, other methods achieved plots where the points from the different clusters were closer to one another and even more dispersed, as can be observed from the results of applying MDS to 2 dimensions (figure A.13).

Next, the dimensionality reduction methods were applied to the gene features of TCGA patients. After inputting the data into the algorithms and getting the output, it was inserted a column with the tumor type of each instance for visualization.

Considering the big amount of data and the different plots generated for each technique, five algorithms were applied: the PCA, the truncated SVD, the PCA down to 25 dimensions followed by the t-SNE, the PCA down to 25 dimensions followed by the MDS,

and the PCA down to 25 dimensions followed by the Isomap. Figure A.14 has the legend of the tumor types of the plots with TCGA patients data.

The output of applying the PCA to 25 dimensions followed by the t-SNE to two dimensions using TCGA patients data can be seen in figure A.15. Most of the points of the different tumor types are overlapped, with some instances of the same cancers presenting a wider dispersion.

The algorithm identified associations between related tumors: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), and also uterine carcinosarcoma (UCS) and uterine corpus endometrial carcinoma (UCEC), with similar points being closer or following a similar pattern, as should be expected.

The data agglomeration in a big cloud of points without a structure makes the analysis more difficult. Kernel density estimation was applied to have an idea of where the instances were more and less concentrated.

The density of the different regions is represented using a color map, with darker colors indicating a lower density and lighter colors standing for a higher density (figure A.16). A closer view of the densities was plotted using contour lines with a heatmap, as depicted by figure A.17.

Finally, the mice samples and TCGA patients data were used together as input, with the mice points being represented on top of the instances of patients, with some transparency applied to these for better visualization (for example, figure 3.4). Again, the different methods overlapped the two mice group points, suggesting that the treatment did not have any particular consistent effect.

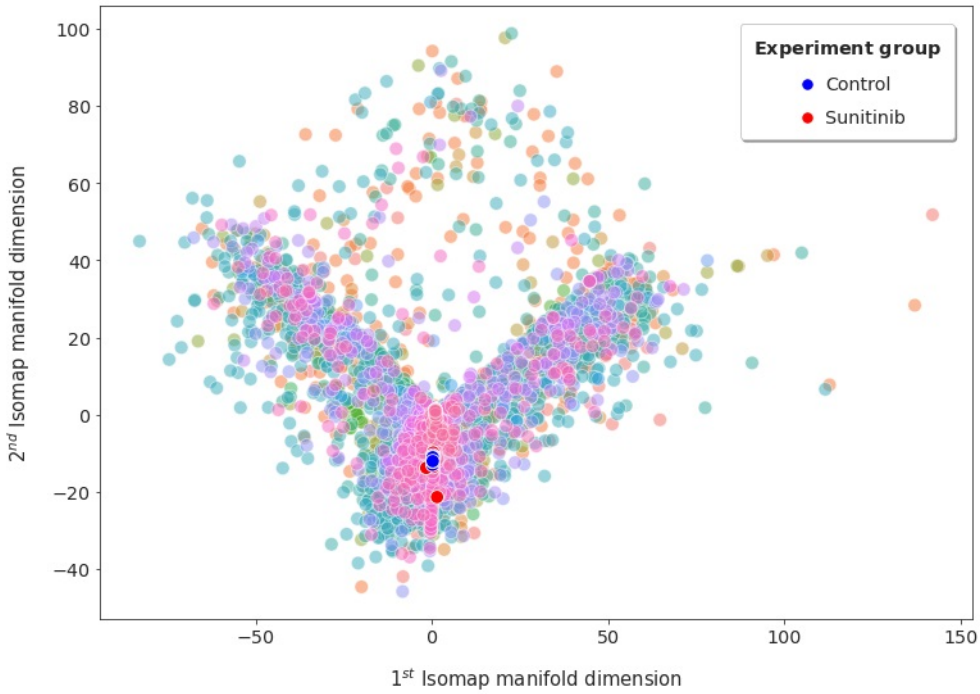

Figure 3.4: PCA to 25 dimensions followed by the Isomap down to 2 dimensions using together the mice samples and the tumor data of the patients stored in TCGA.

Chapter 4

Conclusion

This dissertation investigated the effect that the inhibition of angiogenesis can have in the ITH, based on a specific experiment with mice. Before the work development stage, as it was summarized in the introduction, extensive research on the topic background was carried out.

Chapter 1 started by describing how tumors proliferate; highlighted the role of next-generation sequencing technologies in the study of cancer evolution; defined the ITH estimation-related concepts; approached angiogenesis and its adaptation by tumors; discussed the benefits and drawbacks of anti-angiogenic therapies; in the end, it presented the mice case study and the dissertation goals.

Two characteristics of the tumor sequencing data were studied to distinguish between the mice of the two groups (control and treatment): the heterogeneity and the mutational profiles of the genes. In chapter 2, three classes of methods that are applied to compute ITH were outlined, emphasizing the most popular techniques. The two categories of dimensionality reduction tools were then reviewed.

The two parts of the practical work of the thesis, including the input processing, the specification of the algorithms and their execution, the output, and the generation of plots to analyze the results obtained, were detailed in chapter 3.

In line with previous studies [15, 54], the methods explored to calculate ITH exhibited inconsistent results, including reaching distinct conclusions when using the same technique with different parameter values. Such can be explained by the challenges in quantifying heterogeneity [31] and the quality of the data used in this project.

One strength of the mice case study was the use of mouse xenograft models with human cancer cell lines, which facilitated the computation of the heterogeneity by mitigating the need to estimate the tumor purity. On the other hand, there were also some limitations related to the mice tumor region sampling and sequencing.

Only four sections of each tumor were extracted, restricting the number of examples to be input into the methods. Moreover, the low depth (20×) applied in the multi-region WES performed, together with the higher noise that is associated with it [25, 32], could contribute to the detection of false-positive mutations.

Incoherent results were also obtained using dimensionality reduction methods to analyze if the mutational profiles of the genes could distinguish between the mice. Adding TCGA patients to observe, with more examples, the positioning of the data points of the two mice groups on the manifold did not change the conclusions.

It can be assumed that, in both parts of the work, the algorithms did not find an explicit distinction between the two experimental groups. This way, relying only on the outputs of these methods, we would conclude that the inhibition of the angiogenesis did not increase the heterogeneity in the tumors of the treated mice (it did not trigger the selection of partial or fully resistant tumor cells to therapy).

One question can arise from this: had the outcome been different if other techniques had been chosen? Naturally, there were other options to approach this task, but, unfortunately, these could not all be covered because of limited time. Thus, some future work directions can be considered.

Concerning the estimation of the ITH, more recent computational models, for example, *HATCHet* [85] and *CloneSig* [86], offer new perspectives towards the inference of SNVs [86] and CNAs [85], introducing algorithmic innovations that address the limitations of previous methods and surpass their performances [85, 86].

Regarding analyzing the gene mutational profiles to find differences between the mice (if they exist and which are they), rather than using their variants and copy number alterations, we could have grouped the genes according to specific characteristics. For instance, the metabolic processes in which the genes participate or whether they are impacted by a mutation that alters protein sequences.

Instead of applying conventional dimensionality reduction methods, other more powerful tools from *deep learning* could have been tested, for example, *autoencoders*. The *neural networks* would learn a manifold representation, being trained with TCGA data. Then, this model could be applied to the mice data to see how the points of the two groups would lay out on the manifold learned.

Furthermore, the neural network could also be trained with simulated frequency profiles of mutations to solve the inverse problem of estimating the tumor clones. Knowing the composition of the tumor populations, one can infer the frequency distribution of the different variants. However, as we have seen, performing clonal deconvolution from the frequencies of mutations is a complex task.

A simulator would have to be trained with real data of mutation frequency profiles to achieve this, labeled with the corresponding tumor clonal compositions. Based on the examples of frequency distributions of variants generated by the simulator, the neural network would learn how to compute the composition of the clones for any input, and then the model would be applied to the mice data.

Bibliography

[1] M. A. Zaimy et al. "New methods in the diagnosis of cancer and gene therapy of cancer based on nanoparticles". In: *Cancer gene therapy* 24.6 (June 2017), pp. 233–243. ISSN: 1476-5500. DOI: 10.1038/cgt.2017.16. URL: https://www.nature.com/articles/cgt201716 (cit. on pp. 1, 2).

[2] N. McGranahan and C. Swanton. "Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future". In: *Cell* 168.4 (Feb. 2017), pp. 613–628. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2017.01.018. URL: https://www.sciencedirect.com/science/article/pii/S0092867417300661 (cit. on p. 1).

[3] C. R. Hanna, K. A. Boyd, and R. J. Jones. "Evaluating cancer research impact: lessons and examples from existing reviews on approaches to research impact assessment". In: *Health Research Policy and Systems* 19.1 (Mar. 2021). DOI: 10.1186/s12961-020-00658-x. URL: https://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-020-00658-x (cit. on p. 1).

[4] Global Burden of Disease Cancer Collaboration. "The Global Burden of Cancer 2013". In: *JAMA Oncology* 1.4 (July 2015), pp. 505–527. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2015.0735. URL: https://jamanetwork.com/journals/jamaoncology/fullarticle/2294966 (cit. on p. 1).

[5] W. Kibbe, J. Klemm, and J. Quackenbush. "Cancer Informatics: New Tools for a Data-Driven Age in Cancer Research". In: *Cancer Research* 77.21 (Oct. 2017), pp. e1–e2. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-17-2212. URL: https://cancerres.aacrjournals.org/content/77/21/e1.long (cit. on pp. 1, 2, 10, 16, 25).

[6] R. E. Amor et al. "Breath analysis of cancer in the present and the future". In: *European Respiratory Review* 28.152 (June 2019). DOI: 10.1183/16000617.0002-2019. URL: https://err.ersjournals.com/content/28/152/190002 (cit. on p. 1).

[7] H. Varmus and H. Kumar. "Addressing the Growing International Challenge of Cancer: A Multinational Perspective". In: *Science translational medicine* 5.175 (Mar. 2013). DOI: 10.1126/scitranslmed.3005899. URL: https://www.science.org/doi/10.1126/scitranslmed.3005899 (cit. on p. 1).

[8] T. Helleday. "Chemotherapy-induced toxicity—a secondary effect caused by released DNA?" In: *Annals of Oncology* 28.9 (Sept. 2017), pp. 2054–2055. ISSN: 0923-7534. DOI: 10.1093/annonc/mdx349. URL: https://www.sciencedirect.com/science/article/pii/S0923753419352688 (cit. on p. 1).

[9] N. A. Seebacher et al. "Clinical development of targeted and immune based anti-cancer therapies". In: *Journal of Experimental & Clinical Cancer Research* 38.1 (Apr. 2019). DOI: 10.1186/s13046-019-1094-2. URL: https://jeccr.biomedcentral.com/articles/10.1186/s13046-019-1094-2 (cit. on p. 1).

[10] L. Zhong et al. "Small molecules in targeted cancer therapy: advances, challenges, and future perspectives". In: *Signal Transduction and Targeted Therapy* 6.1 (May 2021). DOI: 10.1038/s41392-021-00572-w. URL: https://www.nature.com/articles/s41392-021-00572-w (cit. on p. 1).

[11] P. R. Prasetyanti and J. P. Medema. "Intra-tumor heterogeneity from a cancer stem cell perspective". In: *Molecular Cancer* 16.1 (Feb. 2017). DOI: 10.1186/s12943-017-0600-4. URL: https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-017-0600-4 (cit. on pp. 1, 4, 11).

[12] A. E. Kersh et al. "Targeted Therapies: Immunologic Effects and Potential Applications Outside of Cancer". In: *The Journal of Clinical Pharmacology* 58.1 (Jan. 2018), pp. 7–24. ISSN: 0091-2700. DOI: https://doi.org/10.1002/jcph.1028. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5972536/ (cit. on p. 1).

[13] M. Baliu-Piqué, A. Pandiella, and A. Ocana. "Breast Cancer Heterogeneity and Response to Novel Therapeutics". In: *Cancers* 12.11 (Nov. 2020). DOI: 10.3390/cancers12113271. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7694303/ (cit. on pp. 2, 14).

[14] M. Greaves. "Evolutionary Determinants of Cancer". In: *Cancer Discovery* 5.8 (Aug. 2015), pp. 806–820. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-15-0439. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4539576/ (cit. on pp. 2, 4, 5).

[15] J. Abécassis et al. "Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data". In: *PLOS ONE* 14.11 (Nov. 2019). DOI: 10.1371/journal.pone.0224143. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224143 (cit. on pp. 2, 4–8, 17, 21, 23, 49).

[16] F. Vandin. "Computational Methods for Characterizing Cancer Mutational Heterogeneity". In: *Frontiers in genetics* 8 (June 2017). DOI: 10.3389/fgene.2017.00083. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5469877/ (cit. on pp. 2, 5, 7–10, 16, 18, 19, 21–24).

[17] J. M. Heather and B. Chain. "The sequence of sequencers: The history of sequencing DNA". In: *Genomics* 107.1 (Jan. 2016), pp. 1–8. ISSN: 1089-8646. DOI: 10.1016/j.ygeno.2015.11.003. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4727787/ (cit. on p. 2).

[18] X. Jiang et al. "Computational Advances in Cancer Informatics (A)". In: *Cancer informatics* 13.Suppl 1 (Oct. 2014), pp. 45–48. ISSN: 1176-9351. DOI: 10.4137/CIN.S19243. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4216040/ (cit. on p. 2).

[19] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". In: *Contemporary Oncology (Poznan, Poland)* 19.1A (Jan. 2015), A68–A77. ISSN: 1428-2526. DOI: 10.5114/wo.2014.47136. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4322527/ (cit. on p. 2).

[20] Z. Wang, M. A. Jensen, and J. C. Zenklusen. "A Practical Guide to The Cancer Genome Atlas (TCGA)". In: *Statistical Genomics: Methods and Protocols*. Ed. by E. Mathé and S. Davis. 1st ed. Springer New York, Mar. 2016. Chap. 6, pp. 111–141. ISBN: 978-1-4939-3578-9. DOI: 10.1007/978-1-4939-3578-9_6. URL: https://doi.org/10.1007/978-1-4939-3578-9_6 (cit. on p. 2).

[21] N. Auslander, A. B. Gussow, and E. V. Koonin. "Incorporating Machine Learning into Established Bioinformatics Frameworks". In: *International journal of molecular sciences* 22.6 (Mar. 2021). DOI: 10.3390/ijms22062903. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8000113/ (cit. on pp. 2, 25).

[22] S. M. D. A. C. Jayatilake and G. U. Ganegoda. "Involvement of Machine Learning Tools in Healthcare Decision Making". In: *Journal of healthcare engineering* 2021 (Jan. 2021). ISSN: 2040-2309. DOI: 10.1155/2021/6679512. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7857908/ (cit. on pp. 2, 25).

[23] A. Glielmo et al. "Unsupervised Learning Methods for Molecular Simulation Data". In: *Chemical reviews* 121.16 (Aug. 2021), pp. 9722–9758. ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.0c01195. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8391792/ (cit. on pp. 2, 25–28).

[24] Z. Qu et al. "Visual Analytics of Genomic and Cancer Data: A Systematic Review". In: *Cancer informatics* 18 (Mar. 2019). ISSN: 1176-9351. DOI: 10.1177/1176935119835546. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6416684/ (cit. on p. 2).

[25] H. Chen et al. "ALLELE-SPECIFIC COPY NUMBER ESTIMATION BY WHOLE EXOME SEQUENCING". In: *The Annals of Applied Statistics* 11.2 (June 2017), pp. 1169–1192. ISSN: 1932-6157. DOI: 10.1214/17-AOAS1043. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5627665/ (cit. on pp. 3, 6–8, 50).

[26] L. K. Sundermann. "Lineage-Based Subclonal Reconstruction of Cancer Samples". PhD thesis. Universität Bielefeld, Jan. 2019. DOI: 10.4119/unibi/2935248. URL: https://pub.uni-bielefeld.de/record/2935248 (cit. on pp. 3, 5–8, 10, 22–24, 43).

[27] M. J. Williams, A. Sottoriva, and T. A. Graham. "Measuring Clonal Evolution in Cancer with Genomics". In: *Annual Review of Genomics and Human Genetics* 20 (Aug. 2019), pp. 309–329. ISSN: 1527-8204. DOI: 10.1146/annurev-genom-083117-021712. URL: https://www.annualreviews.org/doi/10.1146/annurev-genom-083117-021712 (cit. on pp. 3–6, 21).

[28] M. Tarabichi et al. "A practical guide to cancer subclonal reconstruction from DNA sequencing". In: *Nature Methods* 18.2 (Jan. 2021), pp. 144–155. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01013-2. URL: https://doi.org/10.1038/s41592-020-01013-2 (cit. on pp. 4–9, 20).

[29] J. Brosnan and C. Iacobuzio-Donahue. "A new branch on the tree: next-generation sequencing in the study of cancer evolution". In: *Seminars in cell & developmental biology* 23.2 (Apr. 2012), pp. 237–242. ISSN: 1096-3634. DOI: 10.1016/j.semcdb.2011.12.008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3314157/ (cit. on p. 4).

[30] S. Dentro, D. Wedge, and P. Loo. "Principles of Reconstructing the Subclonal Architecture of Cancers". In: *Cold Spring Harbor perspectives in medicine* 7.8 (Aug. 2017). DOI: 10.1101/cshperspect.a026625. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5538405/ (cit. on pp. 4–9, 18).

[31] Z. Hu, R. Sun, and C. Curtis. "A population genetics perspective on the determinants of intra-tumor heterogeneity". In: *Biochimica et biophysica acta. Reviews on cancer* 1867.2 (Apr. 2017), pp. 109–126. ISSN: 0304-419X. DOI: 10.1016/j.bbcan.2017.03.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5623098/ (cit. on pp. 4, 5, 7, 12–14, 17–19, 22, 49).

[32] A. Alizadeh et al. "Toward understanding and exploiting tumor heterogeneity". In: *Nature medicine* 21.8 (Aug. 2015), pp. 846–853. ISSN: 1546-170X. DOI: 10.1038/nm.3915. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4785013/ (cit. on pp. 4, 50).

[33] K. Sprouffske et al. "Genetic heterogeneity and clonal evolution during metastasis in breast cancer patient-derived tumor xenograft models". In: *Computational and structural biotechnology journal* 18 (Jan. 2020), pp. 323–331. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2020.01.008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7026725/ (cit. on pp. 4, 8, 12, 14).

[34] E. A. Mroz et al. "Intra-tumor Genetic Heterogeneity and Mortality in Head and Neck Cancer: Analysis of Data from The Cancer Genome Atlas". In: *PLOS Medicine* 12.2 (Feb. 2015). DOI: 10.1371/journal.pmed.1001786. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4323109/ (cit. on pp. 4, 5, 17).

[35] M. W. Fittall and P. V. Loo. "Translating insights into tumor evolution to clinical practice: promises and challenges". In: *Genome Medicine* 11.1 (Mar. 2019). DOI: 10.1186/s13073-019-0632-z. URL: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0632-z (cit. on pp. 5, 7).

[36] D. Sims et al. "Sequencing depth and coverage: key considerations in genomic analyses". In: *Nature Reviews Genetics* 15.2 (Jan. 2014), pp. 121–132. ISSN: 1471-0064. DOI: 10.1038/nrg3642. URL: https://www.nature.com/articles/nrg3642 (cit. on pp. 6, 32, 33).

[37] E. Talevich et al. "CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing". In: *PLOS Computational Biology* 12.4 (Apr. 2016). DOI: 10.1371/journal.pcbi.1004873. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004873 (cit. on pp. 8, 14, 30).

[38] R. I. Teleanu et al. "Tumor Angiogenesis and Anti-Angiogenic Strategies for Cancer Treatment". In: *Journal of clinical medicine* 9.1 (Dec. 2019). DOI: 10.3390/jcm9010084. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7020037/ (cit. on pp. 10–14).

[39] P. Carmeliet and R. K. Jain. "Molecular mechanisms and clinical applications of angiogenesis". In: *Nature* 473.7347 (May 2011), pp. 298–307. ISSN: 1476-4687. DOI: 10.1038/nature10144. URL: https://www.nature.com/articles/nature10144 (cit. on pp. 10, 11, 14).

[40] C. Viallard and B. Larrivée. "Tumor angiogenesis and vascular normalization: alternative therapeutic targets". In: *Angiogenesis* 20.4 (June 2017), pp. 409–426. ISSN: 1573-7209. DOI: 10.1007/s10456-017-9562-9. URL: https://link.springer.com/article/10.1007/s10456-017-9562-9 (cit. on pp. 10–12).

[41] I. Zuazo-Gaztelu and O. Casanovas. "Mechanisms of Tumor Angiogenesis". In: *Tumor Angiogenesis: A Key Target for Cancer Therapy*. Ed. by D. Marmé. 1st ed. Springer International Publishing, Apr. 2019. Chap. 1, pp. 3–31. ISBN: 978-3-319-33673-2. DOI: 10.1007/978-3-319-33673-2_1. URL: https://doi.org/10.1007/978-3-319-33673-2_1 (cit. on pp. 10–12).

[42] E. Lampri and E. Ioachim. "Angiogenesis: Something Old, Something New". In: *Angiogenesis: Insights from a Systematic Overview*. Ed. by G. Santulli. 1st ed. Nova Science Publishers, Incorporated, July 2013. Chap. 1, pp. 1–30. ISBN: 978-1-62618-114-4 (cit. on p. 11).

[43] E. Maj, D. Papiernik, and J. Wietrzyk. "Antiangiogenic cancer treatment: The great discovery and greater complexity (Review)". In: *Int J Oncol* 49.5 (Nov. 2016), pp. 1773–1784. ISSN: 1791-2423. DOI: 10.3892/ijo.2016.3709. URL: https://www.spandidos-publications.com/10.3892/ijo.2016.3709#9 (cit. on pp. 12, 13).

[44] R. J. Bodnar. "Anti-Angiogenic Drugs: Involvement in Cutaneous Side Effects and Wound-Healing Complication". In: *Advances in Wound Care* 3.10 (Oct. 2014), pp. 635–646. ISSN: 2162-1918. DOI: 10.1089/wound.2013.0496. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4183909/ (cit. on p. 12).

[45] J.-S. Park, I. Park, and G. Y. Koh. "Benefits and Pitfalls of Tumor Vessel Normalization". In: *Tumor Angiogenesis: A Key Target for Cancer Therapy*. Ed. by D. Marmé. 1st ed. Springer International Publishing, Apr. 2019. Chap. 3, pp. 51–71. ISBN: 978-3-319-33673-2. DOI: 10.1007/978-3-319-33673-2_46. URL: https://doi.org/10.1007/978-3-319-33673-2_46 (cit. on pp. 13, 14).

[46] A. Marusyk, M. Janiszewska, and K. Polyak. "Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance". In: *Cancer cell* 37.4 (Apr. 2020), pp. 471–484. ISSN: 1535-6108. DOI: 10.1016/j.ccell.2020.03.007. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7181408/ (cit. on p. 13).

[47] A. B. Kunnumakkara et al. "Cancer drug development: The missing links". In: *Experimental Biology and Medicine* 244.8 (Apr. 2019), pp. 663–689. ISSN: 1535-3702. DOI: 10.1177/1535370219839163. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6552400/ (cit. on p. 14).

[48] D. Benjamin et al. "Calling Somatic SNVs and Indels with Mutect2". In: *bioRxiv* (Dec. 2019). DOI: 10.1101/861054. URL: https://www.biorxiv.org/content/10.1101/861054v1.full (cit. on p. 14).

[49] W. Lu et al. "Patient-derived xenograft models in musculoskeletal malignancies". In: *Journal of Translational Medicine* 16 (Apr. 2018). DOI: 10.1186/s12967-018-1487-6. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5913806/ (cit. on p. 15).

[50] T. Goto. "Patient-Derived Tumor Xenograft Models: Toward the Establishment of Precision Cancer Medicine". In: *Journal of personalized medicine* 10.3 (July 2020). DOI: 10.3390/jpm10030064. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7565668/ (cit. on p. 15).

[51] K. M. Hardiman et al. "Intra-tumor genetic heterogeneity in rectal cancer". In: *Laboratory Investigation* 96.1 (Jan. 2016), pp. 4–15. ISSN: 1530-0307. DOI: 10.1038/labinvest.2015.131. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4695247/ (cit. on p. 17).

[52] A. Gough et al. "Biologically Relevant Heterogeneity: Metrics and Practical Insights". In: *SLAS discovery* 22.3 (Mar. 2017), pp. 213–237. ISSN: 2472-5560. DOI: 10.1177/2472555216682725. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5464733/ (cit. on pp. 17, 18, 28).

[53] A. G. Deshwar et al. "PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors". In: *Genome Biology* 16.1 (Feb. 2015). DOI: 10.1186/s13059-015-0602-8. URL: https://doi.org/10.1186/s13059-015-0602-8 (cit. on pp. 18, 22, 23).

[54] W. M. Ismail, E. Nzabarushimana, and H. Tang. "Algorithmic approaches to clonal reconstruction in heterogeneous cell populations". In: *Quantitative biology* 7.4 (Dec. 2019), pp. 255–265. ISSN: 2095-4697. DOI: 10.1007/s40484-019-0188-3. URL: https://doi.org/10.1007/s40484-019-0188-3 (cit. on pp. 18–20, 22–24, 49).

[55] S. Gillis and A. Roth. "PyClone-VI: scalable inference of clonal population structures using whole genome data". In: *BMC Bioinformatics* 21 (Dec. 2020). DOI: 10.1186/s12859-020-03919-2. URL: https://doi.org/10.1186/s12859-020-03919-2 (cit. on pp. 19–21, 23, 24, 35, 36).

[56] R. Schwartz and A. A. Schäffer. "The evolution of tumour phylogenetics: principles and practice". In: *Nature Reviews Genetics* 18.4 (Feb. 2017), pp. 213–229. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.170. URL: https://doi.org/10.1038/nrg.2016.170 (cit. on pp. 19, 20, 22–24).

[57] L. Oesper, A. Mahmoody, and B. J. Raphael. "THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data". In: *Genome biology* 14.7 (July 2013). DOI: 10.1186/gb-2013-14-7-r80. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4054893/ (cit. on pp. 19, 22).

[58] G. Ha et al. "TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data". In: *Genome research* 24.11 (Nov. 2014), pp. 1881–1893. ISSN: 1549-5469. DOI: 10.1101/gr.180281.114. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4216928/ (cit. on pp. 19, 22).

[59] H. Zare et al. "Inferring Clonal Composition from Multiple Sections of a Breast Cancer". In: *PLOS Computational Biology* 10.7 (July 2014). DOI: 10.1371/journal.pcbi.1003703. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003703 (cit. on pp. 19, 20, 22).

[60] P. Deveau et al. "QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction". In: *Bioinformatics* 34.11 (June 2018), pp. 1808–1816. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty016. URL: https://doi.org/10.1093/bioinformatics/bty016 (cit. on pp. 19, 20, 22).

[61] A. Roth et al. "PyClone: statistical inference of clonal population structure in cancer". In: *Nature methods* 11.4 (Apr. 2014), pp. 396–398. ISSN: 1548-7105. DOI: 10.1038/nmeth.2883. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4864026/ (cit. on pp. 19, 21).

[62] Y. Jiang et al. "Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.37 (Aug. 2016), E5528–E5537. ISSN: 1091-6490. DOI: 10.1073/pnas.1522203113. URL: https://www.pnas.org/doi/10.1073/pnas.1522203113 (cit. on pp. 19, 20, 22–24, 39–44, 67).

[63] B.-J. Yoon. "Hidden Markov Models and their Applications in Biological Sequence Analysis". In: *Current genomics* 10.6 (Sept. 2009), pp. 402–415. ISSN: 1875-5488. DOI: 10.2174/138920209789177575. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2766791/ (cit. on p. 19).

[64] S. Zaccaria et al. "The Copy-Number Tree Mixture Deconvolution Problem and Applications to Multi-sample Bulk Sequencing Tumor Data". In: *Research in Computational Molecular Biology - 21st Annual International Conference, RECOMB 2017, Proceedings*. Ed. by S. C. Sahinalp. Springer International Publishing, Apr. 2017, pp. 318–335. ISBN: 978-3-319-56970-3. DOI: 10.1007/978-3-319-56970-3_20. URL: https://link.springer.com/chapter/10.1007/978-3-319-56970-3_20 (cit. on p. 19).

[65] V. Gómez-Rubio. "Mixture models". In: *Bayesian Inference with INLA*. 1st ed. Boca Raton, FL: Chapman and Hall/CRC Press, Feb. 2020. Chap. 13, pp. 310–330. ISBN: 9781138039872. URL: https://becarioprecario.bitbucket.io/inla-gitbook/ch-mixture.html (cit. on pp. 20, 21, 24, 33).

[66] M. El-Kebir et al. "Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures". In: *Cell Systems* 3.1 (July 2016), pp. 43–53. ISSN: 2405-4712. DOI: 10.1016/j.cels.2016.07.004. URL: https://doi.org/10.1016/j.cels.2016.07.004 (cit. on pp. 22, 23).

[67] F. Raynaud et al. "Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability". In: *PLOS Genetics* 14.9 (Sept. 2018). DOI: 10.1371/journal.pgen.1007669. URL: https://doi.org/10.1371/journal.pgen.1007669 (cit. on pp. 22, 23).

[68] Z. Ghahramani, M. Jordan, and R. P. Adams. "Tree-Structured Stick Breaking for Hierarchical Data". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty et al. Curran Associates, Inc., June 2010. ISBN: 9781617823800. URL: https://papers.nips.cc/paper/2010/file/a5e00132373a7031000fd987a3c9f87b-Paper.pdf (cit. on p. 23).

[69] S. Sun et al. "Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis". In: *Genome biology* 20 (Dec. 2019). DOI: 10.1186/s13059-019-1898-6. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6902413/ (cit. on pp. 25, 26).

[70] J. Lehmann et al. "Machine learning in oncology—Perspectives in patient-reported outcome research". In: *Der Onkologe* 27.2 (Nov. 2021), pp. 150–155. ISSN: 1433-0415. DOI: 10.1007/s00761-021-00916-9. URL: https://doi.org/10.1007/s00761-021-00916-9 (cit. on p. 25).

[71] A. Géron. "Dimensionality Reduction". In: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Ed. by R. Roumeliotis and N. Tache. 2nd ed. O'Reilly Media, Sept. 2019. Chap. 8, pp. 213–234. ISBN: 9781492032649 (cit. on pp. 25, 26, 28).

[72] A. A. Patel. "Dimensionality Reduction". In: *Hands-On Unsupervised Learning Using Python*. 1st ed. O'Reilly Media, Mar. 2019. Chap. 3, pp. 96–127. ISBN: 9781492035640 (cit. on pp. 25–28, 47).

[73] R. Xiang et al. "A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data". In: *Frontiers in genetics* 12 (Mar. 2021). ISSN: 1664-8021. DOI: 10.3389/fgene.2021.646936. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8021860/ (cit. on pp. 26, 27).

[74] P. Rózsa. "CHAPTER 1 - MATHEMATICAL PRELIMINARIES". In: *Applied Dimensional Analysis and Modeling*. Ed. by T. Szirtes and P. Rózsa. 2nd ed. Butterworth-Heinemann, Apr. 2007. Chap. 1, pp. 1–26. ISBN: 978-0-12-370620-1. DOI: https://doi.org/10.1016/B978-012370620-1.50007-1. URL: http://csc.knu.ua/media/study/asp/dsm_zhuk/book/book1.pdf (cit. on p. 26).

[75] S. V.S and S. Surendran. "A Review of Various Linear and Non Linear Dimensionality Reduction Techniques". In: *International Journal of Computer Science and Information Technologies* 6.3 (June 2015), pp. 2354–2360. ISSN: 0975-9646. URL: http://ijcsit.com/docs/Volume%206/vol6issue03/ijcsit2015060383.pdf (cit. on pp. 26–28).

[76] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In: *SIAM Review* 53.2 (June 2011), pp. 217–288. DOI: 10.1137/090771806. URL: https://doi.org/10.1137/090771806 (cit. on pp. 26, 27).

[77] X. Lin and P. C. Boutros. "Optimization and expansion of non-negative matrix factorization". In: *BMC Bioinformatics* 21.1 (Jan. 2020). DOI: 10.1186/s12859-019-3312-5. URL: https://doi.org/10.1186/s12859-019-3312-5 (cit. on p. 27).

[78] B. Ghojogh et al. "Multidimensional scaling, sammon mapping, and isomap: Tutorial and survey". In: *ArXiv* abs/2009.08136 (Sept. 2020). DOI: https://doi.org/10.48550/arXiv.2009.08136. URL: https://arxiv.org/pdf/2009.08136.pdf (cit. on p. 28).

[79] B. Y. Oh et al. "Intratumor heterogeneity inferred from targeted deep sequencing as a prognostic indicator". In: *Scientific Reports* 9.1 (Mar. 2019). DOI: 10.1038/s41598-019-41098-0. URL: https://doi.org/10.1038/s41598-019-41098-0 (cit. on pp. 29, 31).

[80] P. Danecek et al. "The Variant Call Format and VCFtools". In: *Bioinformatics* 27.15 (Aug. 2011), pp. 2156–2158. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr330. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/ (cit. on p. 30).

[81] T. K. Kim. "T test as a parametric statistic". In: *Korean journal of anesthesiology* 68.6 (Dec. 2015), pp. 540–546. ISSN: 2005-6419. DOI: 10.4097/kjae.2015.68.6.540. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4667138/ (cit. on p. 38).

[82] D. T. Larose and D. Chantal. "Simple linear regression". In: *Data mining and predictive analytics*. 2nd ed. New Jersey: John Wiley & Sons, Mar. 2015. Chap. 8, pp. 171–235. ISBN: 978-1-118-11619-7. URL: https://www.wiley.com/en-us/Data+Mining+and+Predictive+Analytics%2C+2nd+Edition-p-9781118116197 (cit. on p. 38).

[83] N. Nachar. "The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution". In: *Tutorials in Quantitative Methods for Psychology* 4.1 (Mar. 2008). DOI: 10.20982/tqmp.04.1.p013. URL: https://www.tqmp.org/RegularArticles/vol04-1/p013/p013.pdf (cit. on p. 38).

[84] M. Richard. "Markov Chain Monte Carlo". In: *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC, Mar. 2020. Chap. 9, pp. 263–298. ISBN: 9780429029608. DOI: https://doi.org/10.1201/9780429029608. URL: https://www.taylorfrancis.com/books/mono/10.1201/9780429029608/statistical-rethinking-richard-mcelreath (cit. on p. 43).

[85] S. Zaccaria and B. J. Raphael. "Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data". In: *Nature Communications* 11.1 (Sept. 2020). DOI: 10.1038/s41467-020-17967-y. URL: https://doi.org/10.1038/s41467-020-17967-y (cit. on p. 50).

[86] J. Abécassis, F. Reyal, and J.-P. Vert. "CloneSig can jointly infer intra-tumor heterogeneity and mutational signature activity in bulk tumor sequencing data". In: *Nature Communications* 12.1 (Sept. 2021). DOI: 10.1038/s41467-021-24992-y. URL: https://www.nature.com/articles/s41467-021-24992-y (cit. on p. 50).

Appendix A

Practical work supplementary images

Part I - Tumor heterogeneity index

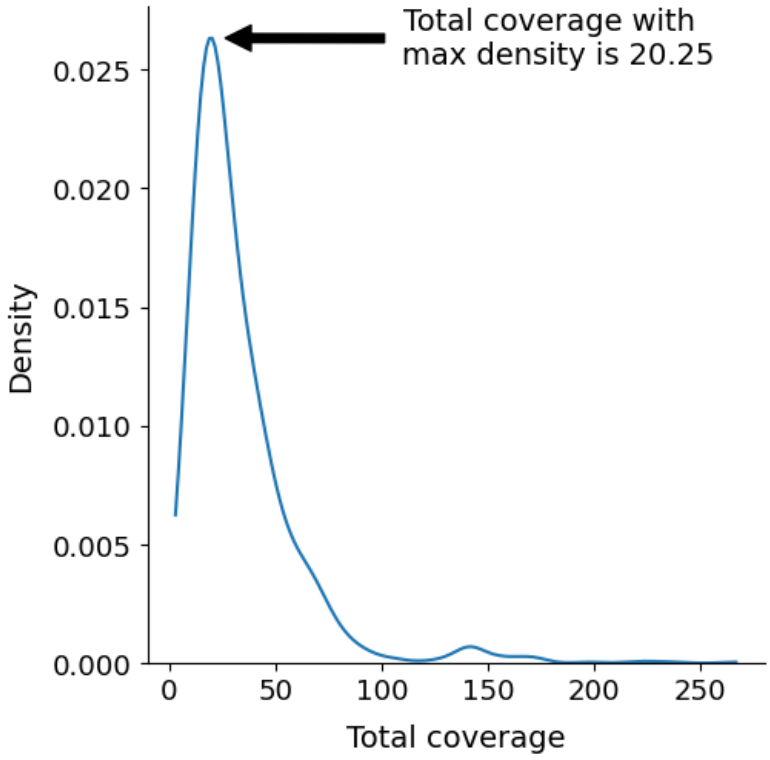

Figure A.1: KDE plot with the coverage distribution of all samples (ref. on p. 33).

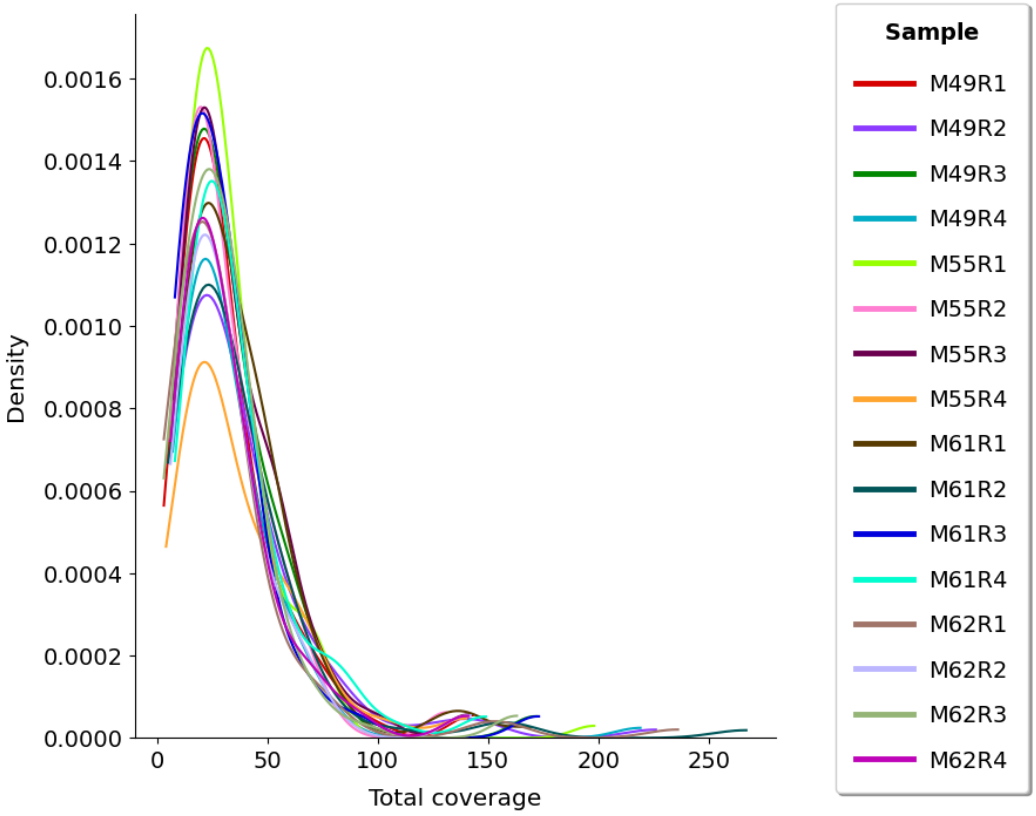

Figure A.2: KDE plot with the coverage distributions of different samples (ref. on p. 33).

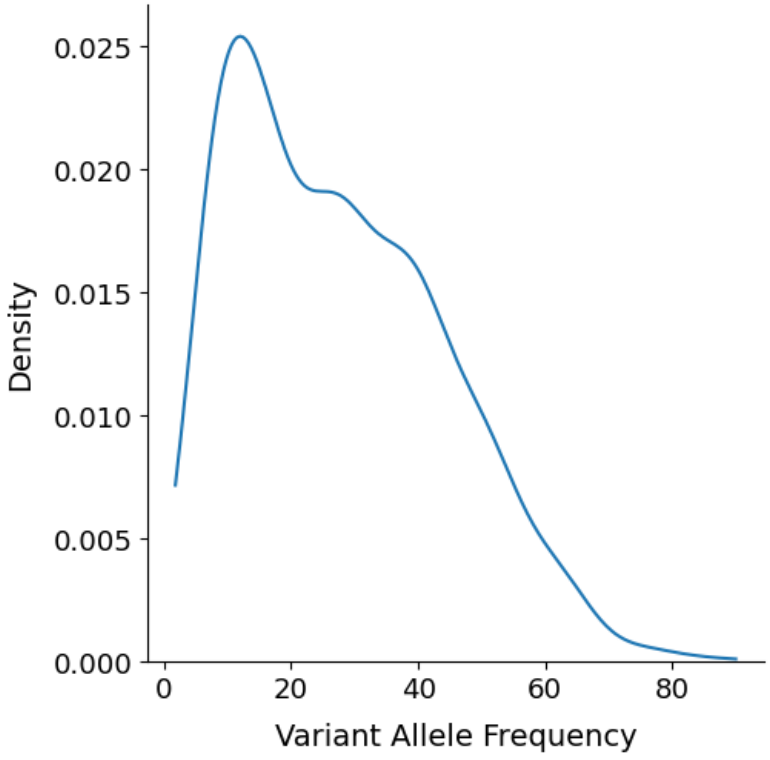

Figure A.3: KDE plot with the average VAF distribution using all samples (ref. on p. 33).

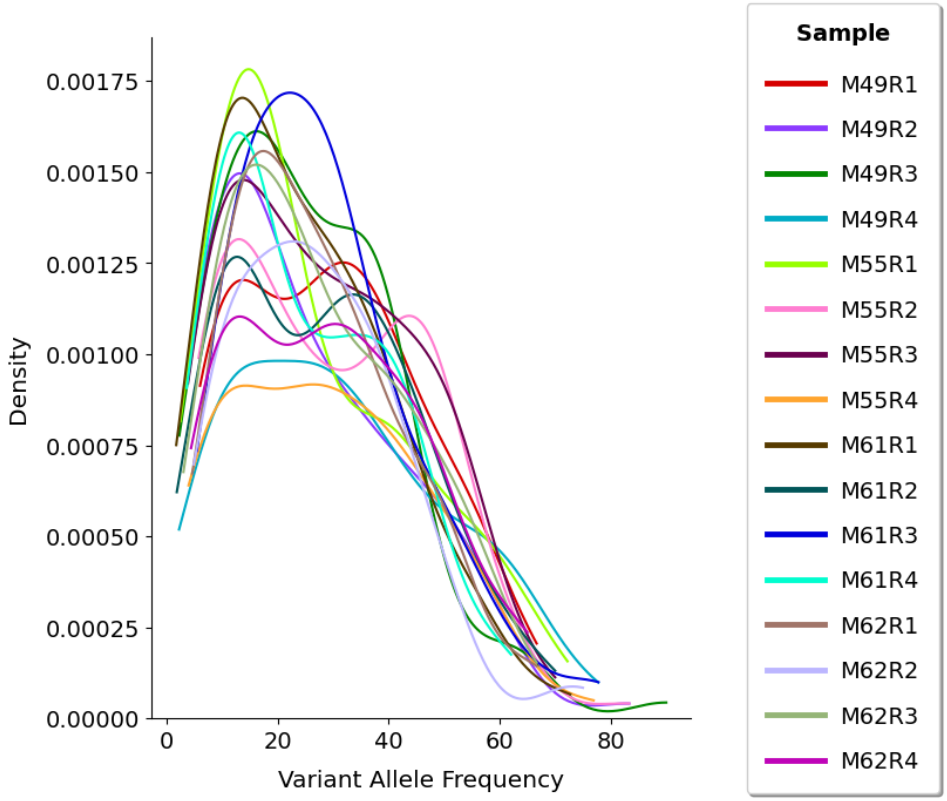

Figure A.4: KDE plot with the VAF distributions of different samples (ref. on p. 33).

Part I - PyClone-VI

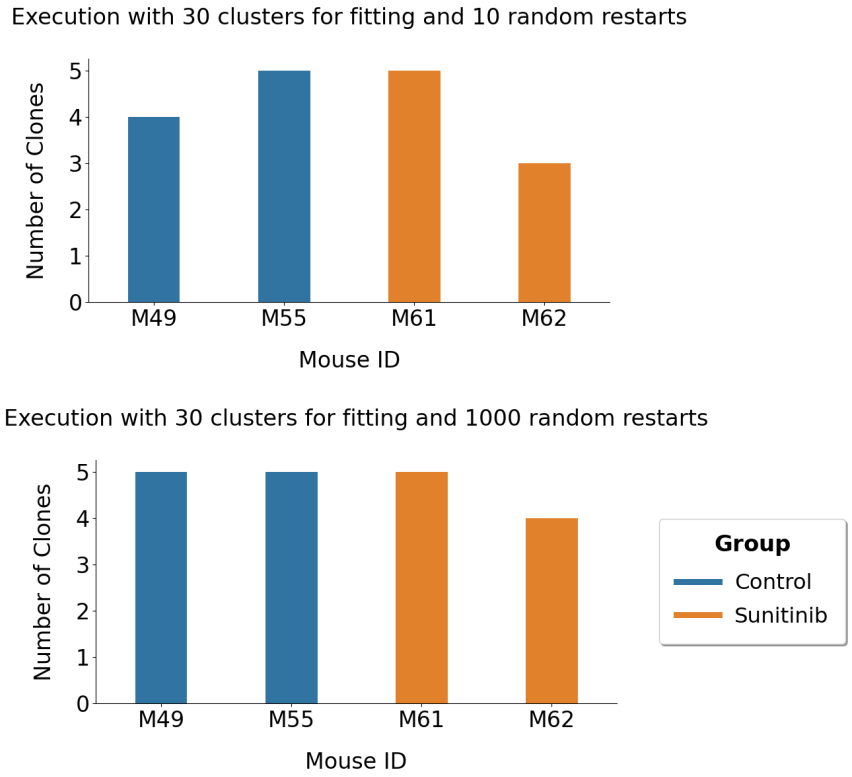

Figure A.5: Comparison between the number of clones found in the two groups of mice when using the samples of each mouse in different files for the execution of the algorithm (ref. on p. 37).

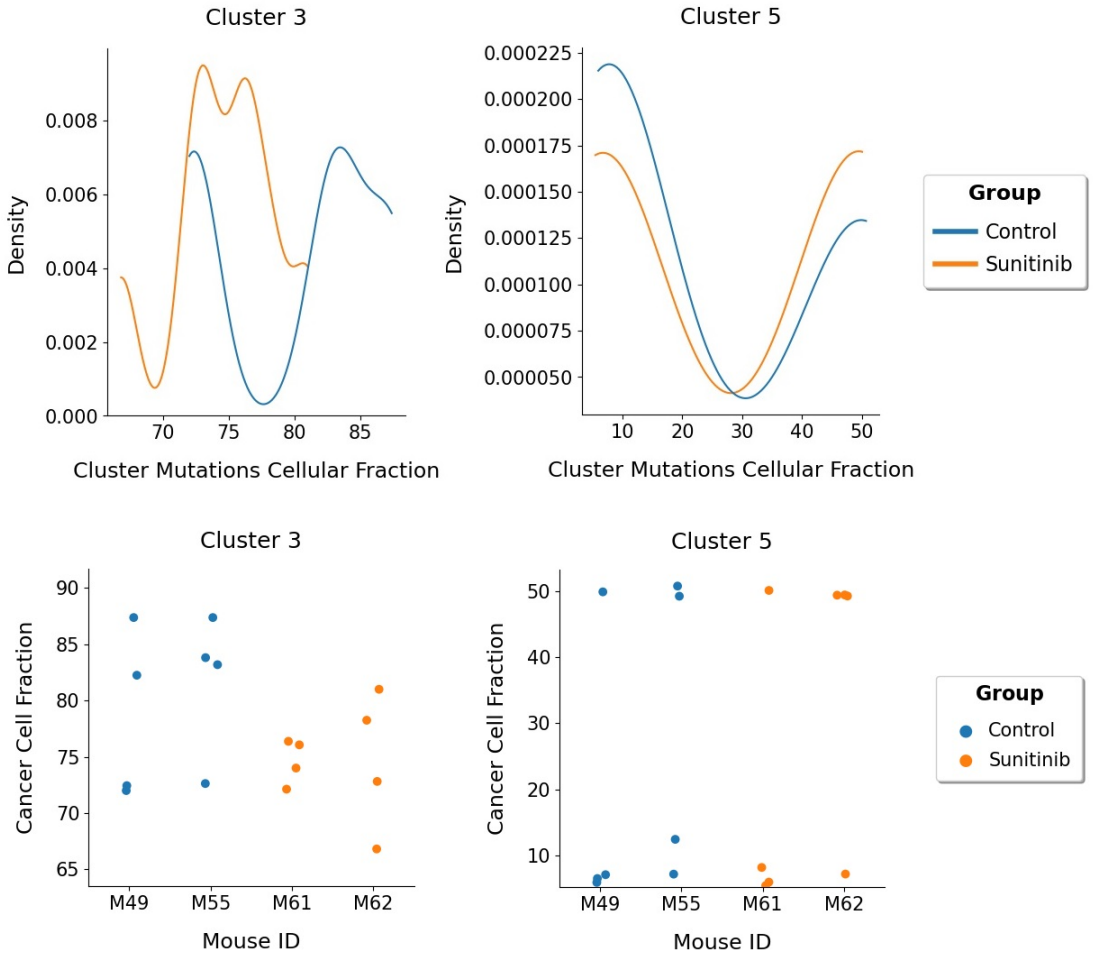

Figure A.6: Comparison of the CCF distribution of the mutations and the average CCF of clusters 3 and 5 between the two groups of mice when using the samples of all mice together in a file for execution (ref. on p. 37).

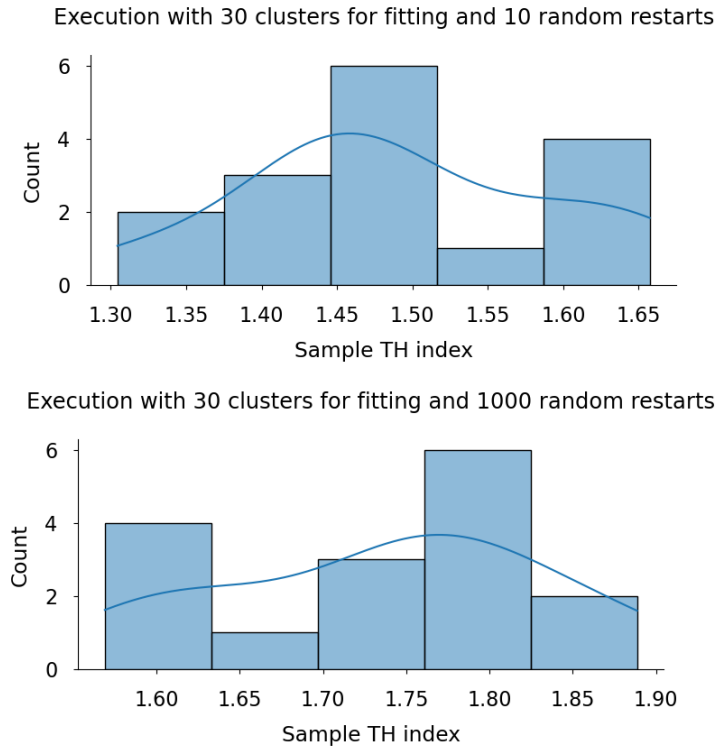

Figure A.7: Histogram and KDE plots to assess the normality of the distribution of the THI values of the mice samples (ref. on p. 38).

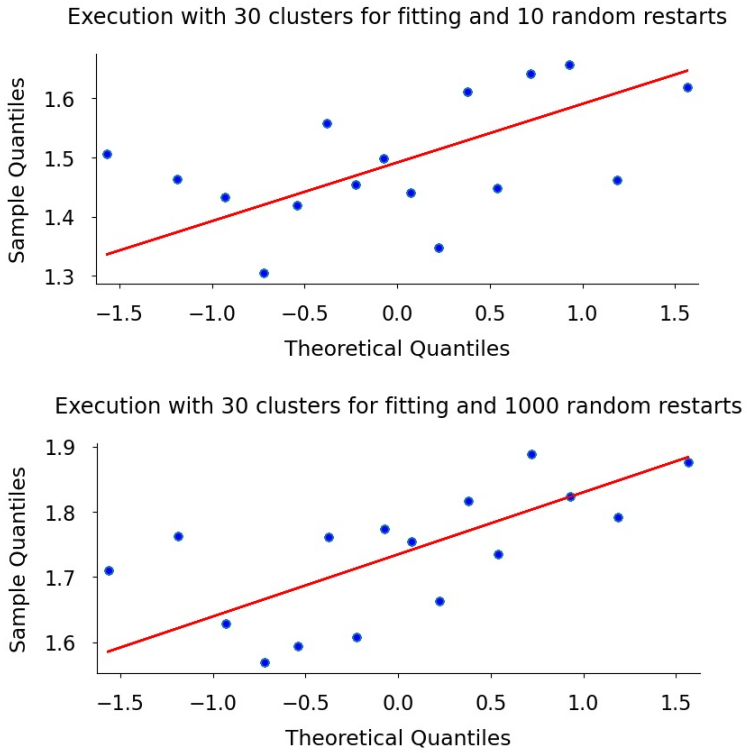

Figure A.8: Quantile-quantile plots of the distribution of the mice samples THI against a normal distribution of data (ref. on p. 38).

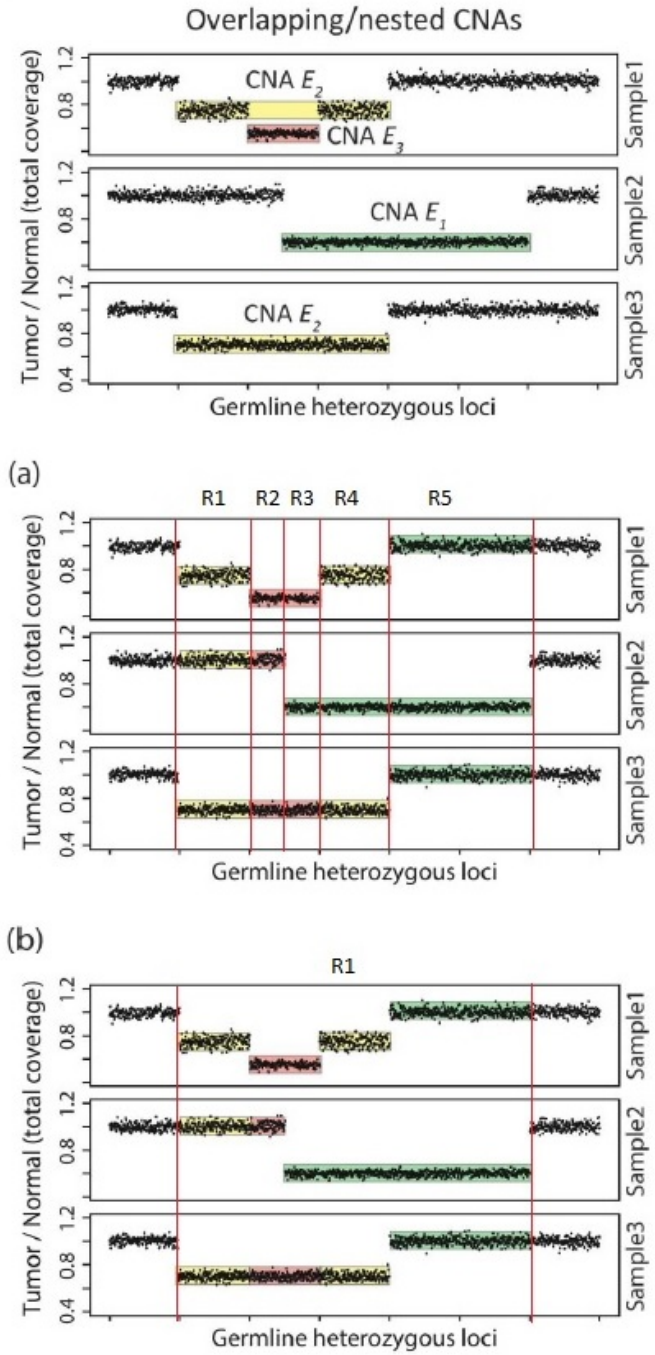

Figure A.9: Two different approaches for defining CNA regions: (a) builds each region as the intersection of different CNA events; (b) considers that the union of the CNAs intersected corresponds to a region. Adapted from the supplementary material of [62] (ref. on p. 40).

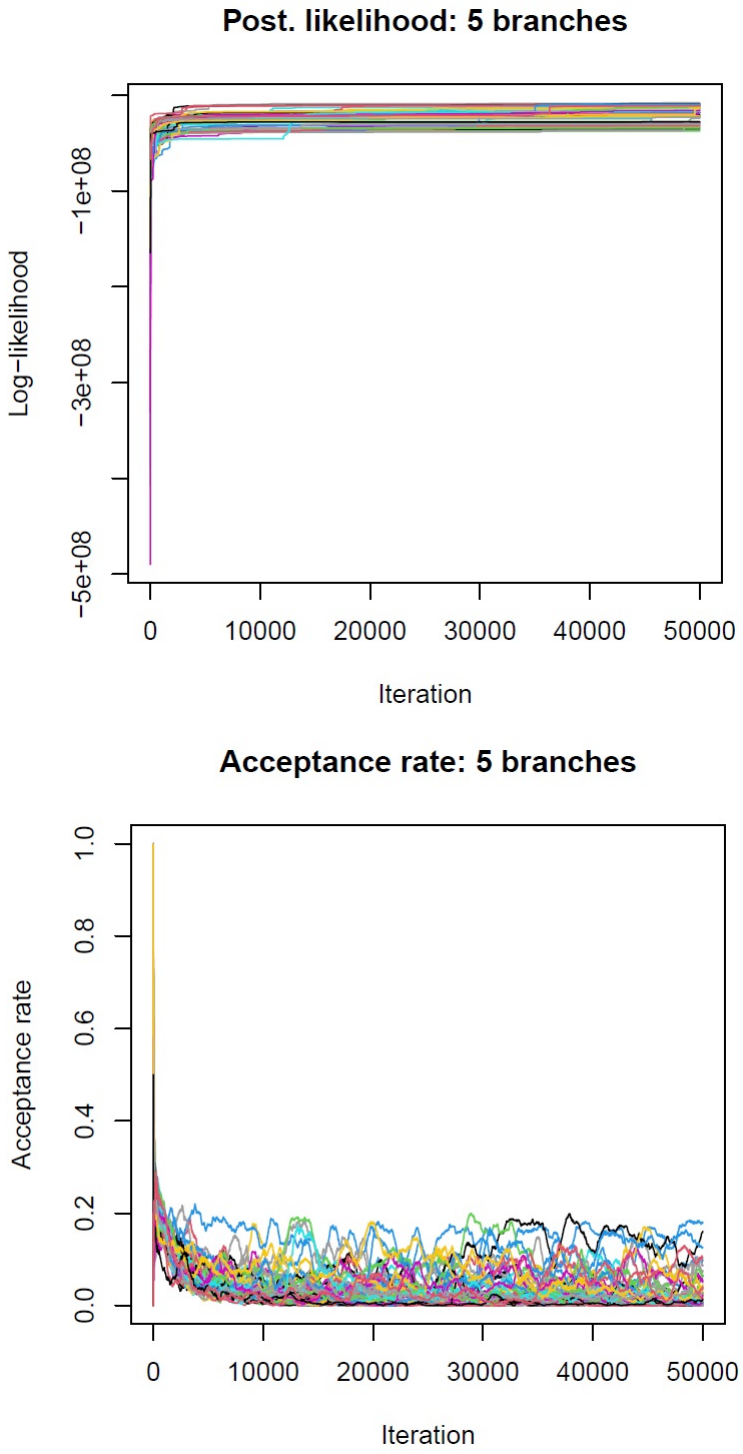

Figure A.10: Plots of the posterior likelihood and the acceptance rate for the tree space sampled using 5 clones with all mice samples input together. The chains are colored differently (ref. on p. 43).

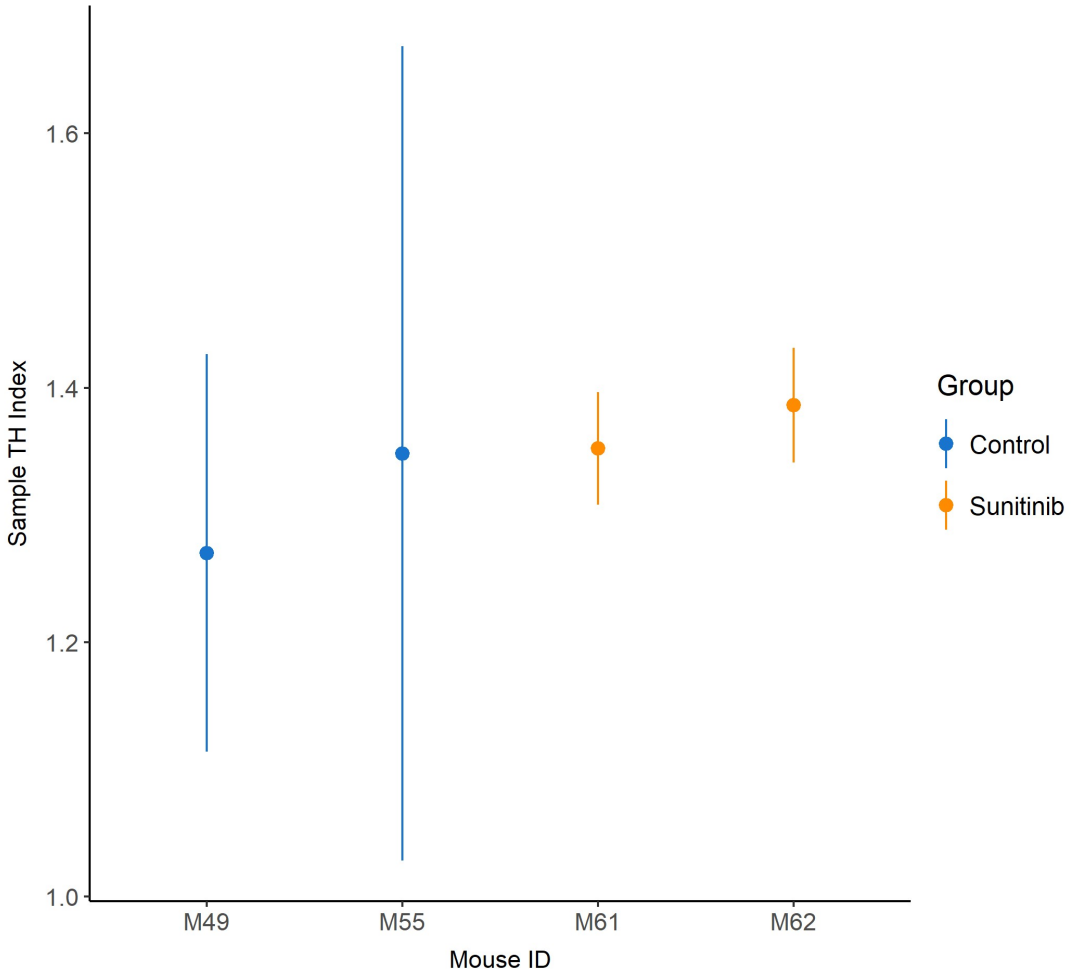

Figure A.11: Point plot with the mean THI values and confidence intervals of 95% when using all mice samples together for the input, comparing between the two mice groups (ref. on p. 44).

Part II - Genetic information feature analysis

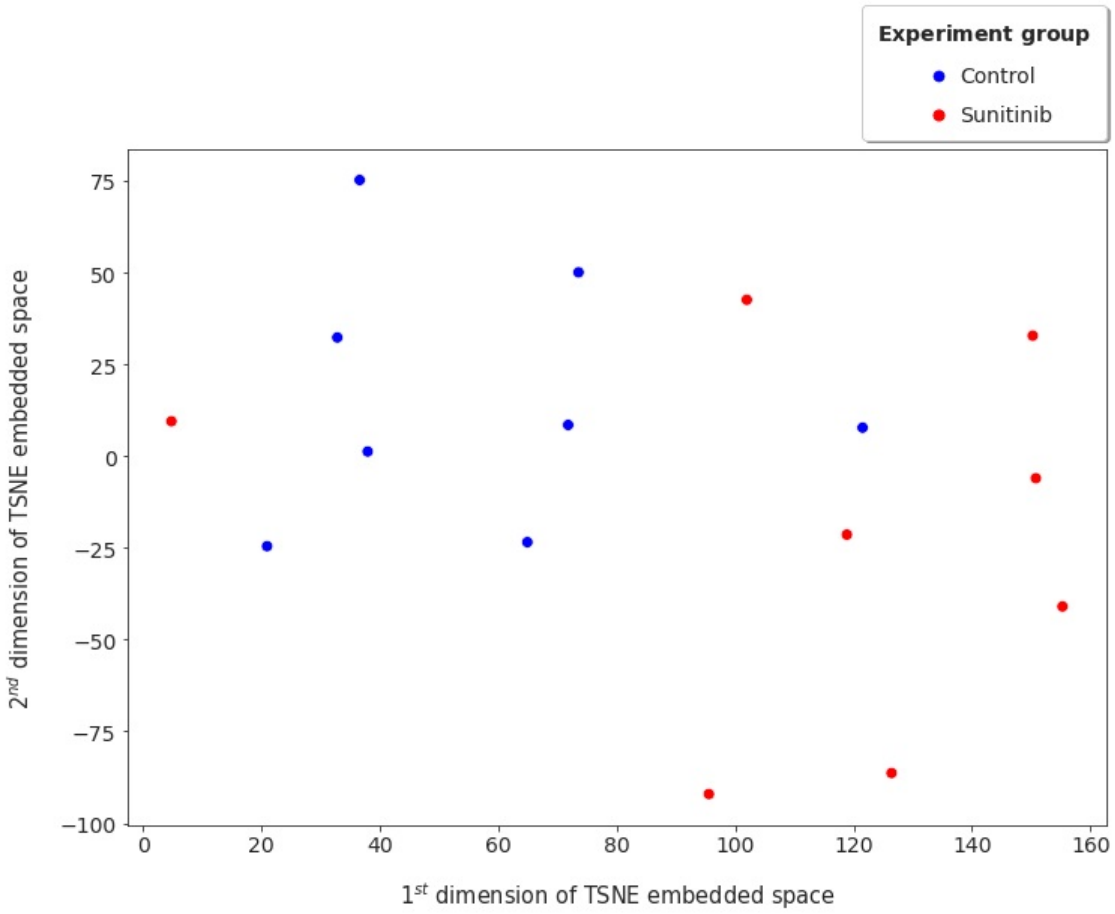

Figure A.12: PCA to 16 dimensions followed by the t-SNE to 2 dimensions using the gene features of the mice samples (ref. on p. 47).

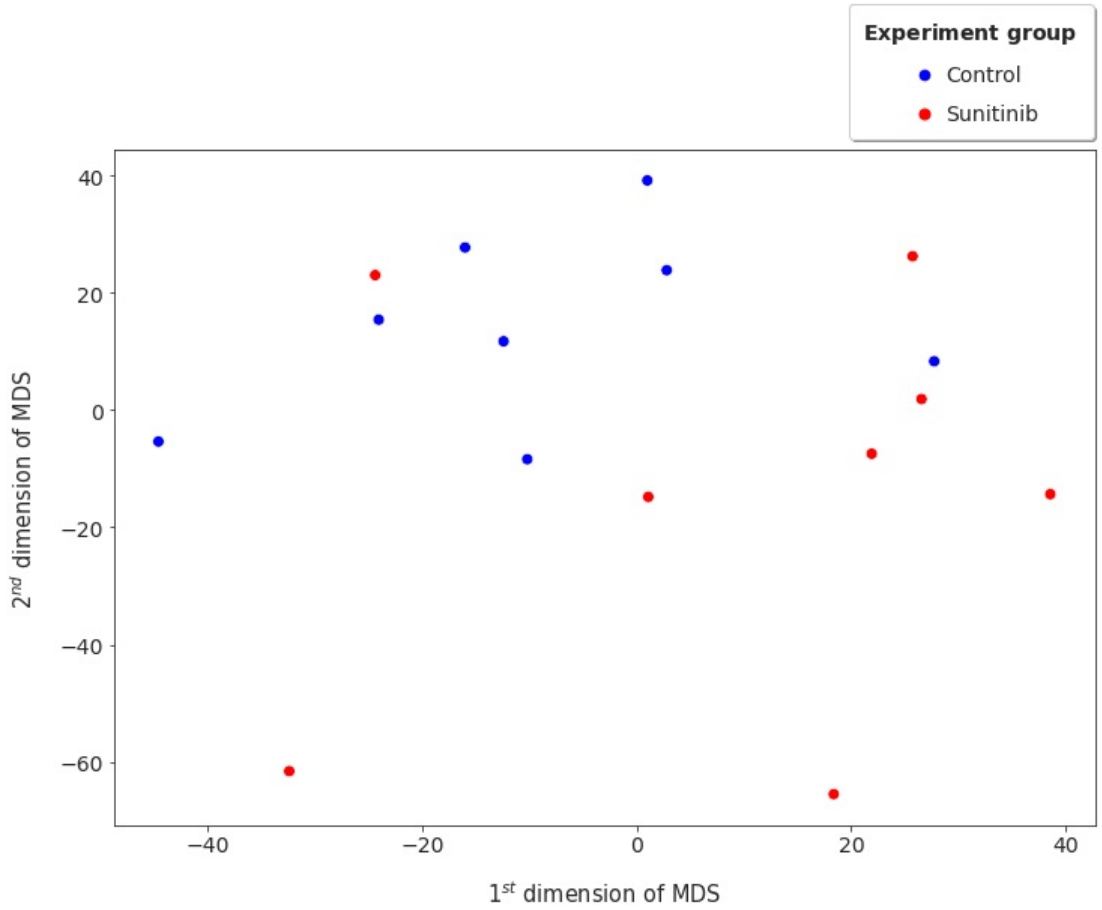

Figure A.13: MDS to 2 dimensions using the gene features of the mice samples (ref. on p. 47).

Figure A.14: Legend with the different tumor types of patients data stored in TCGA. The names of the tumors corresponding to the different acronyms can be found at TCGA study abbreviations (ref. on p. 48).

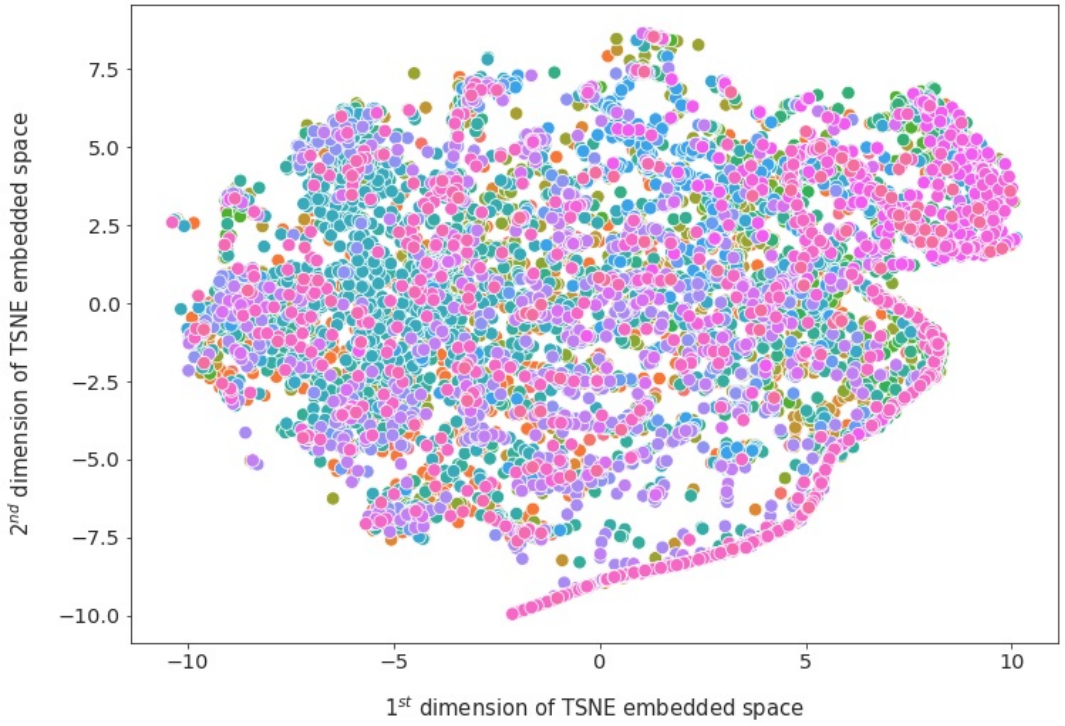

Figure A.15: PCA to 25 dimensions followed by the t-SNE to 2 dimensions using the tumor data of the patients stored in TCGA (ref. on p. 48).

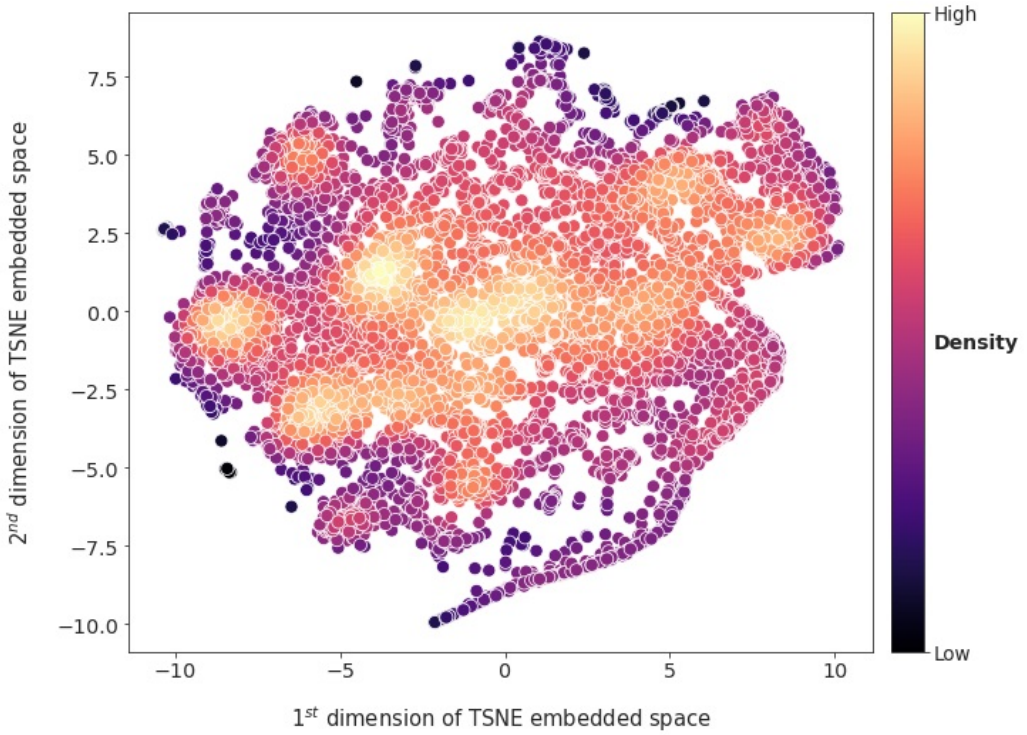

Figure A.16: Density zones of the points in figure A.15 (ref. on p. 48).

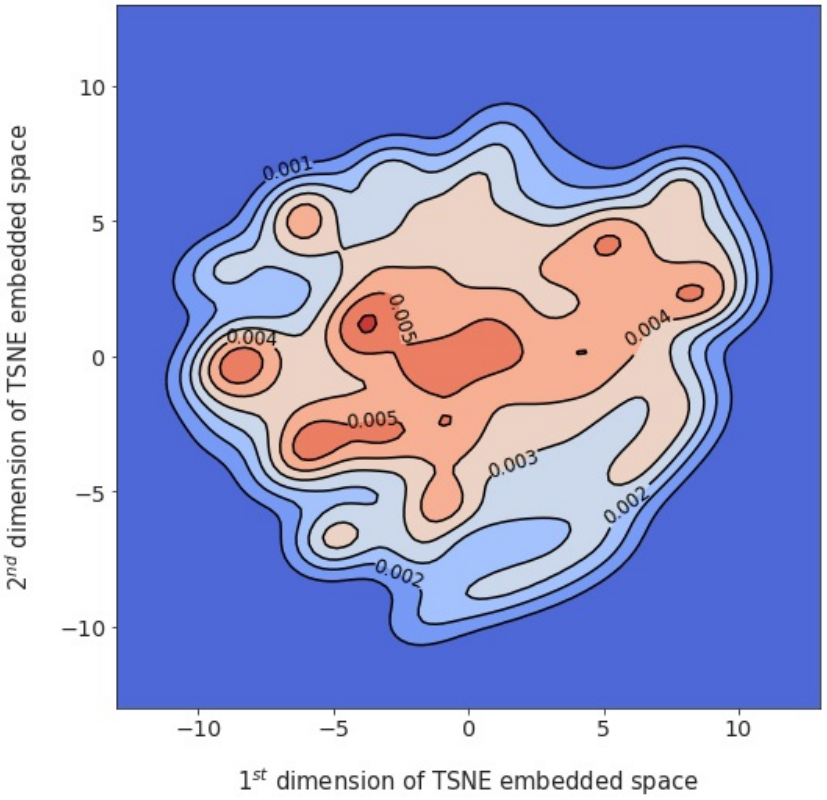

Figure A.17: KDE plot in 2 dimensions of the points in figure A.15 (ref. on p. 48).