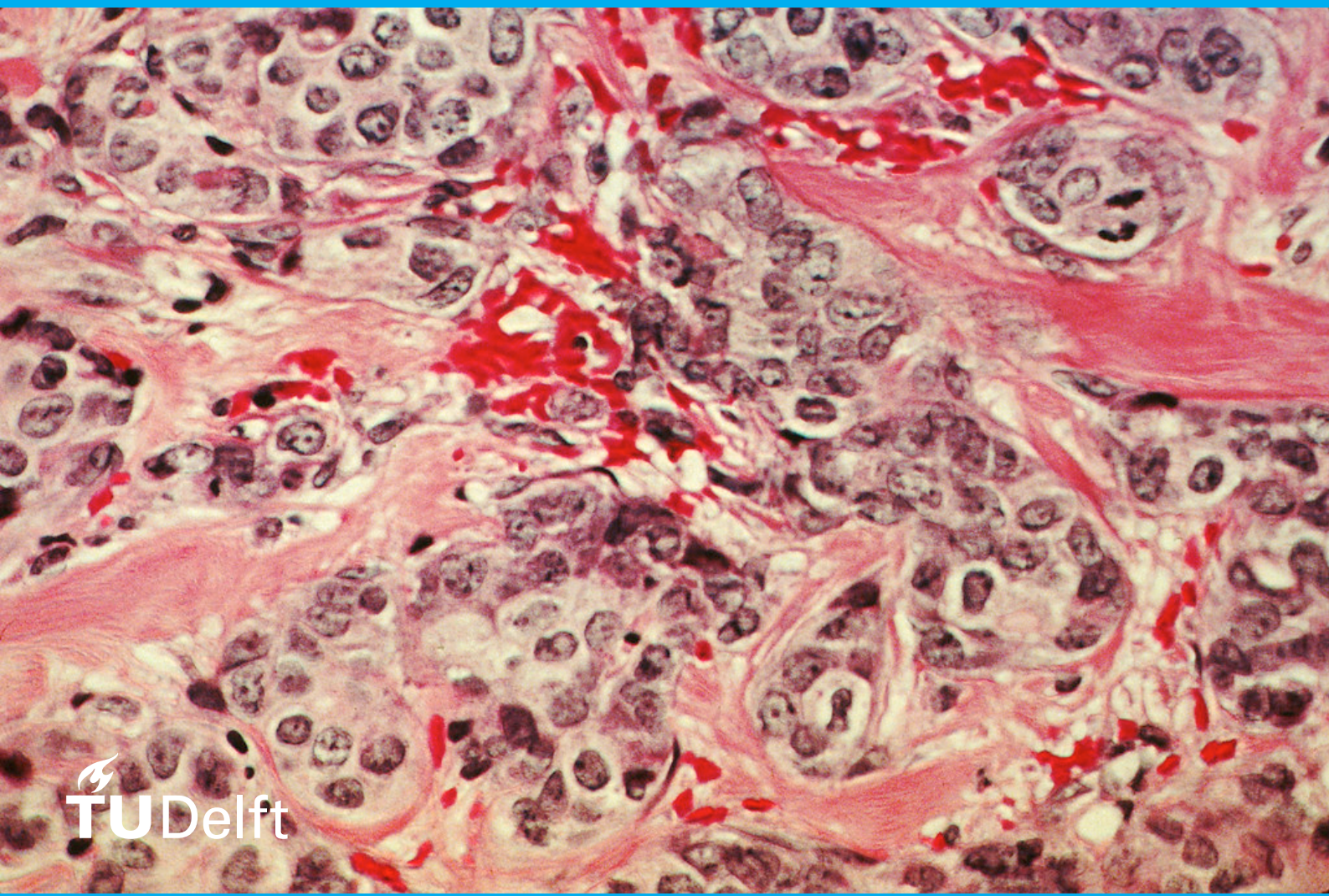
Detecting clonality in contralateral breast cancers

Laura Middeldorp

In collaboration with
Erasmus MC, Rotterdam

Faculty EEMCS, Delft University of Technology



Detecting clonality in contralateral breast cancers

by

Laura Middeldorp

to obtain the degree of Master of Science in Applied Mathematics,
for the specialization Stochastics,
at the faculty of Electrical Engineering, Mathematics and Computer Science,
at the Delft University of Technology,
to be defended publicly on Tuesday January 18, 2022 at 10:00.

Student number:	4484436	
Project duration:	March 1, 2021 – January 18, 2022	
Thesis committee:	Dr. H. P. Lopuhaä	TU Delft, chair
	Dr. H. N. Kekkonen	TU Delft, daily supervisor
	Dr. M. J. Hooning	Erasmus MC, supervisor
	M. Smid, BSc.	Erasmus MC, supervisor

TU Delft

Erasmus MC
University Medical Center Rotterdam

Abstract

When a second tumor arises in the contralateral breast in a patient with a previous or synchronous breast cancer, it is of clinical importance to determine if this tumor is a new unrelated tumor or a metastasis, i.e. clone, of the primary tumor. A new, unrelated tumor may be treated similarly as the first one since treatment was successful, while a distant metastasis demands a change of therapy and has a more adverse prognosis. In clinic, a second tumor is generally regarded as a new primary. If there is clinical suspicion that the second tumor may be a metastasis, clinico-pathological characteristics of the two tumors are used assess the clonality status. Clinico-pathological characteristics, however, are not reliable predictors to determine if a second tumor is a metastasis. Recent studies have investigated tumor clonality using techniques from molecular genetics. These models appear to perform well, but have several drawbacks.

In this thesis a more advanced classification model is being developed that can detect tumor clonality based on SNP array data. For this, two segmentation algorithms, ASCAT and OncoSNP, and two comparison methods, Log LR and adapted SI, have been incorporated. For each tumor, the segmentation algorithms construct a copy number profile based on the SNP array data. Given the copy number profiles, the comparison methods compute a p-value which reflects the probability that a pair is of clonal origin. Both comparison methods are permutation methods which test the null hypothesis of independence against the alternative hypothesis assuming clonality. The proposed model consists of a decision tree which assigns each pair to one of six categories depending on the significance of the four resulting p-values.

The model has been tested on 23 fresh frozen pairs by means of expert judgment. The results were promising: the four pairs which were unanimously labeled as clonal by the experts were also regarded as such by the model. No independent pairs were assigned as clonal by the model. Moreover, the decision tree showed to have a higher sensitivity than the clinical assessments as the latter only managed to detect two out of four clonal pairs. A discordance between the clinico-pathological judgments and decision tree results was found for three out of 18 pairs for which both assessments were available.

The model appears to be suitable in practice, but is not yet applicable as a stand-alone model. There were two ambiguous pairs which were labeled as independent by the model but for which the experts had varying opinions about the clonality status. Until the ambiguous pairs can be reliably categorized, it is advised to take into account both the model results and clinical assessments when determining tumor clonality. Finally, the performance of the model remains to be tested on FFPE pairs.

Keywords: contralateral breast cancer, clonality, SNP arrays, copy number profiles, permutation methods, decision tree.

Preface

Dear reader,

The thesis in front of you completes my journey as a master student at the Delft University of Technology. For the past ten months, I have worked on developing a statistical model that can detect clonality in contralateral breast cancers. The research was conducted in collaboration with the Erasmus MC in Rotterdam. My intrinsic motivation was immediately sparked when encountering this project as my grandmother, Gré, has had contralateral breast cancer. Fortunately, after a double mastectomy and anti-hormonal therapy Gré was declared cancer-free and continued to live for many years. She passed away from dementia in 2015 aged 85. I know that she is looking at me from above and will be very proud of what I have achieved during my master thesis research.

I must admit that the adventure of conducting this research has not been easy at all times. Luckily, I have received great support and supervision while working on my thesis. For this, I would like to thank my supervisors. Hanne, thank you for guiding me through this process. Your countless suggestions and analytical view have really enriched this research and kept me focused to determine the next steps. Geurt, thank you for your supervision, your infinite knowledge of statistics and your help to enable the collaboration with the Erasmus MC for my thesis. Maartje, Antoinette and Marcel thank you for your supervision, enthusiasm and most importantly, for getting me acquainted with the world of DNA and breast cancer. Given the medical knowledge that I have obtained, I feel like half a doctor now. I want to thank Rik Lopuhaä for being the chair of my thesis committee. Kirsten, thank you for preparing the DNA of the contralateral pairs for me.

Next to my supervisors and committee members, I am very grateful for the mental support that I have received from my family and friends the past ten months. Thank you mom and dad for your unconditional love and support and your willingness to always lend an ear whenever I faced difficulties in my research. My boyfriend, Flynn, thank you for making me believe in myself throughout this entire process and for always being there for me whenever I have had a rough day. My roommate, Elise, thank you for listening to my problems during dinner whenever I was stuck and for providing me with books about molecular biology and pathology. A special thanks goes out to my friend Matea, who, even though she was all the way in Croatia, comforted me during difficult times and shared her thesis experiences with me. Moreover, I would also like to thank the departments of Cancer Epidemiology and Medical Oncology at the Erasmus MC for making me feel part of their community. Last but not least, I would like to thank everyone that I have encountered during my master education: you have made this an unforgettable experience.

I wish you a pleasant reading.

Laura Middeldorp
Delft, January 2022

Contents

1 Introduction

The aim of this thesis is to examine how clonality in contralateral breast cancers can be detected. In this chapter the key aspects of breast cancer are introduced as well as why investigating the clonal relatedness between primary and contralateral breast cancers is of interest to be researched. In addition, an overview of the current guidelines used to assess whether two tumors are clonal are presented. The last subsection gives an outline of the report.

1.1. What is breast cancer?

Breast cancer is one of the most prevalent cancers among women both in developed and developing countries. In 2020, a grand total of 2.26 million women were diagnosed with breast cancer and 685,000 deaths due to breast cancer were reported worldwide [34]. It is estimated that one in ten of all cancers diagnosed each year worldwide is cancer of the female breast [14]. Moreover, one in seven women in the Netherlands will develop breast cancer over the course of her lifetime [41].

Breast cancer occurs when breast cells start to grow abnormally. In more than 99% of the cases, breast cancer emerges in the epithelial cells giving rise to *carcinomas*. The epithelial component of the breast consists of cells that line the lobules and ducts. Regarding the histological type of carcinomas, approximately 75% are of the ductal type, 15% are of the lobular type and the remaining 10% are a mixture of other histologies which are less common [25]. It should be noted that the histological type only refers to the growth pattern of the tumor and not to the cell of origin being either from the ducts or lobules. In less than 1% of the cases, breast cancer arises in the non-epithelial cells of the breast, such as the blood vessels and fat cells [28]. These types of breast cancers are called *sarcomas*. The majority of breast cancers thus arise in the ducts or lobules of the breast. Figure 1.1 shows an illustration of the anatomy of the female breast.[1]

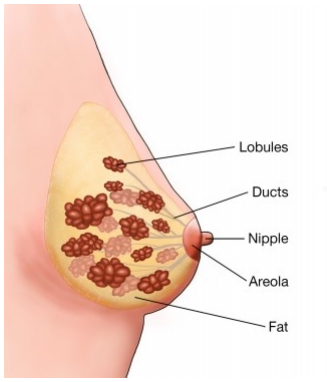


Figure 1.1: Illustration of the anatomy of the female breast.

[1] https://www.oncolink.org/cancers/breast/breast-cancer-the-basics

In the initial stage of breast cancer, also called stage 0, the tumor is *in situ*, meaning that it is encompassed within the duct or lobule. During this stage, the tumor does not cause any symptoms. As time passes, the stage 0 tumor may become more progressive, leaving the duct or lobule and starts invading the surrounding breast tissue. When this happens, the cancer is defined as *invasive* breast cancer. If the cancer manages to reach the lymph nodes, other organs may also be affected by the cancer leading to distant metastases.

Treatment of breast cancer can be very effective if the tumor is diagnosed at an early stage. The treatment of breast cancer generally consists of a combination of multiple therapies such as surgical removal of the tumor (breast-conserving surgery or mastectomy), radiation therapy, chemotherapy or anti-hormonal therapy. Which combination of treatments is most suitable for the patient depends on the stage and grade of the breast cancer and the type (i.e. whether the cancer cells are sensitive to hormones) [10].

1.2. Clonality in contralateral breast cancers

Unilateral breast cancer patients have an increased risk of developing breast cancer in the other, *contralateral*, breast. It is estimated that 2 to 11 percent of women diagnosed with breast cancer will develop contralateral breast cancer (CBC) in their lifetime [9].

For clinical purposes, it is important to determine whether the second tumor in the contralateral breast is a new primary tumor or a metastasis of the first tumor. A new, unrelated tumor may be treated similarly as the first tumor since treatment was successful, while a distant metastasis may ask for a change of therapy and has a more unfavorable prognosis. In the case of a clonal secondary tumor the treatment of the first tumor was not effective meaning that a small amount of cancer cells of the first tumor remained behind in the body of the patient. These cancer cells eventually managed to reach the contralateral breast where the cancer started to develop. Hence, a different therapy is needed to eradicate the cancer when the secondary tumor is a distant metastasis of the first.

In practice, however, contralateral breast cancers are generally regarded as independent tumors [2]. Only when the patient already has metastases from her primary breast cancer in other organs, a metastasis in the other breast will also be considered. In that case, the pathological and clinical characteristics of the two tumors will be compared such as:

- The degree of differentiation: how similar are the tumor cells in appearance compared to the normal cells of the breast. The degree of differentiation determines the grade of the cancer. A low grade tumor has cells that are well differentiated, meaning that the tumor cells look similar to normal cells. A high grade tumor, on the other hand, has cells that do not look like normal cells meaning that the tumor cells are poorly differentiated from the normal cells.
- Time interval between the tumors: how much time has passed between the diagnosis of the primary and secondary tumor. For example, if CBCs are diagnosed 10 years apart, it is likely that the secondary tumor is unrelated to the primary tumor.
- The histological subtype: e.g. ductal or lobular and is the cancer in situ or invasive.
- Hormone receptor status: does the tumor express receptors for hormones such as estrogen and progesterone. If, for instance, the tumor is ER-positive (Estrogen Receptor), the tumor cells depend on the hormone estrogen to grow. Tumors that are receptive to one of the two hormones can be treated with anti-hormonal therapy.
- Human Epidermal growth factor Receptor 2 (HER2) status: if a tumor is HER2 positive, the tumor cells contain a certain protein named HER2. This protein causes the tumor cells to grow uncontrollably.

In [5], a pair is classified as clonal if the two tumors have a concordant histological subtype, ER status and HER2 status. However, using these characteristics to determine tumor clonality is not recommended since the majority of breast cancers have the same characteristics: invasive ductal, ER-positive and HER2-negative [42]. Pathological characteristics are thus not informative to assess tumor

clonality. A discordance in one of the pathological characteristics, for instance the ER-status, may serve as evidence that the two tumors are of independent origin. However, [8] has shown that the estrogen, progesterone and HER2 receptor status frequently change in metastatic breast cancer. Therefore, in order to reliably detect whether a second tumor is clonal or not, molecular features of the two tumors, i.e. the genomic profiles, are needed. This also holds for tumor pairs for which there is no clinical suspicion of a possible metastasis.

Recent studies have investigated tumor clonality using several techniques in the field of molecular genetics (e.g. array Comparative Genomic Hybridization (aCGH), Single Nucleotide Polymorphism (SNP) arrays, DNA methylation and RNA sequencing (RNA-seq)) and emerging evidence suggests that a subset of CBCs (approximately 10%) represent metastases rather than independent primary tumors [3, 5]. These studies have also shown that the current clinical guidelines used for assessing clonality are not well associated with molecular tumor features. For example, [5] investigated 37 pairs of primary and secondary tumors and found a discordance between clinical and molecular assessment in 13 pairs, which is approximately 35%.

In the current literature, several models exist that are capable to quantify the degree of clonal relatedness between two tumors based on their DNA profiles. These models, which are based on comparing the amount of overlap in the two profiles, seem to perform well, but have several shortcomings. The aim of this thesis is, given the existing methods, to develop a more advanced classification model which can detect clonality between contralateral breast cancer tumors given the genomic profiles of the tumors. In this research, the genomic profiles of the primary and secondary breast tumor are generated by means of SNP arrays.

1.3. Outline of the report

This section gives a brief overview of the structure of the report. Chapter 2 gives an introduction to DNA and SNP arrays. This includes the definition of DNA, what happens to the DNA during cancer development, what SNP arrays are and how SNP arrays can be used to determine what has happened to the DNA of tumors. In Chapter 3, an overview of different segmentation algorithms is presented. A segmentation algorithm can be used on the SNP array data to determine where aberrant regions have occurred on the genome. Given the overview, two segmentation algorithms are chosen and explained in more detail. Chapter 4 introduces the comparison methods that will be used on the segmented data. The comparison methods determine whether a pair of tumors is of clonal origin or independent by comparing the profiles coming from the segmentation algorithms. In this thesis, two comparison methods are considered. In Chapter 5, characteristics of the comparison methods are investigated by means of simulations with artificial pairs. In Chapter 6 the performance of the segmentation algorithms and comparison methods is tested on 23 fresh frozen pairs. Given these results, a final model is proposed which can be used by the oncologist. Chapter 7 provides a summary of the conducted research. Finally, Chapter 8 presents the conclusion of this thesis and states recommendations for future research.

2

DNA and SNP arrays

In this chapter, background information regarding DNA and SNP arrays will be given. Section 2.1 gives an introduction to DNA. What is DNA exactly and what are the functions of DNA? Section 2.2 clarifies what happens to the DNA in the case that cancer arises. Finally, Section 2.3 explains the concept of SNP arrays and describes how SNP arrays can be used to determine what has happened to the DNA of the tumor.

2.1. DNA

Deoxyribonucleic acid, also known as DNA, is a complex molecule located in the cells of an organism. DNA contains all the genetic information of a living being. For instance, the observable traits of a living being, also known as phenotypes, are encoded in their DNA. Examples of phenotypes are eye color, hair color, height etc.

In humans, the DNA is located in the nucleus of the cell as illustrated in Figure 2.1.[1]

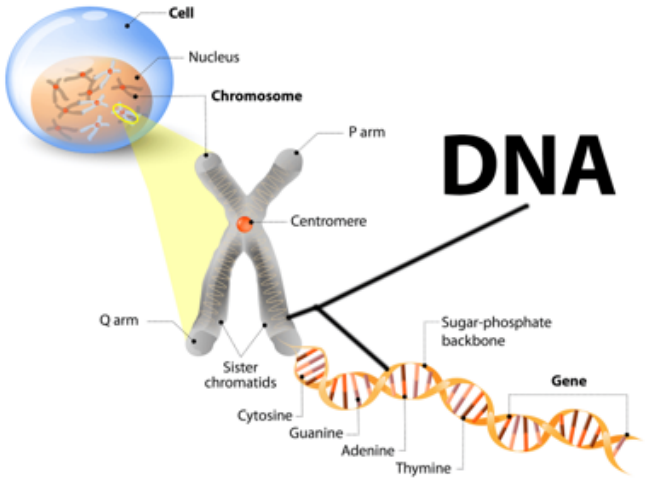


Figure 2.1: Illustration of the composition of DNA.

In the nucleus of human cells, there are 23 pairs of *chromosomes* which contain the genetic information of the individual. The first 22 chromosome pairs are the so-called *autosomes*. The 23rd chromosome pair are the *sex chromosomes* and, as the name suggests, determines the sex of the human being. Figure 2.1 highlights a chromosome. For each pair of chromosomes, a human being inherits one chromosome from the father and one chromosome from the mother. Therefore, the genetic information of a

[1] https://www.istockphoto.com/nl/vector/cell-chromosome-dna-and-gene-gm506992506-84490151

human being consists of 50% of the genetic information of the father and 50% of the genetic information of the mother.

The chromosomes are connected through a centromere which splits each chromosome in a p-arm and a q-arm. Each chromosome is a tightly wound coil of nucleic acid: a large macromolecule made up of millions of nucleotides. A nucleotide is comprised of three different elements: a sugar group, namely deoxyribose, a phosphate group and a nitrogenous base. For each nucleotide, there are four possible nitrogenous bases: cytosine (C), guanine (G), adenine (A) and thymine (T). In DNA, the nitrogenous bases always form solid pairs with each other, where cytosine (C) pairs with guanine (G) and adenine (A) pairs with thymine (T). This phenomenon is also known as *complementary* base pairing: cytosine is complementary to guanine and adenine is complementary to thymine. Figure 2.2 displays how the base pairs bond with each other through hydrogen bonds.[2]

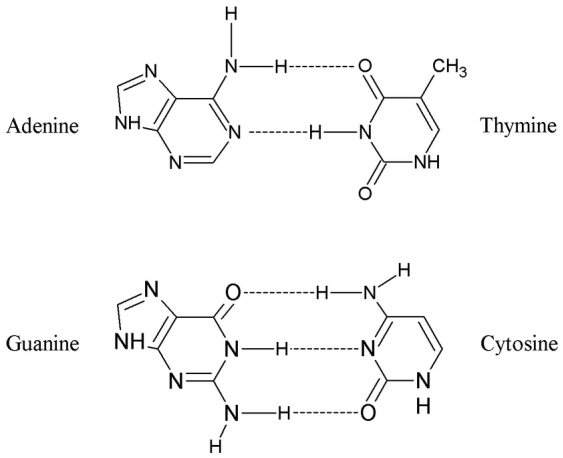


Figure 2.2: Molecular illustration of base pairing.

The formation of the bonds between the base pairs creates the characteristic helical structure of DNA as can be seen in Figure 2.1.

DNA thus consists of nucleic acids which are composed of many nucleotides. A *gene* is a small part of the DNA consisting of hundreds to thousands of base pairs. Each gene has a fixed location on a chromosome, called a locus. It is estimated that humans have approximately 20,000 genes on their chromosomes [24]. The genes contain instructions that tell the cells how to make certain molecules called *proteins* which are vital for the functioning of the human body. Next to that, genes determine the phenotypes of an individual and can also function as predictors for the susceptibility of diseases. An example of a well-known gene that serves a role as a predictor for breast cancer is the *BRCA1* gene. The *BRCA1* gene is located at chromosome 17. Women that inherit a harmful variant of the *BRCA1* gene have an increased risk of developing breast and ovarian cancer. Furthermore, women with a pathogenic variant of *BRCA1* are also more likely to develop cancer at a younger age compared to women who did not inherit a harmful variant.

2.2. DNA and cancer

A human is made up of hundreds of millions of cells that are continuously renewing. The genes in the DNA are responsible for the renewal of the cells: if the genes function properly, the cells will be renewed in an organized way. Sometimes a change can occur in the genes when a cell divides. This change is called a mutation, meaning that the gene has been damaged, lost or is copied too many times. A mutation can happen by chance in the process of cell division, but can also be caused by environmental factors such as UV-radiation, toxic chemicals, diet and smoking. Some mutations are harmless or are repaired by the cell itself. However, when a mutation occurs in a cancer driver gene, such as tumor suppressor genes or proto-onco genes, the cell may start to grow uncontrollably allowing

[2] Francis, Z. & Stypczynska, A. (2013). *Clustering algorithms in radiobiology and DNA damage quantification.*

cancer to develop. Both proto-oncogenes and tumor suppressor genes are genes that are involved in controlling the growth and division of the cell. A mutation in a tumor suppressor gene removes their braking function, causing uncontrolled cell growth and division. When a mutation occurs in a proto-oncogene, it can turn into an oncogene. Oncogenes are genes that send out signals that allow cells to grow and divide when they are not intended to. It is estimated that approximately 6 different mutations in cancer driver genes are needed before a normal cell turns into a cancer cell [46].

Cancer is thus caused by mutations in the cancer driver genes. Once the cells are starting to grow abnormally, they are more likely to undergo additional mutations at other places on the chromosomes as well. There are several types of mutations that can occur in the DNA of a cell, e.g. nucleotide changes (A to G), but also larger structural variations which may cause copy number changes, see Figure 2.3 for a simplified representation.[3] In normal cells, each individual has two copies per chromosome referred to as copy number two. In tumor cells, however, certain modifications occur in the DNA such that the copy number of (part of) the chromosome(s) is no longer equal to two.

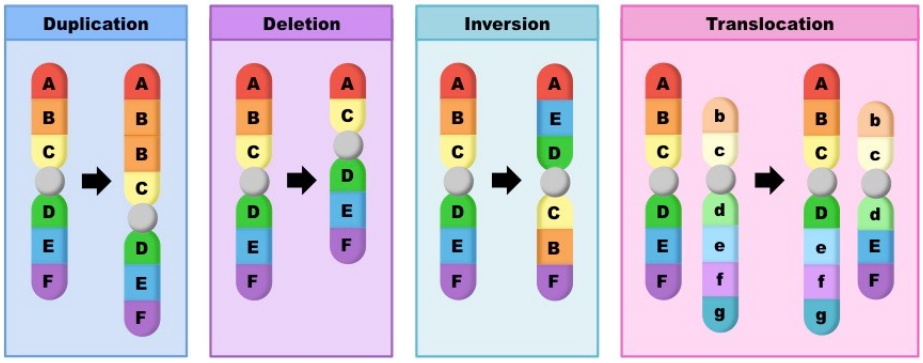


Figure 2.3: Different types of structural variations that can appear in the DNA of an individual.

In Figure 2.3, a chromosome is displayed by means of the letters A-B-C-D-E-F, where each letter stands for a certain region of the chromosome. The grey circle represents the centromere.

- In the case that a duplication takes place in the DNA, (part of) the chromosome is gained. In the figure, the B-region of the chromosome is duplicated. This implies that the individual has an extra copy for this part of the chromosome, resulting in copy number three for the B-region. The other regions of the chromosome still have copy number two. Note that copy numbers higher than three are also possible. Moreover, it is possible that all 23 pairs of chromosomes are duplicated. This phenomenon is known as *whole genome duplication*.
- When a deletion occurs, (part of) the chromosome is lost. In the figure, the B-region of the chromosome is deleted meaning that the copy number is equal to one for the B-region, but still equal to two for all other regions. A deletion on (part of) the chromosome is also called *Loss of Heterozygosity* (LOH). This name comes from the fact that an individual is heterozygous when it comes to the genetic information: the paternal and maternal genetic material will be different. When (part of) one of the two chromosomes is deleted, the heterozygosity in genetic information is lost. Next to a single deletion, a double deletion is also possible. In the case of a double loss, the total copy number for a certain region is equal to zero and all genetic information is lost.
- In the event of an inversion, the genetic information within a chromosome changes positions. In the figure above, the order of the regions has changed from A-B-C-D-E-F to A-E-D-C-B-F. For this mutation, the total copy number remains equal to two.
- For a translocation, two different chromosomes (not belonging to the same pair) exchange their genetic material. In the figure, two chromosomes are seen coded as A-B-C-D-E-F and a-b-c-d-e-f-g. In the case of a translocation, the two chromosomes interchange regions with one another

[3] https://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/block-mutations.html

resulting in two new chromosomes that are combinations of the initial chromosomes: A-B-C-D-E-F becomes A-B-C-D-e-f-g while b-c-d-e-f-g becomes b-c-d-E-F. The total copy number after a translocation is not necessarily equal to two. In Figure 2.3, it can be seen that chromosome b-c-d-e-f-g has lost its g-region after the translocation implying that the individual has lost one copy in part of this chromosome pair.

Next to a duplication, deletion, inversion and translocation, there is also a fifth type of mutation that can occur in the DNA of cancer cells called *Copy-Neutral Loss of Heterozygosity (CN-LOH)*. Figure 2.4 illustrates this type of mutation.

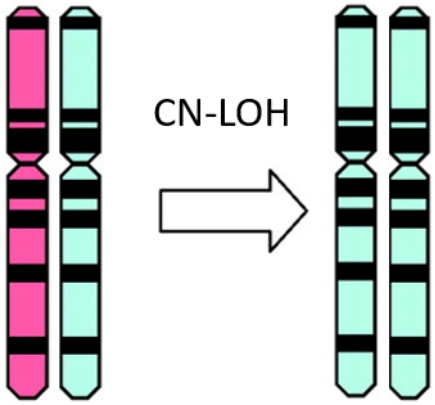


Figure 2.4: Schematic illustration of Copy-Neutral Loss of Heterozygosity.

In the figure, the two chromosomes on the left represent a maternal chromosome (pink) and paternal chromosome (blue). In the case of CN-LOH, (part of) one of the two chromosomes is lost and the remaining chromosome is duplicated. For this type of mutation, the copy number remains equal to two (copy neutral), but the heterozygosity is lost.

The five types of mutations introduced above are not mutually exclusive. For instance, a deletion at chromosome 1 and a duplication at another chromosome can arise during the same cell division. Moreover, combinations of mutations may also happen such as a duplication followed by a translocation or a deletion followed by an inversion. This implies that the DNA profile of a cancer cell may become very disarranged.

2.3. SNP arrays

Single Nucleotide Polymorphism (SNP) arrays are a type of DNA array that can be used to detect copy number changes in the DNA profiles of tumors. This section explains the definition of a SNP, how a SNP array works as well as how the data coming from a SNP array can contribute in detecting copy number changes.

2.3.1. What is a SNP?

The human genome is composed of approximately $3 \cdot 10^9$ base pairs, of which 99.9% is the same between any two individuals. The main form of genetic variation between individual genomes comes from SNPs. A SNP is a single base change in the genomic sequence (i.e. DNA) that occurs in more than 1 percent of the population. It is estimated that there are approximately 10^7 SNPs in total, implying that SNPs occur at a frequency of 1 in 1000 base pairs in the DNA on average. Most SNPs are bi-allelic, meaning that two base variations (also referred to as *alleles*) can occur at the SNP position. As described previously, DNA is composed of 4 different bases, denoted by A (adenine), C (cytosine), G (guanine) and T (thymine). An example of an allelic variation at a SNP is A/C, meaning that an individual either has adenine (A) or cytosine (C) at the specific SNP position on the chromosome.

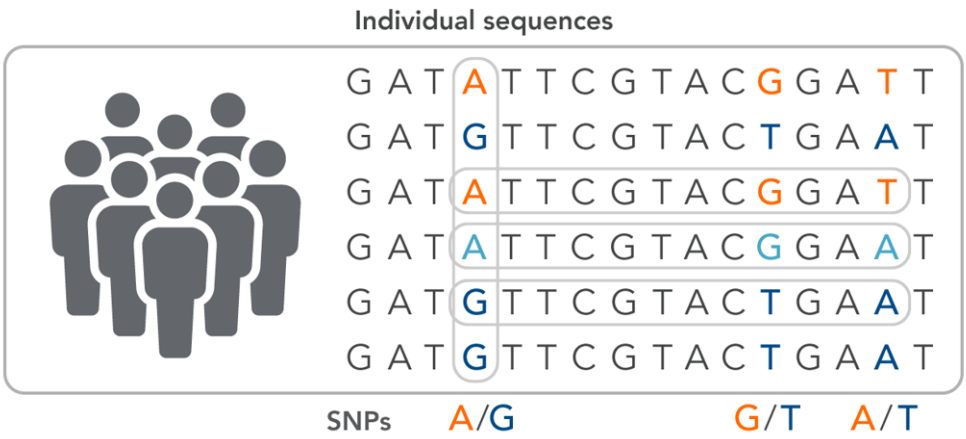


Figure 2.5: An example of how SNPs occur in individual genome sequences.

Figure 2.5 shows a simplified example of SNPs in individual genome sequences.[4] By checking both alleles in an individual, the so-called *genotype* of the SNP can be derived. The genotype specifies the combination of bases, i.e. alleles, that occur at the SNP position in the two chromosomes coming from the father and mother. For instance, in Figure 2.5 the first SNP position is either an A or a G. As a consequence, an individual can have genotype AA, AG or GG at this SNP position. In the case that an individual has genotype AA or GG, the individual is considered to be homozygous at the first SNP position, as the bases at the SNP position that come from both parents are the same. If the individual has genotype AG at the SNP position, the person is classified as heterozygous for that SNP.

2.3.2. How are SNP genotypes derived?

A SNP array, also known as a SNP chip, can be used to determine the genotypes of an individual at different SNP positions. Illumina and Affymetrix are the two most well-known biotechnology companies that produce SNP chips. The number of SNPs that can be tested by a chip ranges from thousands to hundreds of thousands. Which SNP chip to use depends on the research objective: if, for instance, only a part of the genome is of interest (e.g. this part of the genome is a sole predictor for a certain disease), using a SNP chip that tests a lower number of SNPs will suffice to derive a conclusion, while if the entire genome is to be genotyped more detail regarding the SNPs is needed. Which SNPs are included on the chip are commonly determined by the producing company.

An Illumina SNP chip consists of several compartments, where each compartment is composed of multiple *beads*. Each bead is capable to test one SNP position. DNA fragments, called *oligonucleotides*, are attached to each bead that are complementary to the bases before the particular SNP. The length of an oligonucleotide is approximately 50 bases long and does not include the SNP itself. Before the SNP chip can be used, the DNA is amplified to ensure that there are enough molecules to determine the genotype. After amplification, the SNP chip is incubated with a mix of fragmented (the amplified DNA is broken apart into smaller pieces) DNA strands, where the DNA has been *denatured* meaning that the two strands of each chromosome are separated. A DNA strand anneals to a bead if the oligonucleotide attached to the bead has the complementary sequence of the strand. This process is called *hybridization*. For instance, if the oligonucleotide attached to a bead is A-C-T-A-A-A-G and the bead tests for a A/G variation, then a DNA strand having T-G-A-T-T-T-C-*-C-T will bind to this bead where * is equal to A or G. After the strands are attached to the beads, the bead is incubated with fluorescently labeled bases which will bind to the SNP itself. When UV light is consequently shone on the bead, the bead will display a specific color which determines the genotype of the individual at that particular SNP.

[4] https://eu.idtdna.com/pages/applications/genotyping

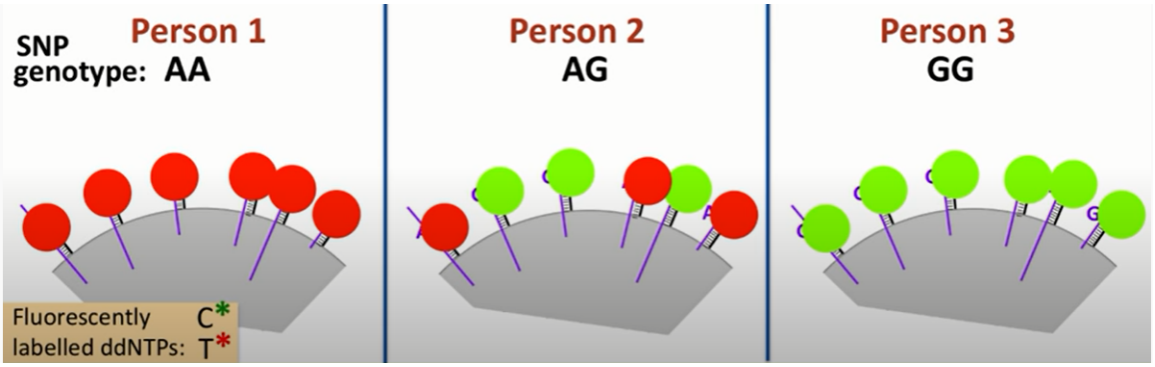


Figure 2.6: A simplified example of a bead testing the genotype of an individual at a SNP position which has an A-G variation.

Figure 2.6 shows an example of how a genotype is determined for a SNP position.[5] In Figure 2.6, the grey half-circle represents a bead on a SNP chip. The black lines are the oligonucleotides and the purple lines depict the DNA fragments that have bound to the bead by means of complementary bases. This bead tests for an A-G variation at a SNP position. The complementary base of adenine, thymine, is labeled red while the complementary base of guanine, cytosine, is labeled green. Figure 2.6 shows all three instances that can occur. If a person is homozygous AA at the SNP position, only the red labeled thymine bases will bind to the SNP in question resulting in the bead turning red when UV light is shone on it. A similar behavior occurs when the person is homozygous GG, but the bead will fluoresce green. If a person is heterozygous AG, the bead will show a mix of red and green.

A specifically designed scanner system detects the fluorescence intensities of each bead, also called the raw signal intensities. Given the raw signal intensities, the genotype of the SNP positions can be determined. Next to genotyping, SNP arrays can also be used to infer copy number changes (duplication, deletion, etc.) on the genome.

2.3.3. Inference of copy number changes

For each SNP position the two possible bases are referred to as the A and B alleles. Note that the A allele is not necessarily equal to the adenine (A) base in DNA. Which base is allele A and which is allele B is determined using a TOP/BOT approach [21]. The raw signal intensity values measured for the A and B allele are normalized using a 5-step procedure, where the signal intensity of all the SNPs are employed. The output of the normalization procedure are X and Y values for each SNP, which represent the normalized raw signal intensities for the A and B alleles, respectively.

Given X and Y, two additional measures can be computed for each SNP:

$$R = X + Y,\ \theta = \frac{\pi}{2} \arctan\left(\frac{Y}{X}\right)$$

R denotes the total signal intensity of the bead while θ displays the relative allelic signal intensity ratio of allele B. If, for example, a person has more alleles of type B, Y will be larger in comparison to X giving a larger θ.

Given R and θ, the Log R Ratio (LRR) and B Allele Frequency (BAF) can be computed. These two quantities together depict the copy number changes that have occurred on the genome. The LRR is calculated for each SNP as follows:

$$\text{LRR} = \log_2\left(\frac{R_{\text{subject}}}{R_{\text{expected}}}\right),$$

where R_{subject} is the observed total signal intensity and R_{expected} is computed from linear interpolation of genotype clusters [38]. Figure 2.7 shows an example of how R_{expected} is computed.

[5] This image is a screenshot taken from https://www.youtube.com/watch?v=Naona1y_I2U. The video in question is a very informative YouTube video that explains the process in more detail.

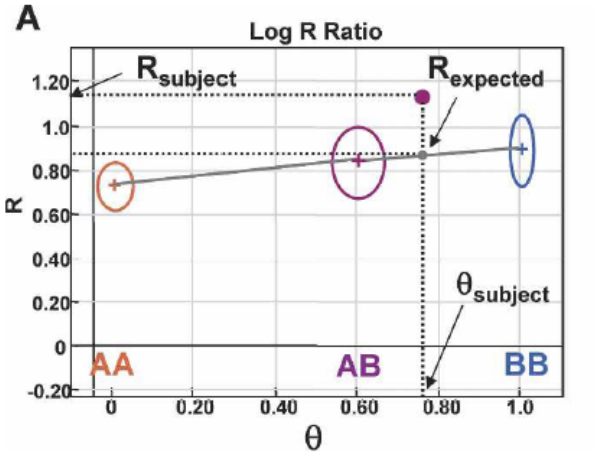


Figure 2.7: Example of how the LRR is computed. Figure retrieved from [38].

In Figure 2.7, three clusters can be seen based on the three possible genotypes at a SNP: AA, AB or BB. The three genotype clusters for a SNP are generated by the producing company of the SNP chip by training on a large set of normal reference samples. The three crosses in Figure 2.7 represent the cluster centroids which are connected through linear interpolation. Given the value of θ at the SNP, θ_{subject}, the value of R_{expected} is set equal to the point where the linear interpolation function evaluated in θ_{subject} intersects the R-axis. For the example in Figure 2.7, θ_{subject} is approximately 0.76. The linear interpolation function intersects the R-axis at approximately 0.87 for this value of θ. Therefore, $R_{\text{expected}} \approx 0.87$.

Next to the LRR, the BAF is computed as follows:

$$\text{BAF} = \begin{cases} 0 & \text{if } \theta < \theta_{AA} \\ \frac{0.5(\theta - \theta_{AA})}{\theta_{AB} - \theta_{AA}} & \text{if } \theta_{AA} \leq \theta < \theta_{AB} \\ 0.5 + 0.5\frac{\theta - \theta_{AB}}{\theta_{BB} - \theta_{AB}} & \text{if } \theta_{AB} \leq \theta < \theta_{BB} \\ 1 & \text{if } \theta > \theta_{BB} \end{cases}$$

In the BAF, θ_{AA}, θ_{AB} and θ_{BB} are the θ values for the three genotype clusters generated from a large set of normal reference samples. Figure 2.8 shows an example of how the BAF is computed.

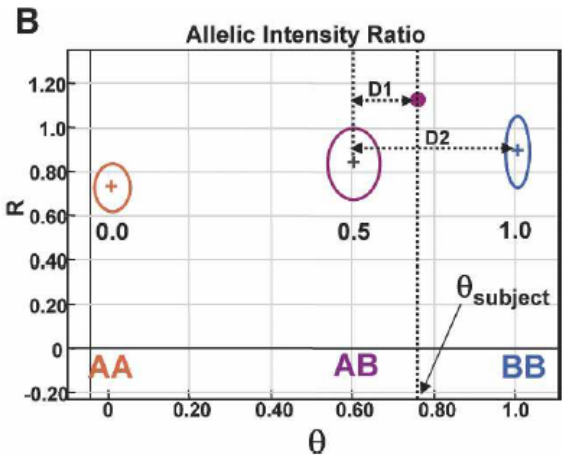


Figure 2.8: Example of how the BAF is computed. Figure retrieved from [38].

In the example in Figure 2.8, the value of θ_{subject} is in between the values of θ_{AB} and θ_{BB}. If $D1$ denotes the distance between θ_{subject} and θ_{AB} and $D2$ the distance between θ_{AB} and θ_{BB}, then it follows that

BAF = $0.5 + 0.5\frac{D1}{D2}$. If a SNP has a BAF value close to zero, it can be regarded as being homozygous AA. For an individual that is homozygous BB at a certain SNP, the corresponding BAF value will be close to 1. Finally, if an individual has genotype AB at a certain SNP its BAF value will be around 0.5. Therefore, given the BAF value, the genotype of the individual at a particular SNP can be estimated.

The LRR and BAF values for the SNP positions can be automatically computed by GenomeStudio, a free software program provided by Illumina. Once both quantities are computed, a LRR plot and a BAF plot can be generated which together give information about where copy number changes (e.g. deletion, duplication) have occurred on the genome. Figure 2.9 shows an example of the LRR and BAF plots from chromosome arm 15q.

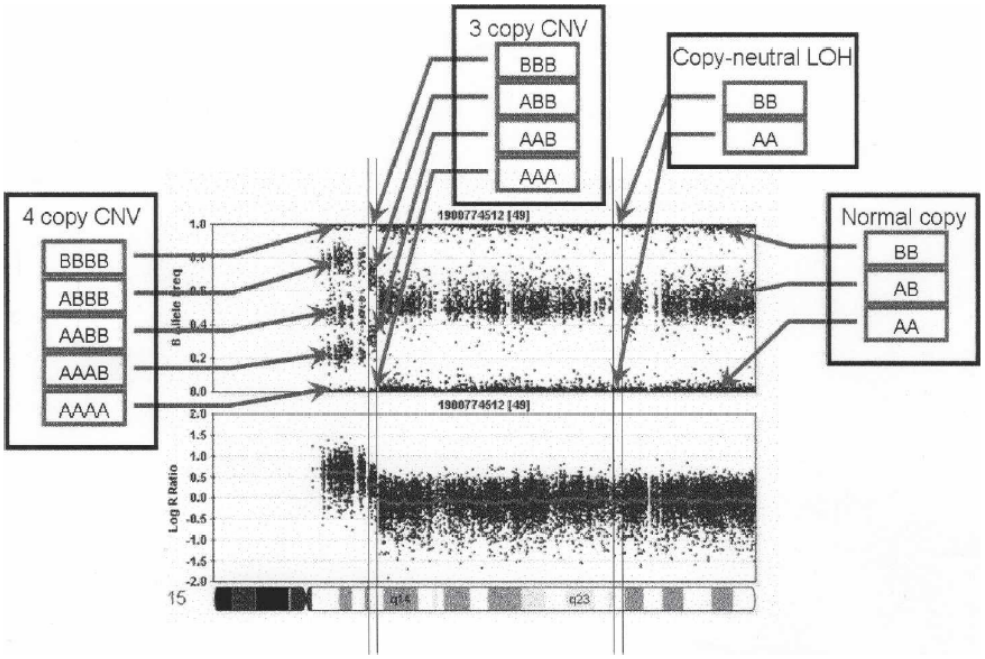


Figure 2.9: BAF (top) and LRR (bottom) plots for the chromosome 15q arm. Figure retrieved from [53].

If no copy number changes have occurred, each SNP will have two copies. The LRR is a log transformed ratio of the observed total signal intensity divided by the expected total signal intensity, where the expected total signal intensity is obtained using the genotype clusters which assume copy number two as a baseline. Therefore, in the normal setting, the LRR will be close to zero. Next to that, as there are 3 different genotypes possible per SNP (AA, AB and BB), the BAF plot will consist of three bands at 0 (AA), 0.5 (AB) and 1 (BB) in the case of two copies. This is also reflected in Figure 2.9.

When a copy number alteration occurs, the LRR and BAF plot will change accordingly as follows:

- If a certain region of a chromosome is deleted (i.e. for SNPs in this part, the individual is of genotype A or B, the SNPs on the other chromosome are missing), there is less DNA that will anneal to the beads representing the region. As a result, the observed total signal intensity in the region is lower than expected resulting in a LRR below zero. Moreover, the BAF plot shows a lack of heterozygotes (one cannot be of type AB in case of a deletion) so that the BAF values center around 0 (A) and 1 (B). This entails that the BAF plot will consist of only two bands in the region where the loss occurred. Next to a single loss, a double loss can also occur. In that case both copies are deleted yielding a copy number equal to 0. A double loss is characterized by a very low LRR and a BAF value that can be anywhere between 0 and 1 [22].

- In the case that a region undergoes a duplication, the LRR values of the region will be above zero. This comes from the fact that there is more DNA that can bind to the beads which increases the

observed total signal intensity making it larger than the expected total signal intensity. Next to a LRR above zero, the BAF plot will consist of more than three bands. In the case of a single duplication (copy number 3), a person can be of genotype AAA, AAB, ABB or BBB. Subsequently, the BAF values will center around 0 (AAA), 0.33 (AAB), 0.67 (ABB) and 1 (BBB) in the case of a single duplication. When a double duplication occurs, the LRR will be even higher compared to a single duplication and the BAF plot will consist of five different bands: 0 (AAAA), 0.25 (AAAB), 0.5 (AABB), 0.75 (ABBB) and 1 (BBBB). In general, the higher the copy number gain, the higher the LRR will be and the larger the number of bands visible in the BAF plot. This is also displayed in Figure 2.9.

- A region containing Copy-Neutral LOH can be detected by means of a LRR value around zero and only two bands in the BAF plot. In a CN-LOH region, an individual still has two copies of each allele, but the alleles in the region consist solely of one of the two parents of the individual. Since there are still two copies, the LRR will not change. However, since the individual is homozygous in the region (AA or BB), the BAF values will be centered around 0 (AA) and 1 (BB). This can also be seen in Figure 2.9.

The LRR and BAF values together thus give an overview of the copy number changes that have occurred in the genome. In this thesis, the objective is to investigate how the degree of clonality between a pair of tumors can be determined based on their genomic features. Two tumors share similar genomic features if the gain, loss and CN-LOH regions have a large overlap. The LRR and BAF plots may give an indication where certain aberrant events occur, but they do not state exactly the boundaries of these events which are needed to determine the amount of overlap in the genomic profiles. In order to quantify the clonal relatedness of a tumor pair, the following two steps need to be taken:

1. Given the LRR and/or BAF data of the two tumors, a segmentation algorithm is applied to discover what has happened to the genomes.
2. After the aberrant regions on the genomes are determined, a statistical method is employed to evaluate the degree of similarity between the two tumors.

3

Segmentation algorithms

Segmentation algorithms are capable of, given the LRR and/or BAF data of the SNP array, estimating the regions where aberrant events, also known as Copy Number Alterations (CNAs), have occurred in the DNA. A wide variety of segmentation algorithms exist such as Hidden Markov Models [12, 45, 53, 57], algorithms that are derived from change point detection algorithms [32, 33, 49] and threshold based algorithms [19, 22]. In order to quantify the clonal relatedness between two tumors, it is of importance that the segmentation algorithm estimates the CNAs of the two tumors as accurate as possible. However, a big challenge in the estimation of CNAs are the presence of confounding variables which may limit the performance of a segmentation algorithm.

Section 3.1 gives an introduction to the confounding variables and how these influence the resulting LRR and BAF data. The second section, Section 3.2, presents an overview of different segmentation algorithms. Given the results of the overview, two segmentation algorithms are chosen for this study, which are explained in more detail in the final two sections of this chapter, Sections 3.3 and 3.4.

3.1. Confounding variables

Confounding variables are factors which occur either in the DNA of the tumor or when the DNA is prepared for the SNP array. The confounding variables influence the LRR and BAF profiles in such a way that CNAs are harder to determine by the segmentation algorithm thereby decreasing the sensitivity of the latter. Three different confounding variables can be defined in this case: normal cell contamination, aneuploidy and tumor heterogeneity.

3.1.1. Normal cell contamination

Normal cell contamination takes place when the DNA is prepared for the SNP array. When DNA is extracted from the tumor tissue, a certain amount of normal cells present within the tumor or surrounding the tumor may be included leading to a mixture of DNA derived from normal and tumor cells. As a consequence, the LRR and BAF profiles will reflect a combination of the tumor genome and the normal genome. Figure 3.1 shows how the tumor percentage influences the LRR and BAF plots in case of a deletion.

Figure 3.1 was created by mixing DNA from a cancer cell line (i.e. cells that can be kept growing/multiplying in laboratory conditions and that only contains cancer cells, without normal cells) and matched normal DNA in different proportions and consequently evaluated by means of SNP arrays. When the tumor percentage is equal to 100%, a loss can be seen in the plots (depicted by the red square). For this region, the LRR values are lower and the BAF values center around 0 and 1. As the percentage of normal cells present in the DNA mixture increases, the LRR values in the loss region go up and the two bands in the BAF plot are coming closer together. The reason why the two bands are coming closer together has to do with the fact that the DNA mixture consists of (A, AB) and (B, AB) genotypes. When the tumor cellularity is equal to 0%, the tumor-specific loss is no longer present and the BAF and LRR plots show the expected profile for a normal cell. In general, the larger the normal

percentage, the less CNAs are apparent in the data. As a result, the segmentation algorithm may struggle to correctly find the boundaries of the aberrant events.

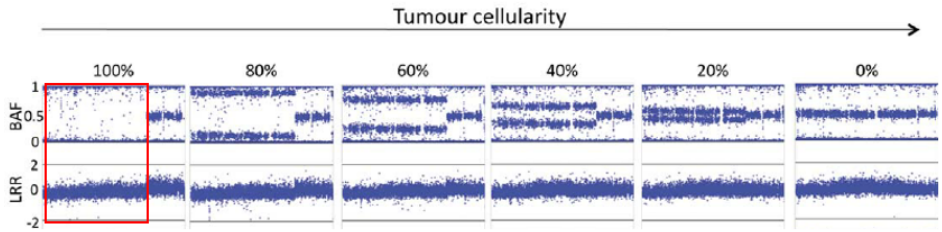


Figure 3.1: BAF and LRR plots for different tumor percentages. Figure retrieved from [43].

3.1.2. Aneuploidy

Aneuploidy is the presence of an abnormal number of chromosomes in the cell. Normal cells are diploid, meaning that there are 23 sets of two chromosomes. However, during cancer development, CNAs can cause the cell to become non-diploid. The ploidy expresses the average number of chromosome sets that a person has. For example, when a whole genome duplication occurs an individual has 4 copies of each chromosome and the individual is considered to be tetraploid. A ploidy of 2.4 could imply that 60% of the chromosomes are diploid and the other 40% is triploid.

In the normalization procedure, the raw signal intensity values are normalized with respect to the overall signal intensity value. In diploid samples SNPs have a LRR equal to zero. In non-diploid samples, on the other hand, the normalization procedure assigns a LRR equal to zero to SNPs whose copy number corresponds to the average copy number found in the sample [1]. As a result, SNPs with fewer copies (i.e. a diploid SNP in a tri- or tetraploid sample) will have a negative LRR.

Figure 3.2 shows the LRR and BAF plots of two chromosomes from a near-triploid neuroblastoma sample.

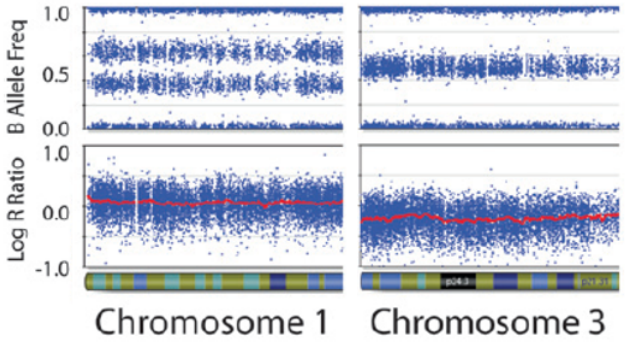


Figure 3.2: BAF and LRR data of two chromosomes from a near-triploid neuroblastoma sample. Figure retrieved from [1].

As the sample is near-triploid, the normalization procedure causes the LRR values to be close to zero for those SNPs whose copy number are three. This can be clearly seen in the LRR and BAF plots of chromosome 1 where the LRR values lie around zero. If it is unknown that the sample is triploid, three bands are expected to arise in the BAF plot. However, the BAF plot does not show three but four bands corresponding to the four different genotypes (AAA, ABB, ABB, BBB) that can occur in case of three copies. Hence, chromosome 1 has copy number three. In the right panel, the LRR values lie below zero for chromosome 3 and, without knowing that the sample is triploid, two bands are expected in the BAF plot. The BAF plot of chromosome 3 shows three bands instead of two, implying that this chromosome has copy number two. In general, aneuploidy can only be detected if both the LRR and BAF values are known.

3.1.3. Tumor heterogeneity

Tumor heterogeneity refers to the case when the tumor consists of several so-called sub-clones. Tumors are continuously mutating. As a consequence, it can happen that some cells within the tumor harbor a certain mutation which is not present in other cells of the tumor. As the cells come from the same origin but show different mutations, the tumor is said to have different subclones. Figure 3.3 shows a simplified example of a tumor consisting of three different subclones evaluated at 5 SNPs. Next to that, the normal cell contamination in the DNA which is assayed on the SNP array is equal to 20%.

Proportion of sample	Cell Type	SNP				
		1	2	3	4	5
20%	Normal	AA	AB	BB	AB	AB
30%	Clone 1	A	AB	BB	AABB	ABB
30%	Clone 2	AA	AB	BBB	AABB	AAB
20%	Clone 3	AA	B	BBB	AABB	AB

Figure 3.3: Simplified example of a tumor consisting of three different subclones and normal cell contamination evaluated at 5 SNPs. Figure retrieved from [57].

Looking at, for instance, SNP 5 in Figure 3.3 it can be seen that the normal genotype is AB. In clone 1 a duplication of the B allele has occurred while clone 2 has an extra A allele. Clone 3 does not show any gain or loss at SNP 5. In total it can be seen that of the DNA evaluated at the SNP array, 60% has a gain at SNP 5 while 40% has no gain. As a result, a less pronounced gain will be present in the LRR and BAF plot. In the same way as normal cell contamination, the LRR and BAF profiles will less clearly reflect genomic aberrations in case of tumor heterogeneity.

3.2. An overview of the segmentation algorithms

As mentioned before, many segmentation algorithms exist which can, given the LRR and BAF data, determine the regions where genomic aberrations have occurred. Table 3.1 gives an overview of several segmentation algorithms. The table specifies for each algorithm which input variables are taken into account (only the LRR or both the LRR and BAF), if CN-LOH can be detected, the type of output of the segmentation algorithm (states or gain/loss) and if the algorithm corrects for any of the three confounding variables. Note that, in case the segmentation algorithm outputs states, the number of states is given in brackets. For instance, PennCNV outputs six different states. ASCAT, on the other hand, outputs the copy number for each SNP.

In Table 3.1, it can be seen that many segmentation algorithms (CBS [33, 49], GLAD [19], BIC [56], L^1 penalization [18], SaRa [32], CGHcall [47], CGHcall* [39], GISTIC [29]) only use the LRR as an input variable. Since the BAF is not considered for these algorithms, CN-LOH cannot be detected by these algorithms. As described in Section 2.3.3, CN-LOH is characterized by LRR values close to zero and a BAF plot consisting of two bands corresponding to the AA and BB genotype. Being able to distinguish between the normal situation and CN-LOH is important when quantifying the clonal relatedness between two tumors. For instance, consider the case where there are two tumors which both show a relatively flat LRR profile, but one tumor has many CN-LOH events and the other tumor is mainly normal. If the clonality between the two tumors is only based on the LRR, the pair will be erroneously classified as very similar. Incorporating the BAF values in the segmentation algorithm will therefore increase the sensitivity to distinguish clonality.

	LRR	BAF	CN-LOH	states/gain&loss	norm. cont.	aneuploidy	subclones
PennCNV	x	x	x	states (6)	-	-	-
QuantiSNP	x	x	x	states (6)	-	-	-
GenoCN	x	x	x	states (6 or 9)	x	-	-
CBS	x	-	-	gain&loss	-	-	-
cnvPartition	x	x	x	states (14)	-	-	-
GLAD	x	-	-	gain&loss	-	-	-
BIC	x	-	-	unknown	-	-	-
L^1 pen.	x	-	-	unknown	-	-	-
SaRa	x	-	-	gain&loss	-	-	-
CGHcall	x	-	-	both	-	-	-
CGHcall*	x	-	-	both	-	x	-
ASCAT	x	x	x	states (CN)	x	x	-
OncoSNP	x	x	x	states (21)	x	x	x
GISTIC	x	-	-	gain&loss	-	-	-

Table 3.1: An overview of different segmentation algorithms. A cross implies that the aspect is taken into account by the segmentation algorithm while a bar entails that the aspect is not considered by the algorithm.

Given the overview of Table 3.1, ASCAT and OncoSNP appear to be the most sophisticated algorithms. These two algorithms are capable to detect CN-LOH and also correct for at least two out of three confounding variables.

[39] has investigated the performance of GISTIC, GenoCN, CGHcall, CGHcall*, ASCAT and OncoSNP on synthetic tumor data to see how well the algorithms were able to detect gains and losses. Of the algorithms examined, OncoSNP proved to be one of the best algorithms in identifying both short and long genomic aberrations. ASCAT performed less well, mainly having trouble in recognizing losses. Therefore, OncoSNP seems like the most suitable candidate algorithm to use for the segmentation of the profiles. However, OncoSNP is more computationally intensive than ASCAT and outputs multiple models for each sample. Moreover OncoSNP is less detailed when it comes to estimating the ploidy and normal cell fraction of the tumor. Hence, both algorithms were employed in this study.

The next two sections explain the two chosen segmentation algorithms in more detail.

3.3. ASCAT

The Allele-Specific Copy number Analysis of Tumors (ASCAT) algorithm takes into account both normal cell contamination of the tumor as well as aneuploidy. Before the workings of the ASCAT algorithm can be introduced, the following background information needs to be known. For a diploid individual whose tumor does not have any normal cell contamination the LRR and BAF value at the i-th SNP, modeled by r_i and b_i respectively, can be approximated as follows:

$$\begin{aligned} r_i &= \gamma \log_2\left(\frac{n_{A,i}+n_{B,i}}{2}\right), \\ b_i &= \frac{n_{B,i}}{n_{A,i}+n_{B,i}}, \end{aligned}$$

where $n_{A,i}$ and $n_{B,i}$ are the number of A and B alleles at SNP i and γ is a constant depending on the technology of the SNP array being used [48]. For Illumina arrays, γ is usually set equal to 0.55.

Tumor aneuploidy causes a shift in the LRR while the BAF remains unaffected. If ψ is the ploidy, the LRR can be rewritten as:

$$r_i = \gamma \log_2\left(\frac{n_{A,i}+n_{B,i}}{\psi}\right).$$

Note that if a person is diploid $\psi = 2$ so that the previous equation is obtained.

Moreover, the influence of normal (non-aberrant) cells on the total copy number at a SNP can be modeled as follows:

$$n_{\text{total}} = \rho n_{\text{tumor}} + (1-\rho) n_{\text{normal}},$$

where ρ is the percentage of tumor (aberrant) cells for $\rho \in [0,1]$, n_{tumor} the copy number in the tumor cells and n_{normal} the copy number in the normal cells. Using the fact that normal cells always have a copy number of two, it follows that:

$$\begin{aligned} r_i &= \gamma \log_2 \left(\frac{2(1-\rho) + \rho(n_{A,i} + n_{B,i})}{\psi} \right), \\ b_i &= \frac{1 - \rho + \rho n_{B,i}}{2(1-\rho) + \rho(n_{A,i} + n_{B,i})}, \end{aligned}$$

where ψ is modeled as $\psi = 2(1-\rho) + \rho\psi_t$, with ψ_t the tumor ploidy. Note that the equation for b_i as stated above only holds for SNPs that are heterozygous (AB) in the normal cells as in that case there is one B-allele present. Being heterozygous in the normal cells is also called *germline* heterozygous, where germline implies that the genotype is hereditary.

In the ASCAT algorithm, germline homozygous SNPs are not informative, as they remain homozygous when a duplication or deletion has occurred. For instance if a person is homozygous AA at a SNP in the normal setting, then this person will remain of type A (for instance, A when deletion occurs, AAA in a single duplication). The BAF value will remain around zero for this SNP. The same applies when the person is germline homozygous BB. Therefore, the ASCAT algorithm focuses on only those SNPs that are germline heterozygous in the normal cells to infer where copy number aberrant events have occurred. Given the system of equations above, the copy numbers for the A and B allele can be expressed in terms of the LRR value r_i, BAF value b_i, tumor ploidy ψ_t, aberrant cell fraction ρ and technology parameter γ as follows:

$$\begin{aligned} \hat{n}_{A,i} &= \frac{\rho - 1 + 2^{\frac{r_i}{\gamma}} (1 - b_i)(2(1-\rho) + \rho\psi_t)}{\rho}, \\ \hat{n}_{B,i} &= \frac{\rho - 1 + 2^{\frac{r_i}{\gamma}} b_i (2(1-\rho) + \rho\psi_t)}{\rho}. \end{aligned}$$

The ASCAT algorithm starts by applying the Allele-Specific Piecewise Constant Fitting (ASPCF) algorithm to segment the data. The input for the algorithm are the LRR and the BAF data. SNPs where the BAF is greater than 0.7 or smaller than 0.3 in the normal cells are excluded as for these SNPs the individual is assumed to be homozygous AA or BB. Note that normal data of the individual is needed for this step. If matched normal data is not available, ASCAT can predict the germline genotypes for certain SNP arrays.[1]

Once the germline homozygous SNPs are deleted from the data, the ASPCF algorithm is applied to each chromosome arm separately. Before the change points are determined, the BAF values that lie above 0.5 are replaced by 1 minus the BAF value, i.e. if $b_i > 0.5$ for SNP i, b_i is replaced by $1 - b_i$ in the ASPCF algorithm. The algorithm fits piecewise constant regression functions simultaneously to the LRR and BAF values. Let $(r_1, \ldots, r_n)$ and $(b_1, \ldots, b_n)$ be the LRR and BAF values at SNPs $x_1, \ldots, x_n$. The ASPCF algorithm seeks an optimal partitioning for each chromosome arm into segments $I_1, \ldots, I_Q$ where Q is unknown [48]. The optimal number of segments Q and corresponding boundaries of the Q segments are found by minimizing the following quantity:

$$\sum_{j=1}^{Q} \sum_{i \in I_j} \left[w(r_i - \text{ave}(\{r_s\}_{s \in I_j}))^2 + (1-w)(b_i - \text{ave}(\{b_s\}_{s \in I_j}))^2 \right] + \lambda Q,$$

where $w = 0.5$ for convenience, ave() is the average function and λ the penalty parameter. Note that in the expression, the minimization is done with respect to the number of segments Q as well as the

[1] A list of supported SNP array platforms can be found at: https://www.crick.ac.uk/research/labs/peter-van-loo/software

assignment of the SNPs to segments. For each value of Q, the optimal boundaries of the segments $I_1, \ldots, I_Q$ are determined by means of trying out all possible combinations of consecutive SNPs in Q segments. Once the optimal change points are determined, piecewise constant functions are fitted to the segments where the function value equals the mean of the LRR and the BAF values in the segment.

The final step of the ASPCF algorithm consists of determining whether the segment shows allelic bias (e.g. has genotype AAB, ABB etc.). If an allelic bias is present, the average BAF value in the segment should be mirrored around 0.5 to reflect the different genotypes that occur in the segment. In order to determine whether allelic bias is present, the mean deviation d and standard deviation s from 0.5 are computed for each BAF segment. For a given constant $\tau > 0$, two values symmetric around 0.5 for the BAF segment are returned (implying allelic bias) if $d \geq \tau s$, otherwise a single value of 0.5 is returned. The default setting in the ASPCF algorithm is $\lambda = 50$, $\tau = \sqrt{3}$ and a minimal segment length of 6. Figure 3.4 shows an example of the output of the ASPCF algorithm.

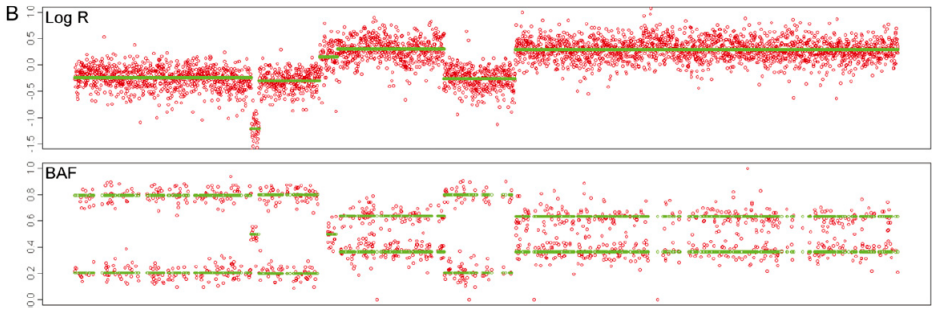


Figure 3.4: Example of the output of the ASPCF algorithm. Figure retrieved from [48].

After the data is segmented, the tumor ploidy ψ_t and tumor purity ρ are estimated. For this, the following steps are taken:

1. A (ρ, ψ_t) grid is defined for values $\rho = (0.1, 0.11, \ldots, 1)$ and $\psi_t = (1, 1.05, \ldots, 5.4)$.
2. For each parameter combination, the total distance to a whole-number solution is calculated by summing over all SNPs:

$$d(\rho, \psi_t) = \sum_i w_i((\hat{n}_{A,i}(\rho, \psi_t) - \text{round}(\hat{n}_{A,i}(\rho, \psi_t)))^2 + (\hat{n}_{B,i}(\rho, \psi_t) - \text{round}(\hat{n}_{B,i}(\rho, \psi_t)))^2),$$

where $w_i = 1$ if the SNP is in a segment showing allelic bias and $w_i = 0.05$ if no allelic bias is present. A larger weight is assigned to the segments showing allelic bias as these are deemed more likely to be aberrant segments [48]. Once the distances $d(\rho, \psi_t)$ are computed for every combination a *sunrise* plot can be constructed.

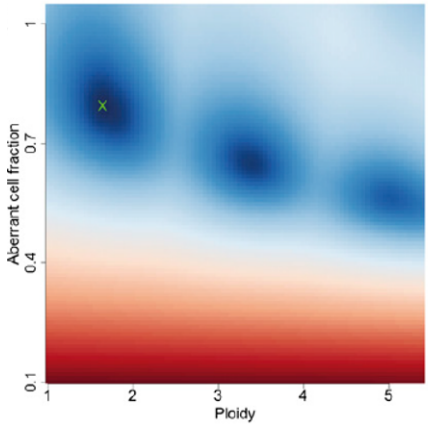


Figure 3.5: Example of a sunrise plot. Figure retrieved from [48].

Figure 3.5 shows an example of a sunrise plot where the ploidy is plotted on the x-axis and the aberrant cell fraction on the y-axis. The color in the plot resembles the size of the distance: a solution with a small distance is colored blue, while a solution with a large distance is colored in red. The smaller the distance, the darker the blue color will be. The cross in the plot indicates the optimal solution (see below).

3. Given the sunrise plot, a goodness of fit g is computed for all local minima using a linear rescaling: $g = 100$ if $d = 0$ and $g = 0$ if $d = \sum_i w_i(2 \cdot 0.25^2)$. According to [48], the value 0.25 is selected as a reasonable maximum distance (averaged over all SNPs), as the goodness of fit is only computed for local minima.

4. After the goodness of fit values are computed for all local minima, ASCAT first attempts to find an optimal local minimum in the internal values of the grid, i.e. $\rho \in (0, 1)$ and $\psi_t \in (1, 5.4)$. The local minima located at the borders are thus excluded from the analysis. For each local minima within the internal grid, ASCAT excludes those which correspond to unlikely interpretations. For this, the following exclusion criteria are used:

   (a) A ploidy ψ_t outside of the range $(1.6, 4.8)$.

   (b) A too low fraction of tumor cells $\rho < 0.2$.

   (c) Floating solutions. Floating solutions are solutions that state that the tumor has a higher ploidy while there is no evidence in the data of a higher ploidy. For instance, a local minima may arise at ploidy 4, but if the LRR values are relatively low and the BAF plot mainly shows three bands, the local minima at ploidy 4 is unlikely to be the optimal solution.

   (d) A goodness of fit below 80 percent.

   If one candidate is left after the exclusion procedure, the optimal solution (ρ, ψ_t) is reported. If multiple candidates are left, the solution with the highest goodness of fit is reported.

   If no solution with a goodness of fit above 80% is found in the internal grid, the procedure above is repeated including the local minima situated at the borders of the grid.

   If including the boundaries still does not yield a solution with a goodness of fit above 80%, the profile cannot be segmented by ASCAT. In that case, ASCAT considers the array to have failed. This mainly happens with LRR and BAF data that is quite noisy and which consists of many short segments. In [48], ASCAT was applied to 112 breast carcinomas. Of the 112 samples, a considerable fraction (19%) does not fit the ASCAT model well enough to construct a segmentation profile.

In Figure 3.5, three local minima can be seen. All local minima lie inside the internal grid. The optimal solution states that the aberrant cell fraction is equal to 80% while the ploidy is estimated at 1.77.

Once the optimal tumor ploidy ψ_t and tumor purity ρ are found, the estimated copy numbers for the A and B allele in the segment can be estimated. Let r_s be the mean of the LRR values in the segment and b_s the mean of the BAF value in the segment. The copy numbers for the A and B allele in segment s are estimated as follows:

$$\hat{n}_{A,s}^{\text{ASCAT}} = \text{round}\left(\frac{\rho - 1 + 2^{\frac{r_s}{\gamma}}(1 - b_s)(2(1-\rho) + \rho\psi_t)}{\rho}\right),$$

$$\hat{n}_{B,s}^{\text{ASCAT}} = \text{round}\left(\frac{\rho - 1 + 2^{\frac{r_s}{\gamma}} b_s(2(1-\rho) + \rho\psi_t)}{\rho}\right).$$

Figure 3.6 shows a plot of an ASCAT profile. In the plot, a red line and a green line are plotted. The green line corresponds to the allele with the lowest copy number, $\min(\hat{n}_{A,s}^{\text{ASCAT}}, \hat{n}_{B,s}^{\text{ASCAT}})$ while the red line corresponds to the allele with the highest copy number, $\max(\hat{n}_{A,s}^{\text{ASCAT}}, \hat{n}_{B,s}^{\text{ASCAT}})$.

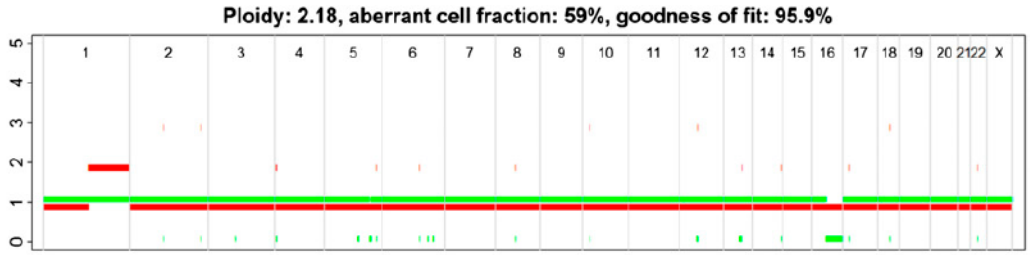


Figure 3.6: Example of an ASCAT profile. On the x-axis we have the chromosome position, on the y-axis the copy numbers. Figure retrieved from [48].

If both the A and B allele have the same estimated copy number, the green and red line will be on the same level. For instance, in the case of copy number two the red and green line will both lie at 1 as can be seen in Figure 3.6. The ploidy number of the sample in Figure 3.6 is estimated at 2.18 and the aberrant cell fraction estimated as 59%.

ASCAT is available as an R package and can be downloaded from Github. In order to install ASCAT directly from R the package `devtools` is required.

3.4. OncoSNP

[57] has developed OncoSNP, which is a Hidden Markov Model (HMM) consisting of 21 states in total. Given the LRR and BAF data, the model tries to predict the most likely state, i.e. copy number, at each SNP position. When predicting, the algorithm takes into account normal cell contamination, tumor ploidy and tumor heterogeneity.

Figure 3.7 shows the 21 different states that occur in the OncoSNP model.

Table 1 OncoSNP tumor states

Tumor states			
Tumor state	Tumor copy number	Allowable tumor-normal genotypes	Description
1	0	(-, AA), (-, AB), (-, BB)	Homozygous deletion
2	1	(A, AA), (A, AB), (B, AB), (B, BB)	Hemizygous deletion
3	2	(AA, AA), (BB, BB), (AB, AB)	Normal
4	3	(AAA, AA), (AAB, AB), (ABB, AB), (BBB, BB)	Single copy duplication
5	4	(AAAA, AA), (AAAB, AB), (ABBB, AB), (BBBB, BB)	4n monoallelic amplification
6	4	(AAAA, AA), (AABB, AB), (BBBB, BB)	4n balanced amplification
7	5	(AAAAA, AA), (AAAAB, AB), (ABBBB, AB), (BBBBB, BB)	5n monoallelic amplification
8	5	(AAAAA, AA), (AAABB, AB), (AABBB, AB), (BBBBB, BB)	5n unbalanced amplification
9	6	(AAAAAA, AA), (AAAAAB, AB), (ABBBBB, AB), (BBBBBB, BB)	6n unbalanced amplification
10	6	(AAAAAA, AA), (AAAABB, AB), (AABBBB, AB), (BBBBB, BB)	6n unbalanced amplification
11	6	(AAAAAA, AA), (AAABBB, AB), (BBBBB, BB)	6n unbalanced amplification
12	2	(AA, AA), (AA, AB), (BB, AB), (BB, BB)	2n somatic LOH
13	3	(AAA, AA), (AAA, AB), (BBB, AB), (BBB, BB)	3n somatic LOH
14	4	(AAAA, AA), (AAAA, AB), (BBBB, AB), (BBBB, BB)	4n somatic LOH
15	5	(AAAAA, AA), (AAAAA, AB), (BBBBB, AB), (BBBBB, BB)	5n somatic LOH
16	6	(AAAAAA, AA), (AAAAAA, AB), (BBBBBB, AB), (BBBBBB, BB)	6n somatic LOH
17	2	(AA, AA), (BB, BB)	2n germline LOH
18	2	(AAA, AA), (BBB, BB)	3n germline LOH
19	2	(AAAA, AA), (BBBB, BB)	4n germline LOH
20	2	(AAAAA, AA), (BBBBB, BB)	5n germline LOH
21	2	(AAAAAA, AA), (BBBBBB, BB)	6n germline LOH

Figure 3.7: Table of the 21 different states considered by OncoSNP. Figure retrieved from [57].

Let $\mathrm{x_i}$ be the tumor state at SNP i and define $(x_{i,t}, x_{i,n})$ as the copy numbers at the i-th SNP in the tumor cells and normal cells respectively. For instance, $(x_{i,t}, x_{i,n}) = (1, 2)$ would imply that a loss has occurred at the i-th SNP since the copy number in the tumor tissue is equal to 1. Note that $x_{i,n}$ is always equal to 2. Moreover, define $\mathrm{z_i} = (z_{i,t}, z_{i,n})$ as the B allele count for the tumor and normal genotypes

respectively.

The combination of $(z_{i,n}, x_{i,n})$ and $(z_{i,t}, x_{i,t})$ define the normal and tumor genotypes at the i-th SNP. For instance, if $(z_{i,n}, x_{i,n}) = (1, 2)$ and $(z_{i,t}, x_{i,t}) = (2, 3)$, then the normal genotype corresponds to AB (2 copies, 1 B allele) and the tumor genotype to ABB (3 copies, 2 B alleles), so that a gain in the B allele has occurred at the i-th SNP.

As can be seen in Figure 3.7, each tumor genotype is associated with one of the three normal genotype classes: AA, AB or BB. For instance, in tumor state 4, tumor genotype AAAB is associated with normal genotype AB as AAAB can only occur if allele A is duplicated twice. The normal genotype class at SNP i will be referred to as l_i. The possible values of the genotype class l_i are 1 (corresponding to genotype AA), 2 (corresponding to genotype AB) or 3 (corresponding to genotype BB). Figure 3.8 shows an example of normal genotype classes for a small set of tumor states.

Tumour Alteration $x_i \rightarrow \{x_{i,n}, x_{i,t}\}$	Genotypes $z_i \vert x_i \rightarrow \{z_{i,n}, z_{i,t}\} \vert x_i$	Genotype Class l_i	Description
$1 \rightarrow \{2,0\}$	$1 \rightarrow \{0,0\}$ $2 \rightarrow \{1,0\}$ $3 \rightarrow \{2,0\}$	$1 \rightarrow$ HomA $2 \rightarrow$ HetAB $3 \rightarrow$ HomB	Homozygous deletion
$2 \rightarrow \{2,1\}$	$1 \rightarrow \{0,0\}$ $2 \rightarrow \{1,0\}$ $3 \rightarrow \{1,1\}$ $4 \rightarrow \{2,1\}$	HomA HetAB HetAB HomB	Hemizygous deletion
$3 \rightarrow \{2,2\}$	$1 \rightarrow \{0,0\}$ $2 \rightarrow \{1,1\}$ $3 \rightarrow \{2,2\}$	HomA HetAB HomB	Normal copy number
$4 \rightarrow \{2,3\}$	$1 \rightarrow \{0,0\}$ $2 \rightarrow \{1,1\}$ $3 \rightarrow \{1,2\}$ $4 \rightarrow \{2,3\}$	HomA HetAB HetAB HomB	Duplication
$5 \rightarrow \{2,2\}$	$1 \rightarrow \{0,0\}$ $2 \rightarrow \{1,0\}$ $3 \rightarrow \{1,2\}$ $4 \rightarrow \{2,2\}$	HomA HetAB HetAB HomB	Copy-neutral LOH

Figure 3.8: Small subset of tumor alterations showing the different genotype classes l_i which are possible for each alteration. Figure retrieved from [57].

Let π_0 denote the percentage of normal cell contamination and let π_i, $i = 1, \ldots, n$ denote the proportion of the tumor cells having the normal genotype at SNP i. Note that π_i models the tumor heterogeneity. The data $\{\mathrm{y_i}\}_{i=1}^n$ is a 2-dimensional vector: $\mathrm{y_i} = [r_i, b_i]'$, where r_i is the LRR value and b_i the BAF value at SNP i.

Define $\mathrm{x} = \{\mathrm{x_i}\}_{i=1}^n$, $\mathrm{z} = \{\mathrm{z_i}\}_{i=1}^n$ and $\pi = \{\pi_i\}_{i=1}^n$ as the vectors representing the tumor states, B allele counts and tumor heterogeneity percentages at all the n SNPs respectively. Given $(\mathrm{x}, \mathrm{z}, \pi, \pi_0)$, the model assumes that $\mathrm{y_i}$ is distributed according to a mixture of Student t-distributions consisting of $K+1$ components. In other words:

$$\mathrm{y_i}|\mathrm{x_i}, \mathrm{z_i}, k_i, \mathrm{m}, \delta, \Sigma = \begin{cases} St(\mathrm{y_i}; \mathrm{m}(\mathrm{x_i}, \mathrm{z_i}, \pi_i) + \delta_{k_i}^{l_i}, \Sigma_{k_i}^{l_i}, \nu), & k_i = 1, \ldots, K, \\ St(\mathrm{y_i}; 0, \Sigma_\eta, \nu), & k_i = 0, \end{cases}$$

where k_i denotes the k-th mixture component of SNP i and $St(\mathrm{y_i}; \mathrm{m}(\mathrm{x_i}, \mathrm{z_i}, \pi_i) + \delta_{k_i}^{l_i}, \Sigma_{k_i}^{l_i}, \nu)$ is the probability density function of the Student t-distribution with mean $\mathrm{m}(\mathrm{x_i}, \mathrm{z_i}, \pi_i) + \delta_{k_i}^{l_i}$, covariance matrix $\Sigma_{k_i}^{l_i}$ and ν degrees of freedom associated with mixture component k_i and genotype class l_i. Moreover, an outlier class is used which assumes that $\mathrm{y_i}$ has a Student t-distribution with mean zero and a large variance Σ_η. Note that the genotype class l_i at SNP i is uniquely defined given $(\mathrm{x_i}, \mathrm{z_i})$ and the label l_i is included for notational convenience [57].

In a mixture model, each component gets assigned a weight which reflects the probability that the

distribution of the component is chosen. For instance, if a mixture model consists of three components and the weights for each component are $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$, then the distributions of components 1 and 3 are chosen with probability $\frac{1}{4}$ and the distribution of component 2 is chosen with probability $\frac{1}{2}$. In the OncoSNP model, the mixture component k_i is drawn from a multinomial distribution:

$$k_i|\mathrm{w}_l \sim Mn(\mathrm{w}_l),$$

where w_l are the mixture weights of the l-th genotype class. As stated previously, given $\mathrm{x_i}$ and $\mathrm{z_i}$, the genotype class is uniquely defined. Hence, the mixture component is drawn from a multinomial distribution where the weights depend on the genotype class. Note that w_l is a vector consisting of $K+1$ elements as there are $K+1$ components. A prior is assigned to the weights of the l-th genotype class which will be introduced later in this section.

Given the tumor alteration state $\mathrm{x_i}$, the conditional distribution of $\mathrm{z_i}$ given $\mathrm{x_i}$ is assumed to follow a multinomial distribution:

$$\mathrm{z_i}|\mathrm{x_i} = j, \mathrm{q} \sim Mn(\mathrm{q}^j),$$

where q^j is a vector of genotype probabilities associated to the j-th tumor state.

Next to that, the mean parameter m of the Student t-distribution can be expressed as a two-dimensional vector corresponding to the LRR and BAF data:

$$\mathrm{m}(\mathrm{x_i}, \mathrm{z_i}, \pi_i) = [m_r(\mathrm{x_i}, \pi_i), m_b(\mathrm{x_i}, \mathrm{z_i}, \pi_i)],$$

where m_r is the mean parameter corresponding to the LRR data and m_b the mean parameter corresponding to the BAF data. It is assumed that:

$$m_r(\mathrm{x_i}, \pi_i) = [\pi_i(1-\pi_0) + \pi_0]\overline{r}_{x_{i,n}} + (1-\pi_i)(1-\pi_0)\overline{r}_{x_{i,t}} + \beta_0 + \beta_1 g_i,$$

where $\overline{r}$ is the average LRR value given that there are $x_{i,t}$ copies in the tumor data and $x_{i,n} = 2$ copies in the normal data, β_0 is the baseline correction of the LRR (taking into account aneuploidy) and g_i is the local GC content (percentage of cytosine-guanine pairs that occur on average) at SNP i. The mean parameter of the BAF data equals:

$$m_b(\mathrm{x_i}, \mathrm{z_i}, \pi_i) = \frac{[\pi_i(1-\pi_0) + \pi_0]z_{i,n} + (1-\pi_i)(1-\pi_0)z_{i,t}}{[\pi_i(1-\pi_0) + \pi_0]x_{i,n} + (1-\pi_i)(1-\pi_0)x_{i,t}}.$$

The other distribution parameters in the OncoSNP algorithm are modeled by a prior distribution. The weights of the l-th genotype class are modeled by a Dirichlet distribution:

$$\mathrm{w}^l \sim \mathrm{Dir}(\alpha^w),$$

where typically $\alpha^w = (1, \ldots, 1)$ so that the prior is uniform.

The other parameters of the Student t-distribution are modeled as follows:

$$\begin{aligned} \delta_k^l|\tau, \Sigma_k^l &\sim N(0, \tau\Sigma_k^l),\ k = 1, \ldots, K,\ l = 1, 2, 3, \\ \Sigma_k^l|\gamma, \Sigma_0 &\sim IW(\gamma, \Sigma_0),\ k = 1, \ldots, K,\ l = 1, 2, 3, \end{aligned}$$

where IW denotes the Inverse Wishart distribution with parameter γ and covariance matrix Σ_0.

The parameter β_0 used for the baseline correction of the LRR and β_1 for the local GC content are modeled by a normal distribution. If $\beta = (\beta_0, \beta_1)$, then:

$$\beta|\lambda_\beta \sim N(0, \lambda_\beta I_2),$$

where I_2 is the 2×2 identity matrix. Furthermore, the normal contamination percentage π_0 and percentage of cells having the normal genotype at SNP i, π_i, are both modeled by a Bèta distribution:

$$\pi_0|\alpha_{\pi_0}, \beta_{\pi_0} \sim \mathrm{Be}(\alpha_{\pi_0}, \beta_{\pi_0}),\ \pi_i|\alpha_{\pi_i}, \beta_{\pi_i} \sim \mathrm{Be}(\alpha_{\pi_i}, \beta_{\pi_i}),\ i = 1, \ldots, n.$$

The parameter η, used in the outlier class, is also modeled by a Bèta distribution.

In the model, it is assumed that $\bar{r}$ is known for the normal and tumor cells. According to [57], these values can be derived in practice.

The tumor states are assumed to form an inhomogeneous Markov Chain. Given the tumor state of x_{i-1}, say $x_{i-1} = k$, the probability that SNP i is in tumor state j is equal to:

$$\mathbb{P}(\mathrm{x_i} = j | \mathrm{x_{i-1}} = k) = \begin{cases} 1-\rho, & \text{if } j = k \\ \rho, & \text{if } j \neq k \end{cases}$$

where

$$\rho = \frac{1}{2}\left[1 - \exp\left(\frac{1}{2L}(s_i - s_{i-1})\right)\right]$$

and s_i is the physical position of the i-th SNP and L is a characteristic length (set to 2.000.000 in [57]).

The unknown model parameters are estimated by means of an Expectation Conditional Maximization (ECM) algorithm. OncoSNP uses multiple restarts to explore different baselines to correct for aneuploidy and the baseline with the greatest likelihood is reported and used for the calculation of the states of the tumors. Given the optimal parameters, the Viterbi algorithm [50] is used to determine the most likely sequence of tumor states. Next to that, OncoSNP also records the Maximum A Posteriori (MAP) estimates of the LRR baseline adjustment β_0 and normal cell contamination π_0.

OncoSNP is an algorithm which can only run on 64-bit Linux systems running `glibc` version 2.6 or higher. The minimal requirements for running OncoSNP is a computer with a dual-core processor and 4 GB of memory, while a multi-core processor and 32 GB of memory is recommended.

During the fitting process OncoSNP considers two initial baseline ploidy assumptions: 1) the data comes from a tumor which is near-diploid and 2) the data comes from a triploid tumor. In the first case, the LRR values are not shifted, whilst in the latter case OncoSNP assumes that the measured LRR values are smaller than in reality (assuming a ploidy above 2) and need to be shifted upwards. Note that OncoSNP implicitly assumes that gains are generally more common than losses. A clear reason on why this assumption is made has not been given by [57]. Figure 3.9 shows the two different settings.[2]

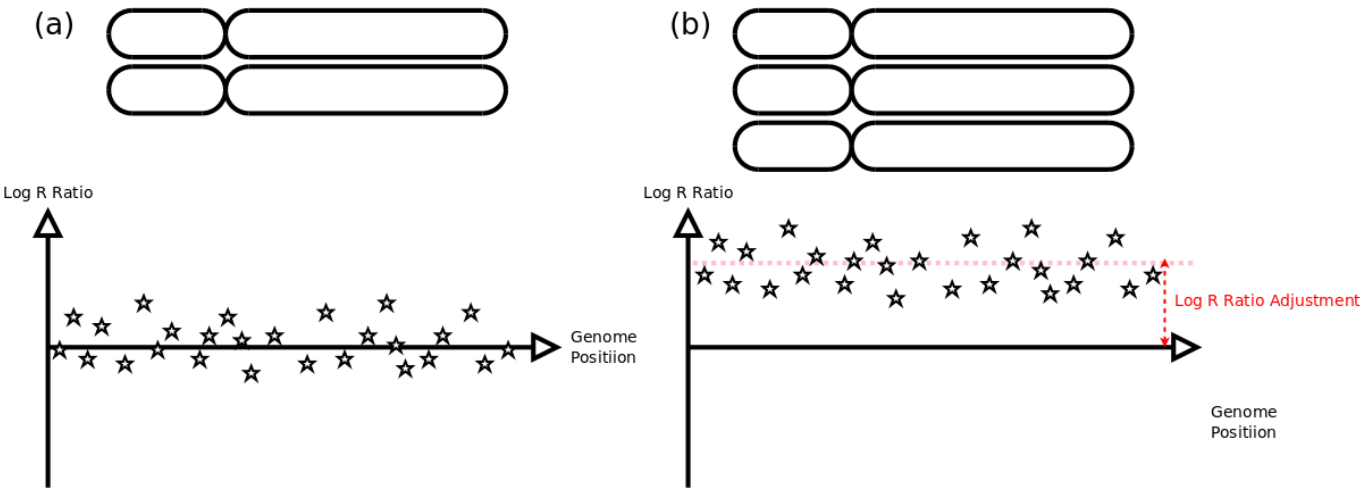


Figure 3.9: Depiction of the two different settings considered by OncoSNP.

In part a) of Figure 3.9, the LRR values of a diploid sample are shown and the LRR values are not shifted. On the other hand, part b) shows a triploid sample and the LRR values are shifted upwards.

[2]https://sites.google.com/site/oncosnp/user-guide/ploidy

During the training phase of the algorithm, OncoSNP will start from each of the two configurations separately and adaptively adjusts the estimates for the normal cell contamination and LRR baseline adjustment until the solution has converged. Therefore, two different solutions will arise. Given the two solutions, two different scenarios may occur at the end of the training phase: the solutions converge to the same solution or the solutions do not converge to the same solution. Figure 3.10 displays the two different scenarios that may occur.[3] In part a) of the plot, the two solutions converge to the same solution, i.e. there is a unique combination of normal cell contamination and LRR baseline adjustment that best explains the data regardless of the ploidy assumption. In other words, the algorithm converges to a unique solution after training the data. This implies that the SNP data encompasses sufficient information to deduce the ploidy and normal cell contamination. In part b) of the plot, the two solutions do not converge and the SNP data cannot uniquely determine the LRR adjustment value and normal cell contamination.

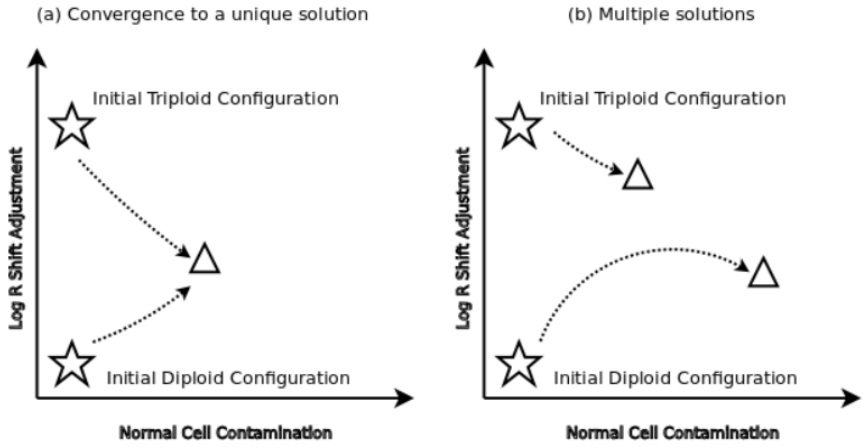


Figure 3.10: Two different scenarios which may occur during the training phase of OncoSNP.

For each setting, OncoSNP computes the likelihood of the model and marks the model with the highest likelihood as Ploidy No 1. The model with the lower likelihood is defined as Ploidy No 2. If the two settings converge to the same solution as in part a) of Figure 3.10, the copy number calls made by OncoSNP are similar. When the two settings lead to different solutions, the user can decide which model to use. A possible way to do this is to go for the model with the highest likelihood. However, [57] notes that the model with the highest likelihood is not the best model per se as there may be extra factors present in the data which are not considered by OncoSNP.

Next to the two different model configurations considered by OncoSNP, the algorithm also defines a rank for each copy number call. In other words, each copy number call which is outputted belongs to a specific rank. In OncoSNP, five different ranks are used in total: the higher the rank, the more detailed the genomic profile becomes. Rank 1 gives a coarse representation of what has happened on the genome, while rank 5 gives the most detailed representation. Figure 3.11 gives an example of how the rank influences the degree of detail of the genomic profile.[4]

In Figure 3.11 it can be clearly seen that the profile becomes more detailed as the rank increases. Which rank is most appropriate to use depends on the research question. Rank 1 would be sufficient to determine whether an entire chromosome has been deleted or not. On the other hand, if a gain or loss at a certain gene is of interest higher ranked calls are more appropriate. According to [57], the suggestion is to be prudent with incorporating higher ranked calls. In general, higher ranked calls include smaller alterations which have a higher classification error rate.

[3] https://sites.google.com/site/oncosnp/user-guide/ploidy
[4] https://sites.google.com/site/oncosnp/user-guide/oncosnpranking

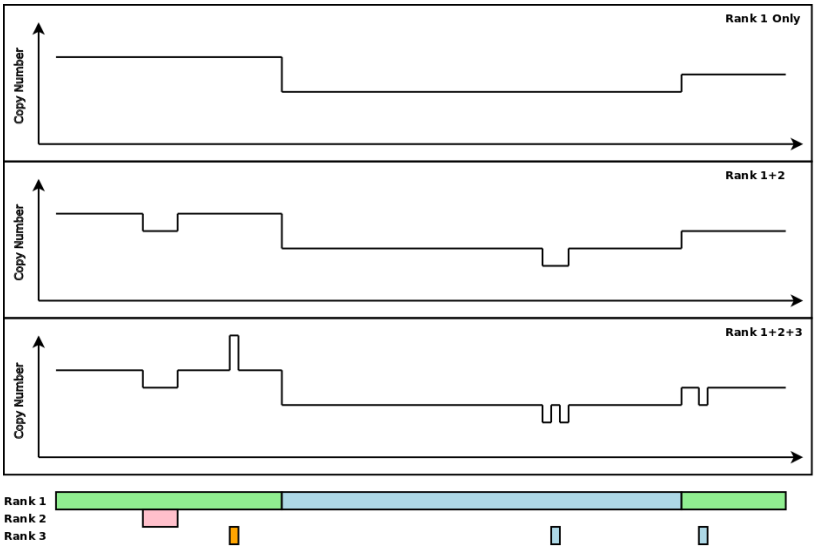


Figure 3.11: An example of how the rank influences the degree of detail of the segmentation profile.

4

Comparison methods

Both ASCAT and OncoSNP give a segmented profile as output: each SNP position belongs to a segment which is assigned a certain copy number. As stated in Section 2.3.3, a statistical method, also called a *comparison method*, will be applied on the segmented profiles to quantify the degree of clonality in a tumor pair. Many comparison methods exist that can be used to assess the clonality between two tumors. Some methods measure the similarity of two segmented profiles based on their overlapping breakpoints [6, 26], e.g. at which SNP positions do the profiles go from say a normal copy number to a gain copy number. A downside of breakpoint methods is that the similarity is based on the exact breakpoints. When noise is present in the raw data, the segmentation algorithm may not assign the breakpoints at the same SNP positions even though the aberrant regions detected by the segmentation algorithm are similar. As a result, the number of exact matches decreases leading in an underestimation of the degree of clonality.

Other comparison methods evaluate clonal relatedness based on the number of concordant events in the two tumors [31, 36, 52], i.e. how many SNPs are labeled as concordant by the segmentation algorithm. In contrast to breakpoint methods, concordant events do not underestimate the degree of clonality. However, they can be time consuming as the number of comparisons can be quite large.

Each comparison method outputs a score for each pair reflecting the degree of clonal relatedness for which a p-value can be computed using permutation methods.

Next to the methods explained above, hierarchical clustering is also a possible candidate that can be employed. Given a defined distance, which measures the closeness of two genomic profiles based on either the raw data or the segmentation profile, hierarchical clustering is applied on the set of all distances of the data. A pair is considered to be clonal if the two tumors of the same pair occur in the leaves of the resulting dendogram. According to [35], hierarchical clustering is not a suitable method to use when the clonality of two tumors is to be assessed. In hierarchical clustering, all the tumor pairs are taken into account in the algorithm. However, [35] states that clonality should primarily be based on the similarity of the tumors from one patient, instead of the relationships they have with the other tumor pairs. Moreover, the more tumors are present in the dataset, the less likely it will become that two clonal tumors will cluster together in a terminal branch. As the size of the dataset increases, the clonal pair has to beat more competitors in order to be classified as clonal. Therefore, the sensitivity of the method decreases as the sample size increases.

This chapter introduces the two comparison methods that will be used in this research. Note that the comparison methods presented in this chapter are methods that are based on the degree of concordance in the segmentation profiles as these give an overall picture of the degree of clonality on the entire genome. Section 4.1 explains how the segmentation output is preprocessed in order to optimize the sensitivity and computational efficiency of the comparison methods. Sections 4.2 and 4.3 describe the two different comparison methods that are going to be used in this thesis. In Section 4.4, alternative comparison methods that have been investigated are introduced and an elaboration on why these

methods are not applied for this purpose are given.

4.1. Preprocessing the segmentation output

When comparing the segmented profiles, a coarsening of the copy numbers is needed in order to improve the sensitivity of the comparison method. For instance, if the first tumor has copy number 3 in a certain region and the second tumor, which is a clone of the first, has copy number 5, then copy number 5 has arisen from copy number 3 by means of two extra duplications. If the profiles were compared based on the copy numbers alone (i.e. only when the copy numbers coincide, the aberrant events are called concordant), important information may be missed and the degree of clonality will be underestimated. Moreover, CN-LOH can not be detected when the copy numbers are used in the comparison methods as the copy number in a CN-LOH event is the same as in the normal setting. In clonal tumor pairs, a gain in the primary tumor will remain with a high probability a gain in the secondary tumor and the same applies for losses. The reason for this has to do with the fact that gains and losses are beneficial for a tumor: once a gain or loss has occurred, the copy number will likely remain a gain or loss and may even further increase or decrease. In other words, aberrant events are not easily reversed. Therefore, coarsening the copy numbers into states increases the sensitivity of the comparison method. Next to that, a coarsening of the output also enables CN-LOH events to be distinguished from normal events.

The copy numbers will be coarsened into the following four different states:

1. loss: copy number 1 or smaller.
2. normal: copy number 2.
3. CN-LOH: copy number 2, but only AA or BB genotypes.
4. gain: copy number 3 or higher.

As a normal cell is diploid, it makes sense to use copy number two as a baseline and refer to this as the normal state. A loss is defined when the copy number is equal to one or zero, while a gain is defined when the copy number is three or higher. CN-LOH is defined when the segmented profile outputs copy number two but only the A or B allele is present in the region, i.e. AB genotypes do not occur in the segment. For example, in the output of ASCAT a CN-LOH region is characterized by a segment having a maximal copy number of 2, i.e. $\max(\hat{n}_{A,s}^{\text{ASCAT}}, \hat{n}_{B,s}^{\text{ASCAT}}) = 2$, and a minimal copy number of 0, i.e. $\min(\hat{n}_{A,s}^{\text{ASCAT}}, \hat{n}_{B,s}^{\text{ASCAT}}) = 0$. Every SNP is assigned to one of the four different states. For instance, if a SNP lies in a segment having copy number one, it gets assigned to state 1. The output of the segmentation algorithm is changed from copy numbers, which range from 0 to possibly very large, to states which range from 1 to 4. This discretization approach is similar to [36], where in this case an extra state for CN-LOH events is included.

After the SNPs are coarsened into the four different states, each tumor pair can be compared SNP by SNP to see if the assigned states match. However, since the number of SNPs on a SNP array is large (in the order of 10^5), comparing the profiles SNP by SNP may not be very computationally efficient. Moreover, a segmentation algorithm may also make errors in the boundaries of the aberrant regions as noise can be present in the LRR and BAF data. The amount of noise is especially large when the DNA is extracted from so-called Formalin-Fixed Paraffin Embedded (FFPE) tumor tissues. FFPE is a preservation method which can be used to store tissue for long periods of time. Analyzing the DNA derived from FFPE samples can be quite challenging due to the fact that the DNA is often degraded and chemically modified during the formalin fixation [4]. In other words, the quality of the DNA decreases in FFPE samples introducing extra noise in the LRR and BAF plots. As the calls of the SNPs may be uncertain, comparing SNP by SNP can also underestimate the degree of clonality between two tumors.

In order to solve both problems, the SNPs can be binned together to give a general overview of what has happened in certain regions on the genome. This smooths out the possible noise and also makes the comparison method less computationally expensive. Note that the binning takes place per chromosome arm. There are two different ways that SNPs can be grouped together in bins:

- The SNPs can be binned together using a fixed bin size. For example, we can choose 500 as a fixed bin size. This implies that SNP 1 to 500 are put together in bin 1, SNPs 501 to 1000 in bin 2 etc.
- Another possibility is to use distance based bins. For this way of binning, the positions of the SNPs on the chromosome arm are taken into account and SNPs that are sufficiently close to one another are binned together. Once a distance is chosen, all SNPs that lie within that distance of the first SNP are binned together. For instance, a distance of 1 Megabases = 1,000,000 bases (abbreviated to 1 MB) can be chosen. One starts with SNP 1 and looks at which SNPs are located within 1 MB of SNP 1. These SNPs are then binned together in bin 1. The first SNP that will be in bin 2 will be the first SNP that is located more than 1 MB from SNP 1 in bin 1. The SNPs in bin 2 are those SNPs that are less than 1 MB away from the first SNP in bin 2 etc.

In general, using distance based bins is more realistic as these bins are located closely together and reflect the state of a certain part at the chromosome quite well. In this study a distance of 1 MB is applied for the distance based bins.

In the same way as the individual SNPs, each bin gets assigned to one of the four states introduced before. As all SNPs in a certain bin are assigned to one of the four states, the state of the bin is defined as the most common state occurring in the bin. For example, if there are 100 SNPs in a bin of which 80 are gains and 20 are normal, then the bin is labeled as a gain. The most common state is easy to determine if there is a true mode, but what if there are ties in the common state? Say that there are 50 gains and 50 normal SNPs in a bin: which state is then assigned to the bin? The following has been decided in case of a tie:

- Gain and normal are tied: assign the gain state to the bin.
- Gain and CH LOH are tied: assign the gain state to the bin.
- Normal and CN-LOH are tied: assign the CN-LOH state to the bin.
- Normal and loss are tied: assign the loss state to the bin.
- CN-LOH and loss are tied: assign the loss state to the bin.
- Gain and loss are tied: assign with probability 0.5 the loss state and with probability 0.5 the gain state to the bin.
- Gain, normal and CN-LOH are tied: assign the gain state to the bin.
- Gain, normal and loss are tied: assign with probability 0.5 the loss state and with probability 0.5 the gain state to the bin.
- Loss, normal and CN-LOH are tied: assign the loss state to the bin.
- Gain, CN-LOH and loss are tied: assign with probability 0.5 the loss state and with probability 0.5 the gain state to the bin.
- All four are tied: assign with probability 0.5 the loss state and with probability 0.5 the gain state to the bin.

In this case, loss and gain states are preferred over the normal and CN-LOH states in case of ties. Moreover, note that when a profile has a bin in which a gain and loss are tied, the final state assigned to the bin is randomized. As a consequence, if a profile has many gain and loss ties in the bins, running the profile several times using the procedure described above gives different outcomes. However, the samples used in this thesis did not show any gain or loss ties in the bins. Therefore, it can be concluded that the randomization process for gain and loss ties is allowed as it has almost no influence on the resulting binned profiles.

Once all the bins are assigned to one of the four different states, bins consisting of less than 10 SNPs are removed from the comparison analysis as these do not yield sufficient information about what has happened on the genomic region in question. The input given to the comparison methods are the states of the bins, where each bin consists of at least 10 SNPs. As the samples are segmented on the same SNPs, the number of bins on which each pair is evaluated in the comparison methods will be the same.

4.2. Log Likelihood Ratio

The Log Likelihood Ratio (Log LR) method is an adapted version of the Likelihood Ratio (LR) method developed by [36]. The LR method is introduced in Section 4.2.1. Section 4.2.2 explains how the Log LR method can be derived from the LR method.

4.2.1. The Likelihood Ratio

The LR method is implemented in the `clonality` package in R. Given the LRR values of the samples, the profiles are segmented using one step of the Circular Binary Segmentation (CBS) [33, 49] algorithm. The CBS algorithm iteratively searches for change points in sequential data until no significant differences can be detected. By using one step of the CBS algorithm, a maximum of one change is assigned to each chromosome arm. In other words, only one gain or loss per chromosome arm is allowed. If no aberrant event is found on the chromosome arm, the arm is considered to be normal. After the segmentation of the profiles is complete, i.e. each chromosome arm is assigned to a gain, normal or loss state, a likelihood ratio approach will be used where the following two hypotheses will be tested against one another:

H_0 : the two tumors arose independently.

H_1 : the two tumors are clonally related, i.e. one is a metastasis of the other.

The Likelihood Ratio in [36] is defined as:

$$LR = \frac{L(c > 0)}{L(c = 0)},$$

where c denotes the percentage of aberrant events that originate from the clonal cell if two tumors are clonal. In the ratio, $L(c = 0)$, corresponds to the null hypothesis assuming independence and $L(c > 0)$ to the metastasis hypothesis. A LR above 1 would imply that the observed genomic aberrations are more likely to occur under the hypothesis of clonality, while a LR below 1 indicates that the observed data is more likely to occur under the hypothesis of independence.

The likelihood $L(c)$ is defined as follows:

$$\begin{aligned} L(c) \quad = \quad & \prod_i \left[cp_{1i} + \frac{(1-c)^2 p_{1i}^2}{1 - cp_{1i} - cp_{2i}}\right]^{r_{11i}} \left[cp_{2i} + \frac{(1-c)^2 p_{2i}^2}{1 - cp_{1i} - cp_{2i}}\right]^{r_{22i}} \left[\frac{2(1-c)^2 p_{1i} p_{2i}}{1 - cp_{1i} - cp_{2i}}\right]^{r_{12i}} \\ & \left[\frac{2(1-c)^2 p_{1i} p_{3i}}{1 - cp_{1i} - cp_{2i}}\right]^{r_{13i}} \left[\frac{2(1-c)^2 p_{2i} p_{3i}}{1 - cp_{1i} - cp_{2i}}\right]^{r_{23i}} \left[\frac{p_{3i}^2}{1 - cp_{1i} - cp_{2i}}\right]^{r_{33i}} \\ & \cdot \prod_{i \in \Psi_g} (b_{1i} f_{M_i}(t_i) + (1 - b_{1i}) f_{I_i}(t_i)) \prod_{i \in \Psi_l} (b_{2i} f_{M_i}(t_i) + (1 - b_{2i}) f_{I_i}(t_i)), \end{aligned}$$

The first part of the likelihood is a multinomial component which characterizes the correlation of the gains and losses that occur on the different chromosome arms. The following indicators are used in the first part of the likelihood:

- $r_{11i} = 1$ if the i-th chromosome arm is a gain for both tumors.
- $r_{22i} = 1$ if the i-th chromosome arm is a loss for both tumors.
- $r_{33i} = 1$ if the i-th chromosome arm is normal for both tumors.
- $r_{12i} = 1$ if the i-th chromosome arm is a gain for one tumor and a loss for the other tumor.
- $r_{13i} = 1$ if the i-th chromosome arm is a gain for one tumor and normal for the other tumor.
- $r_{23i} = 1$ if the i-th chromosome arm is a loss for one tumor and normal for the other tumor.

Let $r_i = [r_{11i}, r_{22i}, r_{33i}, r_{12i}, r_{13i}, r_{23i}]$. The quantity r_i summarizes what has happened on the i-th chromosome arm for the two tumors, i.e. only one indicator in r_i will be equal to 1. Next to that, p_{1i} is

the probability that a gain occurs on the i-th chromosome arm, p_{2i} the probability of a loss on the i-th chromosome arm and p_{3i} the probability that the i-th chromosome arm is assigned as normal. The probabilities p_{ki}, for $k = 1, 2, 3$, in the likelihood $L(c)$ are computed for each pair separately. First, the empirical frequencies of a gain, loss and normal event are determined from the entire cohort. Given these empirical frequencies, the marginal probabilities are rescaled to reflect the overall frequencies of gains and losses that occur in the tumor pair. Hence, each pair in the dataset gets assigned different probabilities for a gain, loss and normal event in the likelihood computations. More information on how the probabilities are computed can be found in the appendix of [36].

The probability that a concordant gain occurs on the i-th chromosome arm (i.e. $r_{11i} = 1$) is by the law of total probability equal to the probability of a clonal gain on the i-th chromosome arm added to the probability that the gains arose independently on the i-th chromosome arm.

The probability of a clonal gain on the i-th chromosome arm is equal to the clonality proportion c times the probability that a gain occurs on the i-th chromosome arm, i.e. cp_{1i}.

The probability that the two gains arose independently on the i-th chromosome arm is equal to:

$$\begin{aligned}\mathbb{P}(\text{gains arose indep. on chr. arm } i) &= \mathbb{P}(\text{gains arose indep. on chr. arm } i|\text{clonal event})\mathbb{P}(\text{clonal event}) \\ &+ \mathbb{P}(\text{gains arose indep. on chr. arm } i|\text{no clonal event})\mathbb{P}(\text{no clonal event}).\end{aligned}$$

As the first term is equal to zero (two gains can not arise independently given a clonal event), the probability simplifies to:

$$\mathbb{P}(\text{gains arose indep. on chr. arm } i) = \mathbb{P}(\text{gains arose indep. on chr. arm } i|\text{no clonal event})\mathbb{P}(\text{no clonal event}).$$

The probability of no clonal event is equal to 1 minus the probability of a clonal event. By the law of total probability, this is equal to:

$$\begin{aligned}\mathbb{P}(\text{no clonal event on chr. arm } i) &= 1 - \mathbb{P}(\text{clonal event on chr. arm } i) \\ &= 1 - \mathbb{P}(\text{clonal gain on chr. arm } i) - \mathbb{P}(\text{clonal loss on chr. arm } i) \\ &= 1 - cp_{1i} - cp_{2i}.\end{aligned}$$

Given this result, the probability of two gains arising independently given no clonal event becomes equal to:

$$\begin{aligned}\mathbb{P}(\text{gains arose indep. on chr. arm } i|\text{no clonal event}) &= \mathbb{P}(\text{indep. gain on chr. arm } i|\text{no clonal event})^2 \\ &= \left(\frac{\mathbb{P}(\text{indep. gain on chr. arm } i \cap \text{no clonal event})}{\mathbb{P}(\text{no clonal event})}\right)^2 \\ &= \left(\frac{(1-c)p_{1i}}{1 - cp_{1i} - cp_{2i}}\right)^2.\end{aligned}$$

Putting all the elements together, it follows that the probability of a concordant gain at chromosome arm i equals:

$$\begin{aligned}\mathbb{P}(\text{concordant gain on chromosome arm } i) &= cp_{1i} + \left(\frac{(1-c)p_{1i}}{1 - cp_{1i} - cp_{2i}}\right)^2 (1 - c_{1i} - cp_{2i}) \\ &= cp_{1i} + \frac{(1-c)^2 p_{1i}^2}{1 - cp_{1i} - cp_{2i}}.\end{aligned}$$

The probability for a concordant loss event is derived similarly. The derivation of the other probabilities in the first part of the likelihood can be found in Appendix A.

The second part of the likelihood only focuses on the chromosome arms where concordant gains and losses have occurred. For this, ψ_g is defined as all the arms on which concordant gains have occurred and ψ_l as all the arms having concordant losses. The quantity t_i is a closeness statistic which reflects

the discrepancy between the start and end points of the concordant gains and losses. Figure 4.1 shows an example of a concordant loss on chromosome arm 10q with different endpoints.

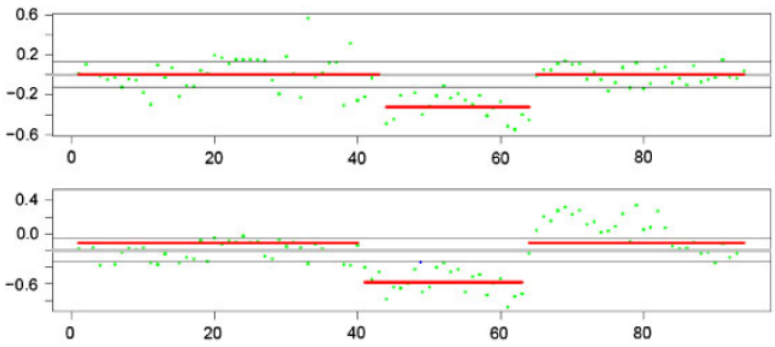


Figure 4.1: A concordant loss on chromosome arm 10q with different endpoints. Figure retrieved from [36].

Let l_k^i and m_k^i be the start and end points of the copy number change on the i-th chromosome arm of the k-th tumor in the pair. The closeness statistic t_i is defined as

$$t_i = |l_1^i - l_2^i| + |m_1^i - m_2^i|.$$

The closeness statistic is evaluated in f_{M_i} and f_{I_i}, which are the probability density functions of the closeness statistic t_i when the concordant events have occurred in the clonal and independent situation respectively. Empirical estimates of the probability density functions are generated using a simulation procedure. See the appendix of [36] for a detailed description on how this is done. Next to that, b_{1i} is the conditional probability that an individual gain on the i-th chromosome arm is a clonal gain. Similarly, b_{2i} is the conditional probability that an individual loss is a clonal loss.

Once the LR score has been computed for each pair, a reference distribution assuming independence is constructed by means of permutation. Permutation methods were originally invented by Fisher [15] and Pitman [40]. The idea behind permutation methods is that since the labels of the observations are exchangeable under the null hypothesis, the distribution of the test statistic under the null hypothesis can be derived by means of looking into every possible permutation of the labels. In other words, the significance of a test statistic T, which can be used to test the null hypothesis against the alternative hypothesis, can be determined by computing the test statistics T^* for each possible permutation of the data. Given the test statistics T^* of the permutation distribution, a p-value for the test statistic T can be derived by looking at the proportion of observations T^* which are smaller or larger (depending on whether smaller or larger values of the test statistic T point into the direction of the alternative hypothesis) than the test statistic T. For example, let's assume that a small value of the test statistic T gives evidence for the alternative hypothesis. Let the value of the test statistic be equal to t, i.e. $T = t$, and define t_i^*, $i = 1, \dots, N$, as the values of the permutation test statistics of all the possible N permutations that can be made of the data. The p-value for the test statistic T is defined as:

$$p = \frac{1}{N} \sum_{i=1}^{N} 1_{\{t_i^* \leq t\}},$$

where 1 is the indicator function which is equal to 1 only if t_i^* is smaller than t. Note that when a large value of the test statistic T points towards the alternative hypothesis, the p-value is defined as the proportion of permutation test statistics which are larger than the test statistic. Given a defined significance level α, the null hypothesis is rejected, i.e. the p-value is found to be significant, if $p \leq \alpha$. In practice, the significance level α is mostly set to 0.05 or 0.01. In other words, the null hypothesis is rejected if less than 5 or 1 percent of the permutation test statistics is smaller (or larger, depending on the test statistic) than the test statistic T.

A permutation test is a type of non-parametric test which makes no assumption about the theoretical underlying distribution of the test statistic under the null hypothesis. The underlying distribution under the null hypothesis is constructed by means of computing the test statistics for all possible permutations of the data. However, a major drawback of permutation tests are that they can become computationally quite intensive when the number of observations increases. This has to do with the fact that the number of possible permutations increases when the data size increases. As a result, more test statistics need to be computed in order to construct the distribution under the null hypothesis which increases the computation time. When the dataset is large, a different technique, such as bootstrapping, can be used instead of permutation to reduce the computation time. The datasets used in this study are relatively small so that this problem is not encountered.

In order to test the significance of the LR scores, the distribution under the null hypothesis will be constructed by means of computing the LR scores for independent pairs. For this, each primary tumor of a patient is paired with a secondary tumor of another patient, which are by definition independent, and the LR score of this artificial pair is computed. If there are n pairs, a total of $n(n-1)$ permutations between different patients can be made. The reference distribution will then be used to assess the significance of the LR scores of the true pairs. For a defined significance level α, the null hypothesis assuming that the two tumors arose independently is rejected if the p-value of the LR score of a pair falls below the significance level α, where the p-value is defined as the proportion of observations in the independence distribution that are greater than or equal to the observed LR score. In other words, the pair is considered as clonal by the LR method if the LR score of the pair falls above the $(1-\alpha)$-th percentile of the reference distribution.

Note that the resulting p-values depend on the sample size that is being used. Given n pairs, the minimal p-value which is larger than zero is equal to $\frac{1}{n(n-1)}$. If the dataset is too small, the minimal non-zero p-value will be larger than the defined significance level α so that the null hypothesis will never be rejected. Therefore, the number of pairs needs to be sufficiently large in order for the permutation method to reject the null hypothesis. The most commonly chosen significance levels are $\alpha = 0.05$ and $\alpha = 0.01$. When 20 pairs are involved, the minimal non-zero p-value is sufficiently small so that clonality can be detected at both significance levels. Next to that, the resulting p-values are also influenced by the similarity of the profiles of two independent tumors. When many independent pairs have similar characteristics, the LR scores for these pairs will be quite large. As a result, the p-values of the pairs may increase as more observations in the reference distribution are larger than the LR score of the pairs.

Figure 4.2 shows an example of a reference distribution consisting of all the LR scores of the permuted independent pairs (black histogram) and distribution of the LR scores of the tumor pairs (red histogram).

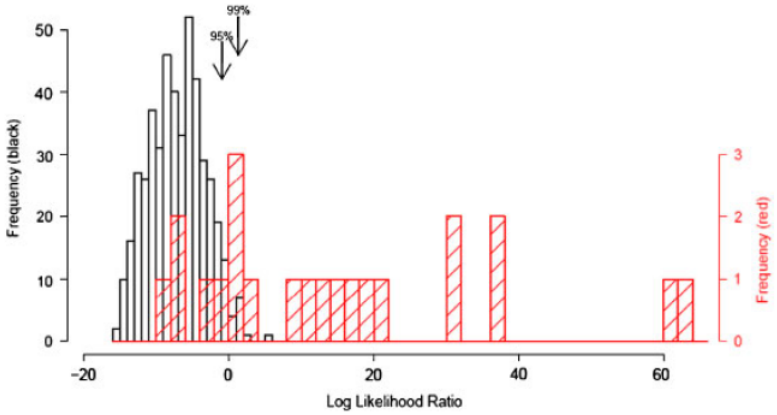


Figure 4.2: LR scores of the independent pairs (black histogram) and the LR scores of the true pairs (red histogram). Figure retrieved from [36].

In Figure 4.2, the 95% and 99% quantiles of the independence distribution are marked by the two arrows. All the pairs whose LR score lies above these two arrows are significant at a 5 and 1% significance level respectively.

The LR method is an intuitive method which can be used to detect clonality in tumor pairs. However, the method has two shortcomings:

- The LR method uses one step of the CBS algorithm to segment the profiles. As a consequence, only one gain or loss per chromosome arm is allowed which is very restrictive. In reality, gains and losses can occur simultaneously on the same chromosome arm.
- The LR method only takes into account the LRR values as input so that CN-LOH can not be detected.

4.2.2. Deriving the Log LR method

As mentioned in Section 4.1, the input that is used for the comparison method are the states of the bins, where each bin consists of at least 10 SNPs. Once the states of the bins are known, the LR method can be modified so that the two genomic profiles of the tumors can be compared bin by bin. Note that, as the profiles are compared bin-wise, different aberrant events can occur simultaneously on the same chromosome arm. Since more than one aberrant event can occur on a chromosome arm, the second part of the likelihood in the LR method, measuring the closeness of the aberrant events on the concordant arms, is removed. The first part of the likelihood is maintained and adapted in order to include the extra CN-LOH state that can occur in a bin. It should be noted that the extended method has not been used in literature before.

In order to extend the existing LR method, the following probabilities are defined:

$$\begin{aligned} p_{1i} &= \mathbb{P}(\text{loss occurs in the } i\text{-th bin}), \\ p_{2i} &= \mathbb{P}(i\text{-th bin is normal}), \\ p_{3i} &= \mathbb{P}(\text{CN-LOH occurs in the } i\text{-th bin}), \\ p_{4i} &= \mathbb{P}(\text{gain occurs in the } i\text{-th bin}), \end{aligned}$$

where $p_{1i} + p_{2i} + p_{3i} + p_{4i} = 1$. Note that the same encoding as introduced in Section 4.1 is used (i.e. a loss is encoded as a 1, a normal as a 2 etc.). In contrast to the original LR method, the probabilities for one of the four events are the same for each pair. When the approach of [36] was used for deriving the probabilities, problems arose in the computation of the Log LR scores. Therefore, we decided that the empirical frequencies in the cohort will be used for the probabilities of a loss, normal, CN-LOH and gain event.

Define r_{11i} as the indicator which equals 1 if both tumors have a loss in the i-th bin, r_{22i} is equal to 1 if both profiles are normal in the i-th bin etc. In other words:

$$r_i = (r_{11i}, r_{22i}, r_{33i}, r_{44i}, r_{12i}, r_{13i}, r_{14i}, r_{23i}, r_{24i}, r_{34i})$$

denotes all the possible combinations that can occur at the i-th bin and only 1 entry in r_i is equal to 1. Let c again be the proportion of mutations that originate from the same cell in case two tumors are clonal. For bin i, the likelihood can be defined as follows:

$$\begin{aligned} L(c)^i = & \left(cp_{1i} + \frac{(1-c)^2 p_{1i}^2}{1 - cp_{1i} - cp_{3i} - cp_{4i}}\right)^{r_{11i}} \left(\frac{p_{1i}^2}{1 - cp_{1i} - cp_{3i} - cp_{4i}}\right)^{r_{22i}} \\ & \left(cp_{3i} + \frac{(1-c)^2 p_{3i}^2}{1 - cp_{1i} - cp_{3i} - cp_{4i}}\right)^{r_{33i}} \left(cp_{4i} + \frac{(1-c)^2 p_{4i}^2}{1 - cp_{1i} - cp_{3i} - cp_{4i}}\right)^{r_{44i}} \\ & \left(\frac{2(1-c)p_{1i}p_{2i}}{1 - cp_{1i} - cp_{3i} - cp_{4i}}\right)^{r_{12i}} \left(\frac{2(1-c)^2 p_{1i}p_{3i}}{1 - cp_{1i} - cp_{3i} - cp_{4i}}\right)^{r_{13i}} \\ & \left(\frac{2(1-c)^2 p_{1i}p_{4i}}{1 - cp_{1i} - cp_{3i} - cp_{4i}}\right)^{r_{14i}} \left(\frac{2(1-c)p_{2i}p_{3i}}{1 - cp_{1i} - cp_{3i} - cp_{4i}}\right)^{r_{23i}} \end{aligned}$$

$$\left(\frac{2(1-c)p_{2i}p_{4i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{24i}}\left(\frac{2(1-c)^2p_{3i}p_{4i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{34i}}.$$

The probabilities of the different events, e.g. a concordant loss, one bin is normal and the other bin a gain etc., are derived similarly as the probabilities in the original LR method.

If there are N bins in total, the likelihood equals:

$$\begin{aligned}L(c) \quad = \quad & \prod_{i=1}^{N}\left(cp_{1i}+\frac{(1-c)^2p_{1i}^2}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{11i}}\left(\frac{p_{2i}^2}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{22i}}\\ & \left(cp_{3i}+\frac{(1-c)^2p_{3i}^2}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{33i}}\left(cp_{4i}+\frac{(1-c)^2p_{4i}^2}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{44i}}\\ & \left(\frac{2(1-c)p_{1i}p_{2i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{12i}}\left(\frac{2(1-c)^2p_{1i}p_{3i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{13i}}\\ & \left(\frac{2(1-c)^2p_{1i}p_{4i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{14i}}\left(\frac{2(1-c)p_{2i}p_{3i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{23i}}\\ & \left(\frac{2(1-c)p_{2i}p_{4i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{24i}}\left(\frac{2(1-c)^2p_{3i}p_{4i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)^{r_{34i}}.\end{aligned}$$

In the same way as the original Likelihood Ratio, the hypothesis of independence is tested against the hypothesis of clonality. The probabilities in the likelihood depend on the constant c, which reflects the probability that a concordant event is of clonal origin, so that the adapted Likelihood Ratio is again equal to:

$$LR=\frac{L(c>0)}{L(c=0)}.$$

However, note that a problem arises if the Likelihood Ratio is defined as above. In contrast to the original Likelihood Ratio, a comparison is not made per chromosome arm, but per bin. This entails that more comparisons are made and since the probabilities are all below 1, an underflow in the likelihood can occur if the number of comparisons becomes large. For instance, 0.1^{40} is already in the order of 10^{-40} which will give an underflow on most computers. Therefore, looking at the likelihood is not a good idea in this case. In order to prevent an underflow, the logarithm of the likelihood can be used instead. The log likelihood is defined as:

$$\begin{aligned}\log(L(c)) \quad = \quad & \sum_{i=1}^{N} r_{11i}\log\left(cp_{1i}+\frac{(1-c)^2p_{1i}^2}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)+r_{22i}\log\left(\frac{p_{1i}^2}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)\\ & +r_{33i}\log\left(cp_{3i}+\frac{(1-c)^2p_{3i}^2}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)+r_{44i}\log\left(cp_{4i}+\frac{(1-c)^2p_{4i}^2}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)\\ & +r_{12i}\log\left(\frac{2(1-c)p_{1i}p_{2i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)+r_{13i}\log\left(\frac{2(1-c)^2p_{1i}p_{3i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)\\ & +r_{14i}\log\left(\frac{2(1-c)^2p_{1i}p_{4i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)+r_{23i}\log\left(\frac{2(1-c)p_{2i}p_{3i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)\\ & +r_{24i}\log\left(\frac{2(1-c)p_{2i}p_{4i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right)+r_{34i}\log\left(\frac{2(1-c)^2p_{3i}p_{4i}}{1-cp_{1i}-cp_{3i}-cp_{4i}}\right).\end{aligned}$$

and the Log Likelihood Ratio becomes equal to:

$$\log LR=\frac{\log(L(c>0))}{\log(L(c=0))}.$$

By taking the log likelihood, all terms in the sum become negative since the probabilities lie between 0 and 1. If the observed events are more likely to occur under the clonality hypothesis, the log likelihood

under the clonality hypothesis becomes less negative than the log likelihood under the independence hypothesis. This is because of the fact that the closer the probability is to 1, i.e. the more likely it is that an event occurs, the less negative the logarithm of the probability becomes. For two probabilities p and q, with $0 < p < q < 1$, the following inequality holds:

$$\log(p) < \log(q).$$

Hence, if the data is more likely to occur under the clonality hypothesis, the Log LR will be below 1. The log likelihood does not give an underflow and thus solves the problem previously mentioned.

Given the Log Likelihood Ratio score for a pair, the significance of the score can be computed by means of constructing an independence distribution in the same way as the Likelihood Ratio. All the scores together form a reference distribution which can be used to assess the significance of the score. The null hypothesis assuming independence is rejected if the p-value of the Log LR score falls below the significance level α, where the p-value of the Log LR score is defined as the proportion of observations in the independence distribution which are smaller than or equal to the Log LR score. In other words, the pair is considered as clonal if the Log LR score of the pair falls below the α-th percentile of the reference distribution assuming independence. Note that the Log LR score needs to fall below the α-th percentile as a lower score implies that the clonality hypothesis is being favored.

[36] has shown that the outcome of the LR is insensitive when the clonality proportion c is within the interval $[0.25, 0.75]$. For simplicity, c is chosen equal to 0.5 in the rest of this report.

In order to show how the Log LR scores vary with different degrees of clonality, a simple example consisting of three pairs has been constructed. Each pair consists of 20 bins in total, where the similarity between the primary and secondary tumor varies per pair. Pair 1 is very similar, having 19 concordant bins (e.g. both bins are a gain, both bins are a loss etc.) and only 1 discordant bin. Pair 2 is somewhat similar, the states of 10 bins are concordant while the other 10 bins have a discordant state. Pair 3, on the other hand, is far from being similar having only 1 concordant bin and 19 discordant bins. Figure 4.3 shows the states of the bins of the three pairs by means of barplots. Note that the height of the bar in the barplot reflects the state of the bin, i.e. a bin of height 1 displays a loss, a bin of height 2 a normal state etc. Moreover, the color of the bin reflects whether the bin is concordant (green) or discordant (red) with the bin of the other tumor in the pair.

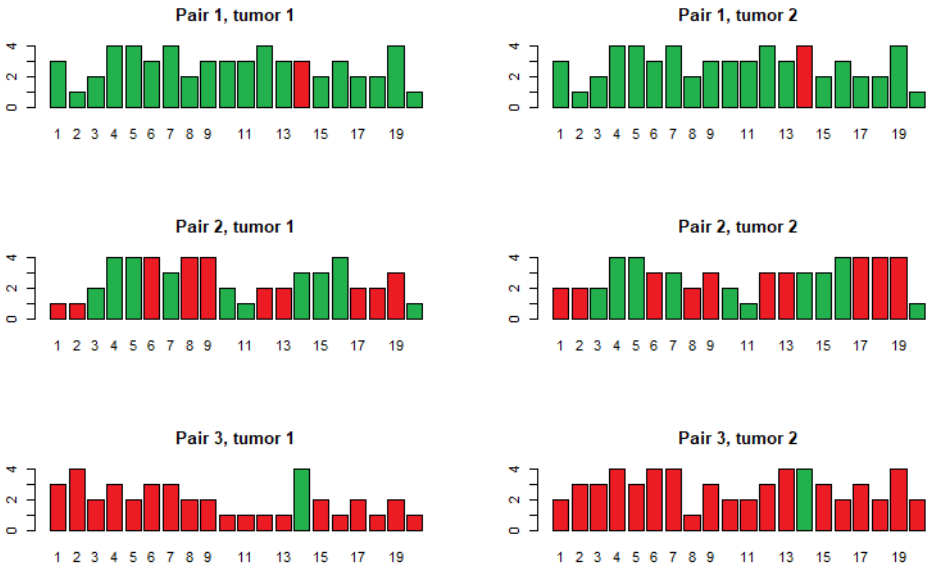


Figure 4.3: Barplots of the states of the bins for three pairs having a varying degree of clonality. The color of each bin reflects whether the bin is concordant (green) or discordant (red) with the bin of the other tumor in the pair.

The Log LR scores for the three pairs are as follows: pair 1 has a score of 0.781, pair 2 has a score of 1.069 and pair 3 has a score of 1.251. From the Log LR scores it can be seen that the larger the number of concordant bins, the smaller the Log LR score becomes. In other words, a larger number of concordant bins implies that the pair is more likely to be of clonal origin. When the pair has 50% of matching bins, the Log LR score is close to 1: the data is equally likely to occur under both hypotheses. Note that for this example, the score is not exactly equal to 1, but slightly above 1. This has to do with the fact that the probabilities of the different events influence the Log LR score as well.

4.3. Adapted Similarity Index

[31] developed the Similarity Index (SI) as a comparison method to identify whether two tumors are of clonal origin. The SI measures the degree of concordance in the aberrant events that are present in the genomic profiles of the two tumors. Note that the SI was originally invented to compare the genomic profiles SNP by SNP, but it can also be applied in our setting. Given the N bins for which the states are determined, define N_S as the number of bins for which the aberrant states coincide, i.e. both bins are losses or gains, N_O as the number of bins for which the outcome is opposite, i.e. one bin is a gain while the other bin is a loss, and N_U as the number of bins for which an unique gain or loss occurs, i.e. one bin is a gain or loss, the other bin is normal. Figure 4.4 taken from [31] shows all possible combinations that may occur in the primary and secondary tumor.

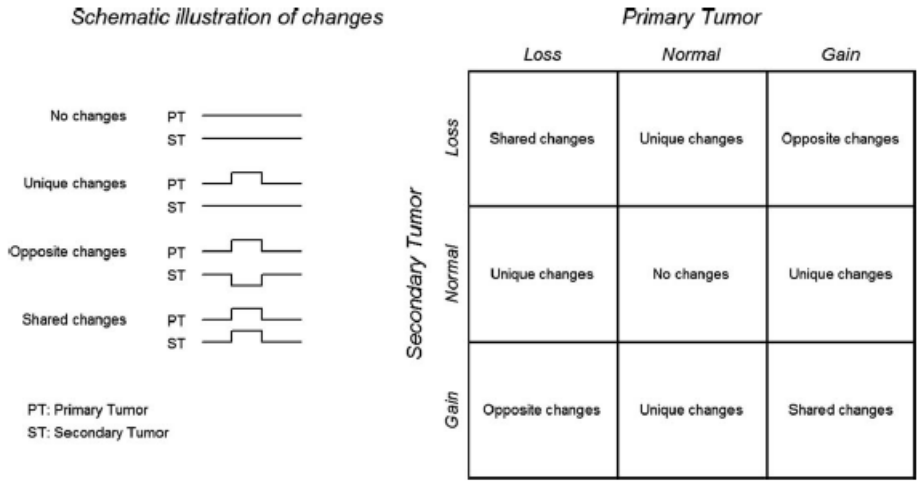


Figure 4.4: All possible combinations that may occur in the primary and secondary tumor. Figure retrieved from [31].

The SI is defined as:

$$SI = \frac{N_S}{N_S + N_U + N_O}.$$

In other words, the SI measures the fraction of aberrant events that is shared between two tumors. Note that the bins which are normal for both tumors are not included in the measure as these are not of interest for determining the degree of clonality. The SI is equal to 1 in case all the aberrant events on the two genomic profiles are shared and 0 if two profiles do not share any aberrant events. The SI tests the same two hypotheses as the Log LR method:

H_0 : the two tumors arose independently.

H_1 : the two tumors are clonally related, i.e. one is a metastasis of the other.

The significance of the SI score can be determined by means of constructing a reference distribution assuming independence as described in Section 4.2. Given the significance level α, the null hypothesis assuming independence is rejected if the p-value of the SI score falls below the significance level, where the p-value is defined as the proportion of observations in the independence distribution which are greater than or equal to the SI score. In other words, the pair is considered as clonal if the SI score falls above the $(1-\alpha)$-th percentile of the reference distribution as a larger SI score implies that the pair is more likely to be clonal.

4.3.1. Extending the Similarity Index

In [31] a distinction is made only between gains, losses and normal states. In this research, however, a fourth state, CN-LOH, is also included in the analysis. The SI can be easily extended so that all four states are incorporated in the measure. For this, define:

- N_S as the number of shared aberrant bins, i.e. both bins are losses, gains or CN-LOH.
- N_O as the number of opposite aberrant bins, i.e. gain-CN-LOH, gain-loss or loss-CN-LOH.
- N_U as the number of unique aberrant bins, i.e. one bin is in one of the three aberrant states and the other bin is normal.

Given these three quantities, the SI score of a pair and the significance of the latter can be computed in the same way as described above.

4.3.2. The adapted Similarity Index

The Similarity Index measures the degree of concordance in the aberrant events that occur in the two genomic profiles of the tumors. The quantities N_O and N_U are the number of opposite and unique events that occur in the bins respectively. However, the quantity N_U does not make any distinction between whether the bin in the first tumor is normal and the bin in the second tumor is aberrant or vice versa. These two events are weighted similarly in the SI. It may be questioned whether this is realistic: secondary tumors often have, in case of a metastasis, more aberrant events than the primary tumor as the tumor cells keep on mutating. Therefore, a lower weight for this event should be applied. On the other hand, a bin going from aberrant in the primary to normal in the secondary tumor may be a sign that the two tumors are not of clonal origin as the mutation is lost. Next to that, no distinction is being made in the number of opposite events as well: a gain-loss is weighted the same in the SI as a gain-CN-LOH or loss-CN-LOH. As a gain-loss is less likely to occur in case of a clonal pair than a gain-CN-LOH or loss-CN-LOH event, a gain-loss should be penalized heavier than the other two events.

An adapted Similarity Index can be constructed by means of making a distinction between the several sub events that may occur within each event. For this, define:

- N_S as the number of shared aberrant bins.
- N_U^1 as the number of bins which are aberrant in the primary tumor and normal in the secondary tumor.
- N_U^2 as the number of bins which are normal in the primary tumor and aberrant in the secondary tumor.
- N_O^1 as the number of bins which are gain-CN-LOH or loss-CN-LOH.
- N_O^2 as the number of bins which are a gain-loss.

The adapted SI can be defined as:

$$SI^{\text{adapted}} = \frac{N_S}{N_S + c_1 N_U^1 + c_2 N_U^2 + d_1 N_O^1 + d_2 N_O^2},$$

for $c_1, c_2, d_1, d_2 \geq 0$. The constants in the adapted SI can be chosen freely.

In this thesis, the following constants are used in the adapted SI: $c_1 = 1$, $c_2 = 0.5$, $d_1 = 1$ and $d_2 = 2$. Given this choice of constants, the adapted SI becomes equal to:

$$SI^{\text{adapted}} = \frac{N_S}{N_S + N_U^1 + 0.5 N_U^2 + N_O^1 + 2 N_O^2}.$$

By choosing the constants this way, the adapted SI puts a larger penalization on bins which are aberrant in the primary and normal in the secondary tumor compared to bins which are normal in the primary and aberrant in the secondary tumor as the former implies that an aberrant region is lost, which rarely occurs in case of clonality. Next to that, bins that are a gain-loss combination get a twice as heavier

weight in the ratio than bins that are gain-CN-LOH or loss-CN-LOH since a gain may become CN-LOH but it is uncommon for a gain to become a loss or vice versa when two tumors are of clonal origin.

Similar to the original SI, the significance of the scores can be determined by means of constructing a reference distribution assuming independence. Given the significance level α, the adapted SI score is considered to be significant if the p-value of the adapted SI score falls below the significance level, where the p-value is defined as the proportion of observations in the independence distribution which are greater than or equal to the adapted SI score.

In order to investigate how the adapted SI score changes for different degrees of clonality, the same three pairs, plotted in Figure 4.3, which were used to study the behavior of the Log LR were employed. As explained in Section 4.2.2, pair 1 is very similar, pair 2 is somewhat similar and pair 3 is not similar at all. The adapted SI scores for the three pairs are as follows: pair 1 has a score of 0.933, pair 2 has a score of 0.5 and pair 3 has a score of 0.057. As expected, a larger number of concordant bins implies a larger adapted SI score. Pair 2, whose bins match in 50% of the cases, has an adapted SI score which is exactly equal to 0.5. However, note that an adapted SI score of 0.5 does not always occur when the bin match percentage is equal to 50%. In the computation of the adapted SI, the concordant normal bins are excluded from the analysis and some discordant events are penalized more than others.

4.4. Alternative comparison methods

The comparison methods introduced in the previous two sections are both quite complicated methods which are extended and adapted from already existing comparison methods. As the comparison methods are fairly sophisticated, it is of interest to see if a simpler comparison method, which measures the similarity of two segmentation profiles solely on the states of the bins, would perform equally well. In other words, the comparison of the segmentation profiles can be seen as a mathematical problem that is application independent: is there a mathematical method that can be applied on this data so that the more complex comparison methods may be outperformed? In the next three subsections, different candidate methods are introduced and the applicability of the methods investigated.

4.4.1. Histogram difference

The first method which was attempted is the histogram difference method. For this method, the states of the bins are converted into histograms. As explained in Section 4.1, the segmentation output for a pair of tumors consists of two vectors of length N (number of bins) existing of the integers 1, 2, 3 and 4, where each number corresponds to one of the four discretized states, i.e.:

$$s^{(k)} = \left(s_1^{(k)}, s_2^{(k)}, \dots, s_N^{(k)}\right),\ k = 1, 2,$$

where $s_i^{(k)}$ is the state at bin i for tumor k, $s_i^{(k)} \in \{1, 2, 3, 4\}$.

Given the vectors $s^{(k)}$ for $k = 1, 2$, a histogram can be constructed for each tumor consisting of N bars where each bar corresponds to a bin and the height of each bar reflects the state of the bin. Figure 4.5 shows an example of two sequences consisting of six bins.

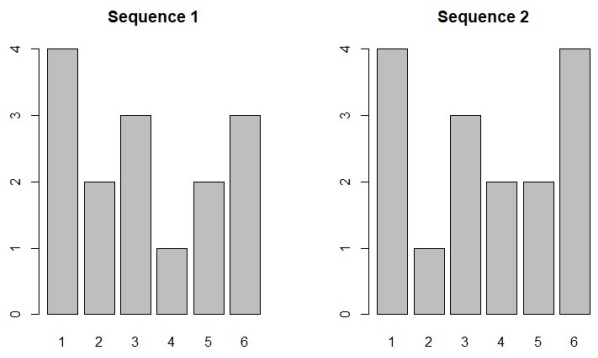


Figure 4.5: Two histograms reflecting the states (1,2,3,4) at $N = 6$ bins.

The histogram difference method computes the absolute difference between the two histograms. Let A and B be two histograms each consisting of N bins that display the states of tumor A and tumor B respectively. Define A_i as the height of the i-th bar in histogram A (reflecting the state of the i-th bin in tumor A) and B_i as the height of the i-th bar in histogram B (reflecting the state of the i-th bin in tumor B). The difference measure D is defined as:

$$D = \sum_{i=1}^{N} |A_i - B_i|.$$

The difference measure of the two histograms in Figure 4.5 is equal to:

$$D = |4-4| + |2-1| + |3-3| + |1-2| + |2-2| + |3-4| = 3.$$

If the two histograms coincide, the difference measure is equal to zero. In general, the smaller the difference measure, the more likely it is that the tumors are clonal. The histogram difference method is a relatively simple method that can be used to measure the degree of clonal relatedness between two tumors.

However, looking more closely at the histogram difference method it can be seen that the method is not completely fair for this purpose as the penalization of discordant events is unevenly distributed. For instance, if for one tumor the bin is a loss (state 1) and for the other tumor the bin is CN-LOH (state 3), then this combination is penalized more than when one tumor has a gain bin (state 4) and the other tumor a CN-LOH bin (state 3). In other words, a gain and CN-LOH combination is regarded as more clonal than a loss and CN-LOH combination which is not correct. A CN-LOH region has copy number two and is therefore equally close in copy number to a gain (copy number larger than three) and loss (copy number smaller than one) region. Therefore, the histogram difference method is not fair. Note that, if the CN-LOH state was excluded from the analysis and a gain would have been encoded as a 3, the histogram difference method would have produced fair comparison scores. However, as mentioned in Section 3.2, being able to distinguish CN-LOH events from normal events is important when determining the clonality status of a tumor pair. Hence, the histogram difference method is not suitable for this problem.

4.4.2. Corrected histogram difference

In order to solve the problem mentioned in the previous subsection, the histogram difference method is slightly modified. Instead of taking the absolute differences of the states, a different scoring system is applied:

- If the i-th bar in both histograms have the same height, the i-th bin is concordant. A concordant event (e.g. both gains, both normal etc.) gets score 0.
- If the i-th bar in one histogram is equal to 1 and in the other histogram equal to 4, a gain-loss combination occurs. A discordant event that is a gain-loss combination gets score 2.
- If the i-th bar in the histograms do not have equal height and are not a gain-loss combination, the i-th bin gets score 1.

Define $N_{\text{gain-loss}}$ as the number of bins that are a gain-loss combination and $N_{\text{other discordant}}$ as the number of bins that are discordant but not a gain-loss combination. The corrected difference measure $D_{\text{corrected}}$ is defined as:

$$D_{\text{corrected}} = 2N_{\text{gain-loss}} + N_{\text{other discordant}}.$$

The scoring system defined above penalizes a gain-loss discordant event more than any other discordant event as gain-loss combinations are not common to occur when two tumors are of clonal origin. The smaller the corrected difference score, the more similar the two tumors are.

Given the corrected difference scores, the significance of the scores can be determined by means of constructing a reference distribution assuming independence as described in Section 4.2. If the p-value of the corrected difference score falls below the significance level α, where the p-value is defined

as the proportion of observations in the independence distribution which are smaller than or equal to the corrected difference score, the null hypothesis assuming independence is rejected. In other words, the pair is considered to be clonal if the corrected difference score falls below the α-th percentile of the independence distribution.

The corrected histogram difference method is a simple comparison method that may serve as an alternative. However, in this study, the corrected histogram difference method performed worse than the Log LR and adapted SI method. The method is more conservative in detecting clonal pairs than the other two comparison methods: if a pair was correctly labeled as clonal by the corrected histogram difference method, it was also outputted as clonal by at least one of the two other comparison methods. In other words, the method did not contribute much to the overall results. Therefore, the corrected histogram difference method was not included as a comparison method in the model.

4.4.3. Wasserstein distance

Another metric that has been examined is the Wasserstein distance. This metric describes the distance between two probability distributions, i.e. how easily can a density function be rearranged to obtain another density function. In order to apply the Wasserstein distance in this setting, the segmentation profiles, consisting of the states of the bins, need to be converted into a probability distribution. This was attempted by transforming the states of the bins into a histogram, where the height of the i-th bar reflects the state of the i-th bin. A cumulative distribution function was constructed by means of taking the cumulative sum of the states of the bins. Given the vectors of states $s^{(k)}$ for $k = 1, 2$, the Wasserstein distance is set equal to the sum of the absolute differences in the cumulative sums:

$$W = \sum_{i=1}^{N} \left| \text{cumsum}_i(s^{(1)}) - \text{cumsum}_i(s^{(2)}) \right|,$$

where cumsum_i is the cumulative sum until element i in the sequence. The lower the Wasserstein distance score, the more similar the two tumors are.

However, a problem arises with the Wasserstein distance as more differences sometimes yield a smaller score. In other words, the Wasserstein distance interprets a pair with more changes as more clonal. For this, consider the following example of two pairs consisting of 4 bins:

- Pair 1: primary tumor has sequence (4,2,3,1), the secondary tumor has sequence (4,1,2,3). There are two differences between the primary and secondary tumor. The cumulative sum of the primary tumor is (4,6,9,10) and the cumulative sum of the secondary tumor is (4,5,8,10). The sum of the absolute difference in the cumulative sums is equal to 2.
- Pair 2: primary tumor has sequence (4,2,3,1), the secondary tumor has sequence (4,3,3,1). There is one difference between the primary and secondary tumor. The cumulative sum of the primary tumor is (4,6,9,10) and the cumulative sum of the secondary tumor is (4,7,10,11). The sum of the absolute difference in the cumulative sums is equal to 3.

In the example above it can be seen that the second pair has a larger score, even though there are less differences between the states of the samples. This is not desirable in this setting as a larger concordance in states should yield a smaller score. The reason why this phenomenon occurs has to do with the fact that the cumulative sum is not truly a cumulative distribution as the states of the bins do not always add up to the same number. In other words, the amount of mass differs per tumor so that the Wasserstein distance can not be employed for this purpose. A more detailed description about the workings of the Wasserstein distance and why it can not be used in this setting can be found in Appendix B.

Nevertheless, the Wasserstein distance method has been tested and similar results as the corrected histogram difference method were obtained. In conclusion, it can be stated that the more complex comparison methods perform considerably better than the simpler proposed alternatives. In the remainder of this thesis, the Log LR and adapted SI will be used as the two comparison methods.

5

Simulations

This chapter investigates the workings of the comparison methods introduced in Chapter 4 by means of simulations in which artificial pairs are constructed. For instance, how does the distribution of a sample influence the resulting p-value of the comparison method? Section 5.1 describes the data that will be used in the simulations and how the data is preprocessed before it is given as input to the segmentation algorithm. Section 5.2 examines the distribution of the aberrant events in the dataset as assigned by the two segmentation algorithms. Are gains or losses more likely to occur in breast cancer? Finally, Section 5.3 explains how the artificial pairs used in the simulations are created from the data and elaborates on the results of the simulations.

5.1. Data description

The dataset that is being used in the simulations consists of 50 *sporadic unilateral breast cancer* patients. In sporadic cancer, the gene mutations that cause the cancer to develop are acquired and not inherited. It is estimated that approximately 90-95% of breast cancers are sporadic [7]. The tumor material of the 50 patients were *fresh frozen*, which means that the tumor was frozen in liquid nitrogen within 30 to 60 minutes after surgery excision. The DNA of the 50 samples was extracted from the tumor material and analyzed on a customized Illumina Infinium Global Screening Array (GSA) V3 chip to which extra SNP positions were added. The total number of SNPs on the chip is equal to 730,059.

It should be noted that the 50 samples that are being used in the simulations are a subset of a larger GenomeStudio project consisting of 283 sporadic unilateral breast cancer patients. As the number of SNPs that are being analyzed is quite large and the comparison methods use permutation techniques to determine the significance of the comparison scores, using all 283 samples would make the simulations computationally very intensive. In order to reduce the computation load of the simulations, a subset of 50 samples is used instead. [31] and [36] have successfully applied comparison methods on 24 and 22 pairs respectively. Therefore, a sample size of 50, resulting in 50 artificial pairs, is sufficient to produce reliable simulation results.

5.1.1. Preprocessing the raw data

Before the LRR and BAF values are given as input to the segmentation algorithms, the data first needs to be preprocessed. The preprocessing procedure consists of the following steps:

1. SNPs that are present on the Y or mitochondrial chromosome are removed from the analysis. As all patients are female, the Y chromosome is not present in females. The mitochondrial chromosome is excluded from the analysis as it is not considered in both segmentation algorithms, i.e. SNPs on the mitochondrial chromosome are not segmented. Hence, the chromosomes included in the comparison analysis consist of the chromosomes 1-22 and X.

2. Some SNPs may not be scanned properly resulting in missing values for the LRR and BAF values. SNPs which have a NA for either the LRR or BAF value for at least one of the samples are

removed. In other words, only SNPs which have a known LRR and BAF value for all samples are included.

3. Some SNPs in the data occur at the same position at the same chromosome. These double SNPs capture two variations, for instance an A/C and an A/G variation. Keeping the double SNPs in the analysis may introduce a bias in the result of the segmentation algorithm. Therefore, only SNPs with a unique location are considered by the segmentation algorithm.

After following the preprocessing steps, 692,101 SNPs (94.8% of all the SNPs on the chip) remain and are given as input to the segmentation algorithms.

5.2. Distribution of the data

The segmentation algorithms divide the raw data into segments, where each segment has its own copy number. Given the segmented profiles, each SNP is assigned to one of the four states and the SNPs are binned together as described in Section 4.1. Using distance based bins with a distance of 1 MB, 2841 bins are defined in total. Figure 5.1 shows the first six bins of the dataframe.

	chr	startpos	endpos	nprobes	chrarm
1	1	565433	1561051	282	1
2	1	1586842	2583272	300	1
3	1	2700215	3698166	360	1
4	1	3700639	4698677	306	1
5	1	4701258	5695978	325	1
6	1	5701323	6699327	365	1

Figure 5.1: First six bins of the dataframe.

For each bin, the chromosome and chromosome arm (coded a 1 fot the p-arm and a 2 for the q-arm) on which the bin is located as well as the start position, end position and the number of SNPs in the bin are reported in the dataframe. As mentioned in Section 4.1, bins that are comprised of less than 10 SNPs are excluded from the comparison analysis, which leaves 2828 bins for analyses. Figure 5.2 shows a histogram of the number of SNPs for the 2828 bins.

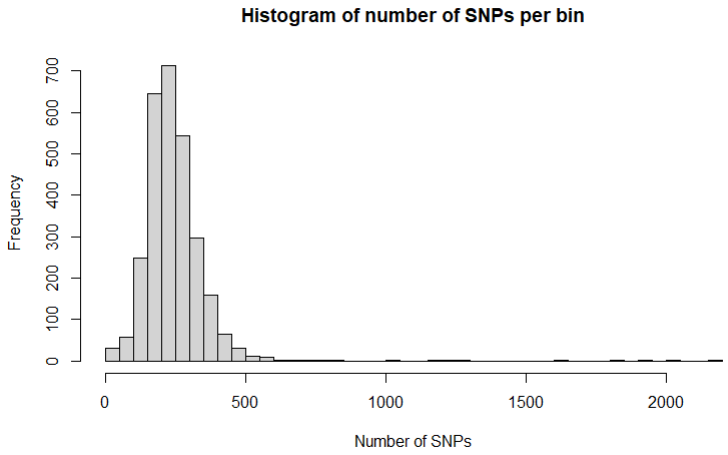


Figure 5.2: Number of SNPs in the 2828 bins taken into account in the analyses.

Of the 2828 bins, 2740 bins (96.9%) consist of more than 100 SNPs. Once the states of the bins are determined, the distribution of gains, losses, normal and CN-LOH states can be derived for each sample.

5.2.1. ASCAT

Of the 50 samples fitted by ASCAT, (total runtime $\approx$ 40 minutes), two samples had a goodness of fit below 80% so that a segmentation profile could not be constructed. Looking at the raw data plots of

the failed samples, a larger amount of noise appeared to be present compared to other samples. As a result, ASCAT had difficulties segmenting the data resulting in a lower goodness of fit. Figure 5.3 shows the raw data of sample 2393, which ASCAT failed to fit, and the raw data of sample 563 which had a goodness of fit of 90.9%.

In order to also obtain a segmentation profile for the two samples which ASCAT failed to fit, the minimal goodness of fit threshold was lowered in the ASCAT source code to 50% after which ASCAT gave a goodness of fit for the previously failed samples of 75.5% (sample 2393) and 76.2% (sample 2328) respectively. Since the goodness of fit of both samples is still quite close to the default threshold of 80%, we decided that the two samples would be included in the analysis of the distribution of the data.

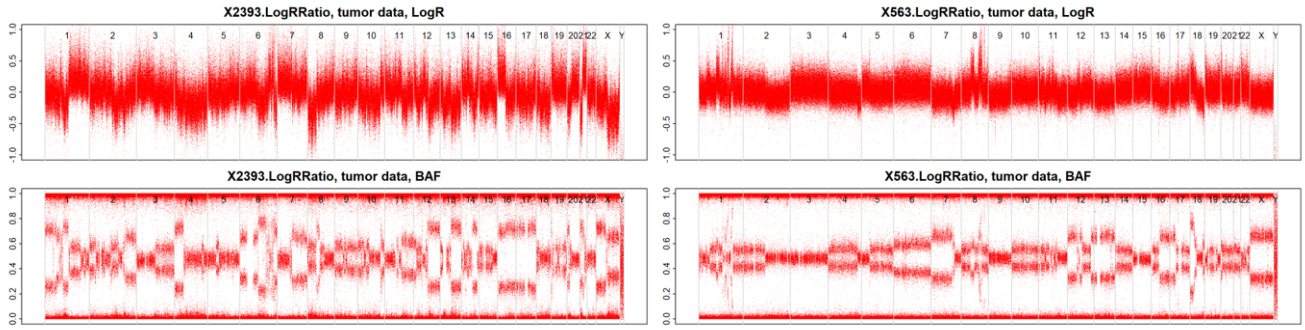


Figure 5.3: Raw data of sample 2393 (left), which ASCAT failed to fit, and sample 563 (right) which ASCAT successfully fitted with a goodness of fit of 90.9%.

For all 50 samples the percentage of bins which are a gain, loss, CN-LOH and normal were determined and summarized in the barplot in Figure 5.4. Each bar in the barplot corresponds to one sample and is colored according the frequencies of the states.

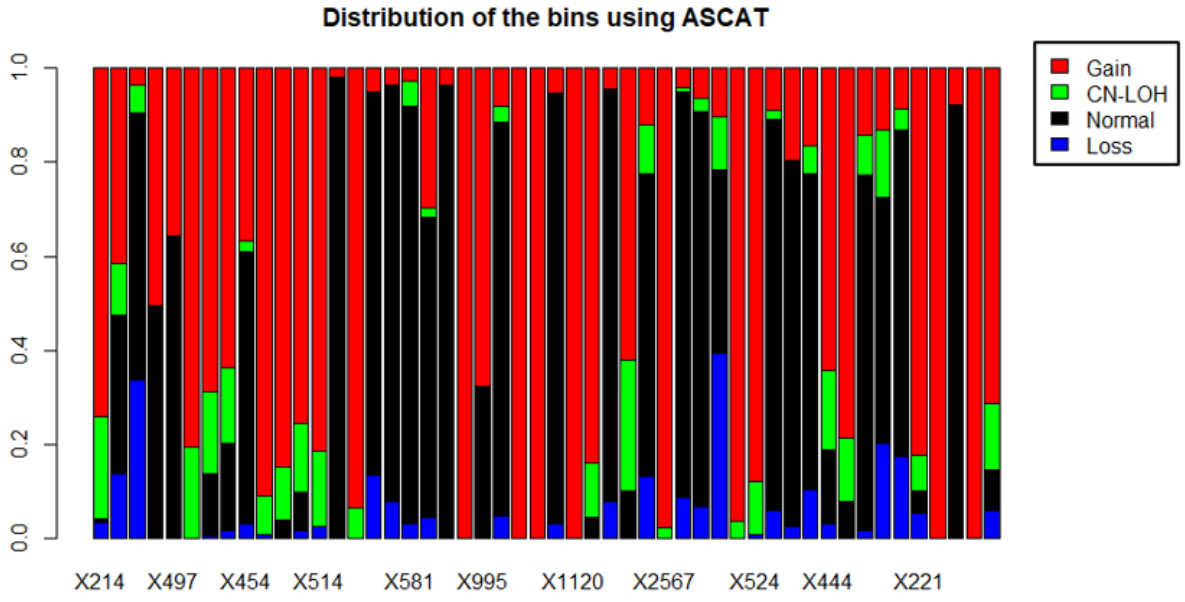


Figure 5.4: Distribution of the bins for all 50 samples fitted by ASCAT.

In Figure 5.4 it can be seen that gains are very common to occur in the 50 samples. 5 out of 50 samples solely consist of bins which are a gain. Moreover, for 26 out of 50 samples the percentage of bins that are a gain is above 50%. Losses, on the other hand, are more rare. 17 out of 50 samples do not have any loss bins and the maximal percentage of loss bins that occur in a sample is equal to 39.2%. Next

to that, there are 22 samples for which the percentage of normal bins is above 50%.

Figure 5.5 shows the histograms of the percentages of bins that are a loss, normal, CN-LOH and gain in all samples. The histograms clearly show that losses are the most rare as there is a peak close to zero. Moreover, gains are even more common to occur than normal states as 18 samples have at least 75% gain bins while only 14 samples have more than 75% normal bins. Next to that, a larger number of samples has no normal states than no gain states. Regarding CN-LOH, the percentages range between 0 and 30%. Comparing the histogram of the CN-LOH percentages with the histogram of the loss percentages, it can be concluded that losses are more rare to occur than copy-neutral states.

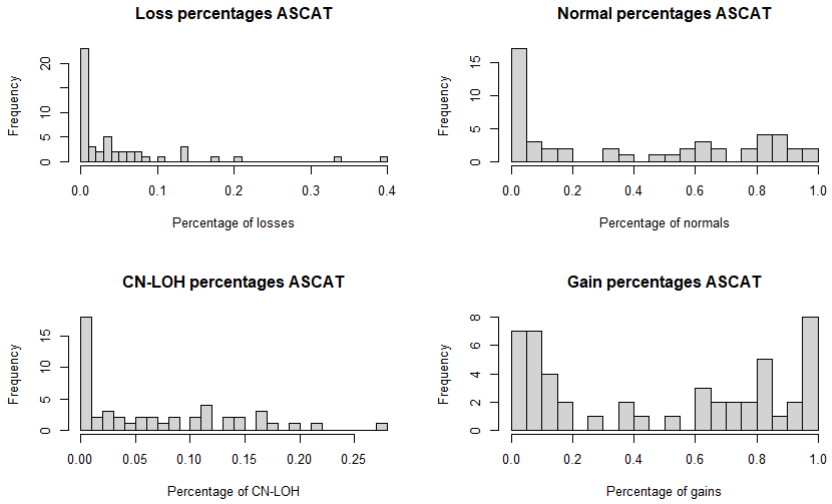


Figure 5.5: Histograms of the loss, normal, CN-LOH and gain percentages.

In general, it can be concluded that if the 50 samples are segmented with ASCAT, gains are most common. This is also reflected in the estimated ploidy numbers: 26 out of 50 samples have a ploidy above 2.5. Figure 5.6 shows the distribution of the ploidy estimates of the 50 samples.

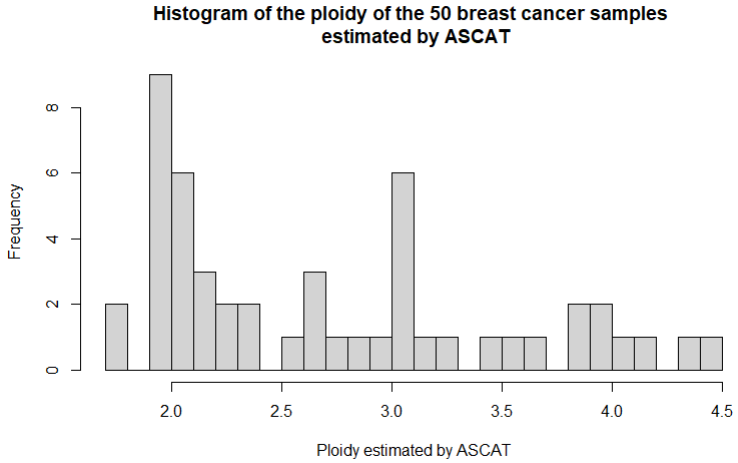


Figure 5.6: Ploidy estimated by ASCAT.

In Figure 5.6, two clear peaks at ploidy 2 and ploidy 3 can be seen. The peak at ploidy 2 is to be expected as not all tumors will show many aberrations and will thus be diploid. The peak at ploidy 3 and the majority of samples having a ploidy above 2 explains why gains are so prevalent. When the ploidy number is above 2, the average copy number lies above two so that, if copy number two is

chosen as a baseline, the majority of the bins will be a gain. As the number of samples with a ploidy number below 2 is quite small, losses do not occur that often.

5.2.2. OncoSNP

In contrast to ASCAT, OncoSNP takes into account all three confounding variables introduced in Section 3.1. However, a downside of using both the normal cell contamination and tumor heterogeneity mode in OncoSNP is that the segmentation can be computationally demanding. When both modes are switched on, OncoSNP takes around 4-6 hours to analyze one sample compared to 10 minutes when only the normal cell contamination mode is used.

In order to investigate if it is worth the additional runtine with regards to the obtained results, the first 10 samples of the 50 samples are run in both instances. As OncoSNP fits 10 different models (two ploidy baselines and five different ranks), the difference in the states of the bins is investigated for both ploidy models and all ranks. Figure 5.7 shows a plot of the match percentage in the bin states per chromosome for the model with the highest likelihood (Ploidy No 1) for sample 530. Each line in the plot corresponds to a rank. The x-axis represents the 23 chromosomes (1-22 and X) which are analyzed by OncoSNP. In the plot, it can be seen that the match percentage is 100% (i.e. all bin states in the two settings coincide) for ranks 1-3 and that slight differences in the states of the bins are present for ranks 4 and 5. Nevertheless, the match percentage is also quite high for ranks 4 and 5.

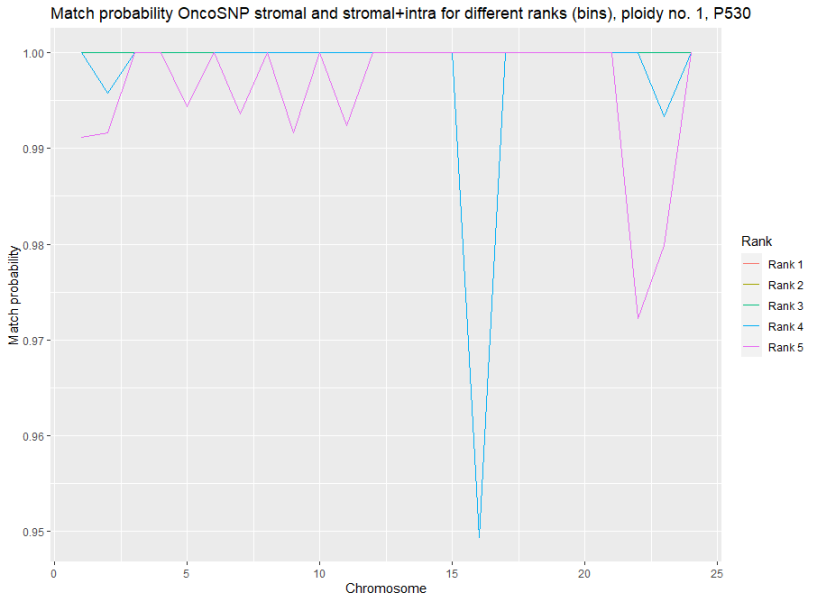


Figure 5.7: Match percentage of the bin states in the two settings for sample 530.

A large concordance was found in the states of the bins in both settings: the mean concordance over all ranks and samples was equal to 97.2% for Ploidy No 1 and 99.8% for Ploidy No 2. Therefore, as a means to reduce the computation time of OncoSNP, we decided that only the aneuploidy and normal cell contamination mode will be used in the analysis. Furthermore, by using only the aneuploidy and normal cell contamination mode for OncoSNP the results can be compared with the output of ASCAT as the latter also corrects for the same confounding variables.

Since OncoSNP fits 10 different models, the question remains which ploidy baseline and rank should be chosen. In this thesis, the model with the highest likelihood (Ploidy No 1) is chosen together with rank 3. As mentioned in Section 3.4, the higher the rank the more detailed the segmentation profile becomes. However, a higher rank also yields smaller segments in the profile which may result in a higher error rate. For that reason, we decided that rank 3 would be used for the segmentation profiles: rank 3 gives a relatively detailed profile while at the same time not including too many small segments in the profile. Figure 5.8 shows the percentage of bins that are a gain, loss, CN-LOH and normal state

for each sample when OncoSNP is used as segmentation algorithm.

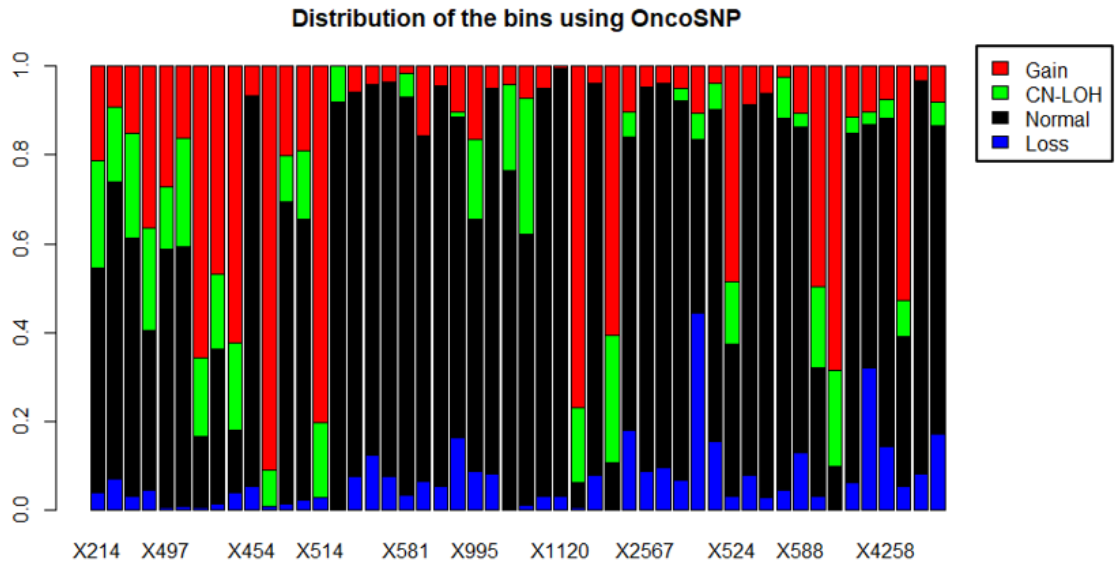


Figure 5.8: Distribution of the bins for all 50 samples fitted by OncoSNP.

Looking at Figure 5.8, it can be seen that the majority of samples mainly consist of normal states: 37 out of 50 samples consist of more than 50% of normal bins. When it comes to the aberrant events, gains appear to be more common than losses. There are 8 samples whose bins consist of more than 50% of gains, while there are no samples with more than 50% losses. Figure 5.9 displays the histograms of the loss, normal, CN-LOH and gain percentages for all samples.

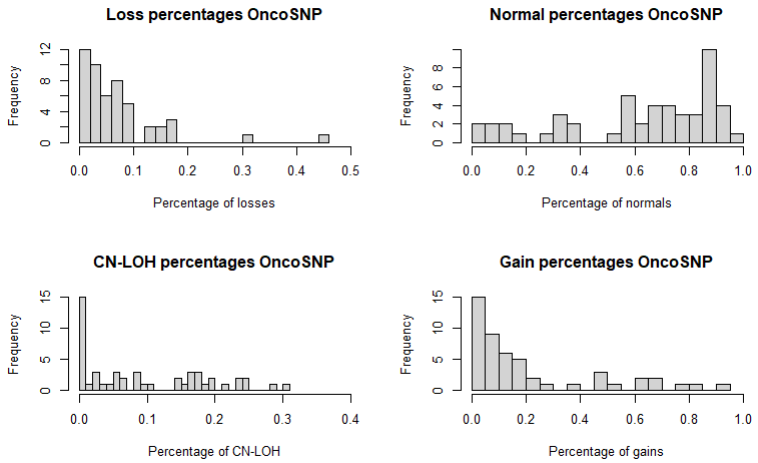


Figure 5.9: Histograms of the loss, normal, CN-LOH and gain percentages in all samples.

The histograms of Figure 5.9 show that the most common state in the 50 samples is the normal state. This has to do with the fact that for most of the samples (45 out of 50) the model with the highest likelihood is the model assuming that the sample is diploid. In the diploid model, the average copy number is assumed to be two so that many SNPs and consequently bins will be assigned as normal. For the other 5 samples, 4 samples have a unique solution, i.e. the diploid and triploid model converge to the same values for the Log R adjustment and normal cell contamination. The converged samples all have a ploidy number close to three. Next to that, there is one sample for which the triploid model is the best fitting model. If copy number two is used as a baseline, the converged samples and the sample whose

best fitting model is the triploid model will mainly consist of gains. As a result, similar to ASCAT, losses do not happen that often as gains.

As mentioned in Section 3.4, when the two models converge to the same solution the SNP data contains sufficient information to deduce the normal cell contamination and necessary Log R adjustment without making any assumption on the ploidy configuration. The normal cell contamination percentages and ploidy numbers for the converged samples as estimated by ASCAT and OncoSNP were very similar, see Table 5.1.

Sample	Normal cell % ASCAT	Normal cell % OncoSNP	Ploidy ASCAT	Ploidy OncoSNP
507	37%	30%	3.19	3.2
514	48%	50%	2.98	3.2
571	13%	10%	2.85	2.9
2393	47%	50%	3.09	3.1

Table 5.1: Normal cell contamination percentages and ploidy estimated by ASCAT and OncoSNP for the converged samples in OncoSNP.

This entails that ASCAT manages to detect the same features in the data as OncoSNP leading to similar estimates. However, for the samples that did not converge in OncoSNP, the ASCAT and OncoSNP estimates are not necessarily close to one another: the ASCAT and OncoSNP estimates may be similar but can also be drastically different.

5.2.3. Prevalence of gains

Comparing the results of ASCAT and OncoSNP with one another, gains are more prevalent in the profiles fitted by ASCAT. The reason for this has to do with the fact that the ploidy estimates of ASCAT are higher than the ploidy estimates of OncoSNP. As the majority of samples is fitted by the model assuming diploidy in OncoSNP, the estimated ploidy for these samples will also be around 2.

Nevertheless, the output of both ASCAT and OncoSNP show that duplications in the DNA are more likely to happen than deletions. The question that arises is: are gains generally the most common aberrant event in breast cancer tumors? Figure 5.10 taken from [48] shows the frequencies of gains (red) and losses (green) in 91 breast cancer patients analyzed by ASCAT where copy number two is taken as a baseline.

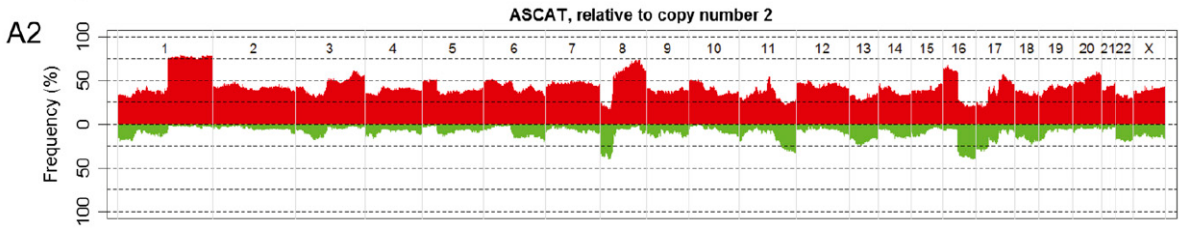


Figure 5.10: Gain and loss frequencies determined by ASCAT for each SNP position for 91 breast cancer patients.

The SNP chip used for this case is the Illumina 109K SNP array. Figure 5.10 shows that, similar to the 50 sporadic breast cancer patients, gains are more common to occur than losses. Moreover, specific gain and loss regions can be defined: gains are likely to occur on chromosome 1q, 8q and 16p, while losses are mostly present on chromosome 8p and 16q.

A possible hypothesis which explains why gains are more common than losses might be that gains can keep on gaining whereas losses can more easily be turned into a copy-neutral or (very rarely) a gain state. If a gain occurs in the DNA and this gain is beneficial for the tumor, then it can keep on acquiring additional gains in subsequent cell divisions to for example copy number four or five. Once it has reached a relatively high copy number, it will remain a gain with a high probability. Losses, on the other hand, consist of copy numbers one and zero. If the copy number is one and the tumor cell does not benefit from this loss, the loss might relatively easy return to a higher copy number during later cell

divisions. For regions having copy number zero, the copy number will remain zero as all alleles within the region are lost.

The hypothesis stated above describes a possible cause why gains are more common than losses. However, it is unknown whether losses are indeed more easily reversed than gains. Hence, the prevalence of gains may also be a characteristic of breast cancer tumors in general.

5.3. Creating artificial pairs

In this section the behavior of the comparison methods will be examined by means of simulating artificial pairs. Given the segmentation profiles of the 50 sporadic breast cancer patients, changes are made to the segmentation profiles. By doing this, artificial pairs are constructed which can be used to test how the output of the comparison methods are influenced. In other words, to which extent can mutations be added to a sample so that the comparison method still detects the artificial pair as clonal and which distribution characteristics may influence this? Note that the objective of the simulations is to discover certain features of the comparison methods and not to compare the two comparison methods in terms of sensitivity.

5.3.1. Workflow description

First, adding mutations means that normal bins go to an aberrant state but aberrant events can also be changed back to a normal state or another aberrant state. In other words, a mutation can be seen as a state change that occurs in a bin.

In order to add the mutations to the samples as realistic as possible, the probability of being in a particular state in a bin are being derived from a larger dataset consisting of copy number data from 201 ER-positive sporadic breast cancer patients. The tumor DNA of the patients was examined on an Affymetrix 100K SNP chip which consists of approximately 100,000 SNP positions. For each sample, only the SNP positions for which both the copy number and BAF value are known are taken into account. The latter is needed to discriminate the normal state from the CN-LOH state. The copy number for each SNP is determined by means of applying a threshold on the LRR data. Given a threshold a, the copy number for a SNP is larger than two if the LRR value lies above a and smaller than two if it is below a. For each sample, the most common state in each bin is determined in a similar way as in Section 4.1. Once the bin states are determined for all 201 samples, the frequency that a gain, loss, normal or CN-LOH state occurs in a bin can be computed. These frequencies can be used as the probabilities of being in a particular state in a certain bin. Figure 5.11 shows part of the probability matrix that is derived from the 201 samples.

	Chr	start	end	Gain	Neutral	CN.LOH	Loss	est.nr.SNPs
1	chr1	2700215	3698166	0.9104478	0.000000000	0.014925373	0.074626866	18
2	chr1	3700639	4698677	0.6965174	0.000000000	0.059701493	0.243781095	23
3	chr1	4701258	5695978	0.6865672	0.000000000	0.019900498	0.293532338	38

Figure 5.11: Part of the probability matrix consisting of the frequencies of a gain, loss, normal and CN-LOH state for each bin in the Affymetrix dataset.

Note that the normal state is called neutral in the probability matrix. Figure 5.12 shows the histograms of the loss, normal, CN-LOH and gain probabilities of the Affymetrix dataset. The y-axis of the histograms reflect the frequency with which the probabilities occur in the Affymetrix dataset for the different states. Figure 5.12 clearly shows that normal and CN-LOH states are less likely to occur than gain and loss states. The probabilities of the gains and losses look evenly distributed: the probabilities of both states are spread uniformly over the interval $[0, 1]$. This is not in line with the observations of Section 5.2 where gains were found to be more common than losses for both segmentation algorithms. Next to that, the probabilities of the normal states in the Affymetrix dataset are centered around zero which is not the case for the distribution of the 50 sporadic samples. The reason why the results are different for the Affymetrix dataset has to do with the fact that the segmentation is done differently. The distinction between a gain, loss and normal state for the Affymetrix dataset is made based on the LRR values of the SNPs, where each SNP is considered separately. As there is noise present in the LRR data, this may result in different states for consecutive SNPs. For instance, a gain SNP may be followed by a

SNP that is assigned as normal and then followed by a loss SNP. The output of ASCAT and OncoSNP, on the other hand, divide the genome into segments where all the SNPs in the segment have the same copy number. As a result, the most common states in the bins are different for the Affymetrix segmentation approach than for the other segmentation approaches. Even though the probabilities of the Affymetrix dataset are not similar to the distributions found for the 50 sporadic samples, they can still be employed in the simulations to discover the workings of the comparison methods. However, it should be noted that the way in which mutations are added to the samples may not be realistic.

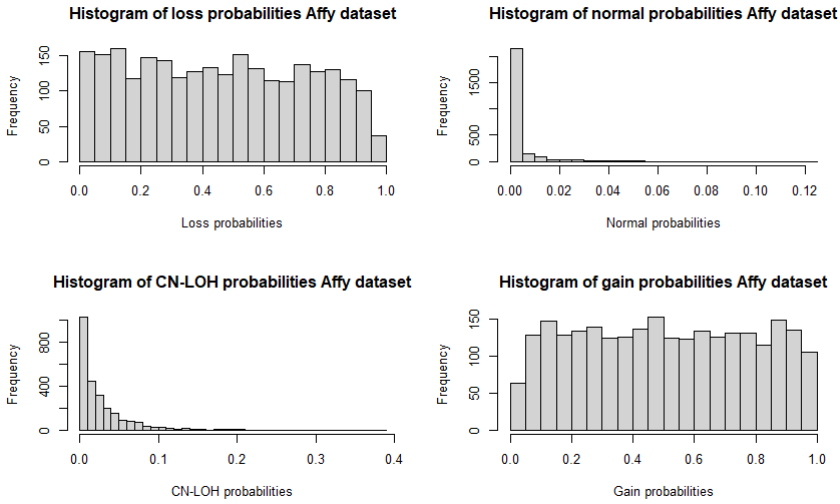


Figure 5.12: Histogram of loss, normal, CN-LOH and gain probabilities of the derived from the Affymetrix dataset.

Since the number of SNPs on the Affymetrix chip is significantly lower than the SNP chip of the 50 sporadic breast cancer patients, some bins defined in Section 5.2 will not contain any SNPs. As a consequence, the probabilities for these bins can not be derived. Next to that, for some bins the number of SNPs within the bin may be very small, so that the probabilities are not accurate enough to be used in the simulations. For this reason, bins which contain less than 10 SNPs in the Affymetrix data are excluded from the simulations. Given the 2828 bins defined in Section 5.2, only 2546 bins will be taken into account in the simulation procedure. The other 282 bins are not considered by the comparison methods in the simulations, i.e. the artificial pairs are only compared on the 2546 bins for which the probabilities of being in a certain state can be accurately derived.

Mutations will be added to the segmentation profiles of the samples using a specific workflow. The workflow of the simulation is as follows:

- For each percentage $x \in (0, 1, 2, \ldots, 99, 100)$:
  - For each simulation i in 1,…, N, where N is the total number of simulations to be conducted for each percentage x:
    - For each sample in the cohort, add mutations to the states as follows:
      - Take a subset of $x\%$ of the bins at random, where the subset is uniformly sampled.
      - For each bin that is within the subset, look at the state of the bin.
        1. If the state of the bin is a loss, the state of the bin changes with a certain percentage $a_1\%$ which are defined below. If the loss state is changed, it can only go a CN-LOH state (state 3). Note that a loss region can not go back to a normal state: a loss can only go back to copy number two if the remaining allele is copied yielding either AA or BB genotypes. Next to that, it is assumed that a loss can not become a gain in the case of clonal tumors.
        2. If the state of the bin is normal, the state of the bin changes with probability $1 - \mathbb{P}(\text{bin is in state 2})$, where $\mathbb{P}(\text{bin is in state 2})$ is the probability that the bin

is normal taken from the Affymetrix dataset. If the state is changed, it can go to a loss (state 1), CN-LOH (state 3) or a gain (state 4). The probabilities of going to either state 1, 3 or 4 is determined by the probabilities resulting from the Affymetrix chip where rescaling is used so that the probabilities add up to 1. For instance, if for a certain bin the probabilities are as follows:

$$\mathbb{P}(\text{loss}) = 0.1,$$
$$\mathbb{P}(\text{CN-LOH}) = 0.1,$$
$$\mathbb{P}(\text{gain}) = 0.3,$$

the probabilities will be rescaled to:

$$\mathbb{P}(\text{loss}) = 0.2,$$
$$\mathbb{P}(\text{CN-LOH}) = 0.2,$$
$$\mathbb{P}(\text{gain}) = 0.6.$$

3. If the state of the bin is CN-LOH, the state of the bin changes with probability $1 - \mathbb{P}(\text{bin is in state 3})$, where $\mathbb{P}(\text{bin is in state 3})$ is the probability that a bin is CN-LOH in the Affymetrix dataset. If the state is changed, it can go to either a loss (state 1) or a gain (state 4). Note that a CN-LOH region can never go back to normal as the other allele is lost in the case of a CN-LOH event. The probabilities of going to a loss or a gain state are determined by the probabilities of the Affymetrix dataset where the probabilities are rescaled.
4. If the state of the bin is a gain, the state of the bin changes with a certain percentage $a_4\%$ which are defined below. If the state is changed, it can either go to a normal (state 2) or a CN-LOH state (state 3). The probabilities of going to one of the two states is derived from the Affymetrix dataset where the probabilities are rescaled. It is assumed that a gain never turns into a loss when two tumors are of clonal origin.

◦ Once the mutations are added to the states for each sample, the corresponding Log LR and adapted SI scores are computed for each artificial pair.

◦ The significance (i.e. the corresponding p-value) of the scores are computed by means of constructing a reference distribution assuming independence.

The segmentation profiles used in the simulations will be the profiles fitted by ASCAT. Note that the chosen segmentation algorithm has a minor influence on the simulation results as we are mainly interested in how the distribution of a sample influences the results of the comparison methods. Therefore, even though ASCAT and OncoSNP may yield different profiles, the behavior of the comparison methods will be similar when artificial clonal pairs are created. Hence, only the ASCAT profiles are used as input for the simulations.

As described in the simulation workflow, if a bin is a loss or a gain the state changes with probability a_1 and a_4 percent respectively. In the simulations, three different settings are used:

- Setting 1: $a_1 = a_4 = 10\%$. This setting implicitly assumes that gains and losses are beneficial for a tumor: a gain or loss will not be easily reverted and remains with a high probability (90%) a gain or a loss.
- Setting 2: $a_1 = a_4 = 50\%$. In this setting, the gains and losses are more easily changed to other states.
- Setting 3: $a_1 = 1 - \mathbb{P}(\text{bin is in state 1})$ and $a_4 = 1 - \mathbb{P}(\text{bin is in state 4})$, where the probabilities are derived from the Affymetrix dataset. In contrast to the previous two settings, the change probability differs per bin.

Next to the p-values, which determine whether the null hypothesis stating that the two tumors are independent can be rejected or not, the number of bin changes for each sample and each percentage x

is also recorded in the simulations. For example, if 25 percent of the bins are taken into consideration and almost all bins are a gain (state 4), then the total number of bins that change will be less than 25 percent on average as the states only change with a certain probability.

Once the p-values are computed for each percentage x, the average p-value can be determined for each comparison method, artificial pair and percentage x to see how mutations affect the results of the comparison methods. In other words, the evolution of the p-value can be examined for samples with different distributions. It should be noted that the workflow description above describes a possible way how a secondary clonal tumor may evolve from a primary tumor but it is unsure whether this generally holds for all clonal pairs. As discussed previously, the probabilities coming from the Affymetrix dataset do not mirror the distributions of the 50 sporadic samples. Therefore, it may be questioned whether the way in which mutations are added to the samples are realistic. Nevertheless, the main objective of the simulations is to discover certain characteristics of the comparison methods, e.g. how is the distribution of a sample of influence on the resulting p-value? As the distribution of the sample is already determined before the simulations, the way in which mutations are added to the sample do not necessarily have to be realistic.

As 50 artificial pairs are simulated, $50 \cdot 49 = 2450$ comparisons between independent pairs need to be made in order to construct the independence distribution in each simulation. Since this number is quite large, the runtime of the simulations may become fairly long. In order to make the runtime of the simulations feasible, $N = 50$ simulations are done for each percentage x. This number of simulations had a total runtime between 16 and 24 hours depending on which setting was used.

The results of the third setting, where the change probabilities for the gain and loss states are derived from the Affymetrix dataset, were very similar to the results of the second setting. Investigating this, it appeared that the probability that a certain bin is in a gain or loss state are close to 50% on average: 46.1% and 50.6% for the average probability for a loss and gain, respectively. As a result, the simulation results of the third setting are similar to the results of the second setting and are therefore not included in the result sections.

In the next two subsections, the results of the simulations for the two comparison methods and two settings are discussed in more detail. For each comparison method and each setting the influence of the distribution of the sample on the resulting p-value is examined. A summary of the simulation results is given in Section 5.3.4.

5.3.2. Results Log LR

Figure 5.13 shows the evolution of the average p-values by the Log LR for the two different settings. The x-axis in the plots correspond to the fraction of bins taken into account in the simulation and the y-axis represents the average p-value. Note that the fraction of bins taken into account do not reflect how many changes have truly occurred within an artificial pair as these depend on the states of the bins.

In the left panel of Figure 5.13, where the gain and loss states change with probability 10%, it can be seen that for some samples the average p-value of the artificial pair remains close to zero, even if all bins are taken into consideration, while for other samples the average p-value sharply increases when the percentage of bins considered increases. Samples with a relatively high p-value are mainly samples that consist of many normal and CN-LOH states as these states generally change with a larger probability than gain and loss states. As described in the workflow description, the change probability of a normal and CN-LOH state is defined as 1 minus the probability that the bin is in a normal or CN-LOH state which are derived from the Affymetrix dataset. In Figure 5.12, it can be seen that the probability of being in a normal or CN-LOH state is generally quite small for the Affymetrix dataset resulting in more changes for samples having many normal and CN-LOH states. As more changes are applied to samples with a large number of normal and CN-LOH states, the number of concordant bins within these artificial pairs will decrease leading to a lower Log LR score and hence a higher p-value.

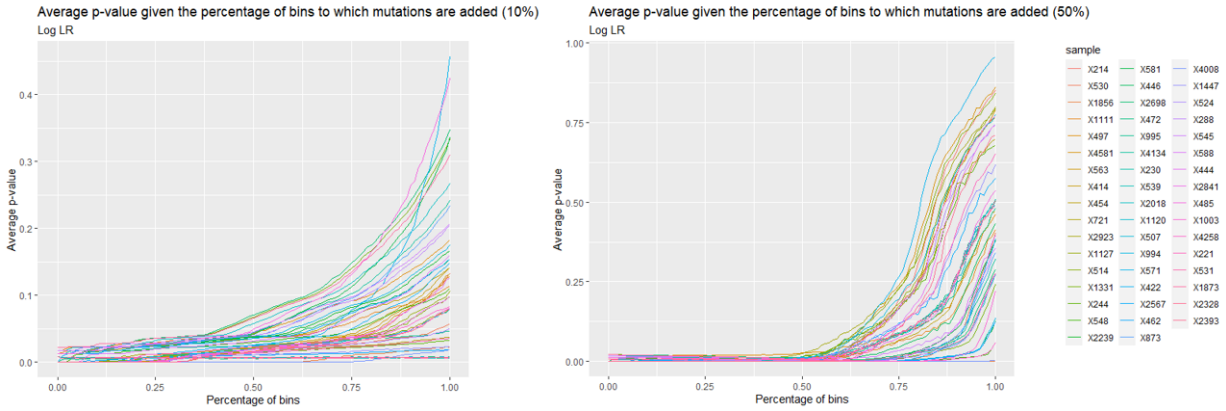


Figure 5.13: Evolution of the average p-values of the 50 artificial pairs for the Log LR method for the first (left panel) and second (right panel) setting.

The right panel of Figure 5.13 shows the average p-values of the artificial pairs when the change probability of a gain or loss is equal to 50%. Comparing this plot to the results of the first setting, it can be seen that the p-values for all artificial pairs remain close to zero for a larger percentage of bins than in setting 1. This is unexpected: as more mutations are added to the samples a higher p-value would be expected. However, when more than 50% of bins are taken into account the p-value suddenly increases even leading to p-values close to 1 for some samples when all bins are considered. The difference in p-values can be explained by the fact that a permutation method is being used to establish the independence distribution. In other words, the reported p-values depend on the distribution of the entire cohort. As mutations are added differently to the samples as in setting 1, the distribution of the simulated pairs will also vary resulting in different p-values. Note that this fact does not only hold for the Log LR method, but for permutation methods in general. It thus appears that for this setting, until a certain threshold of bins the p-value remains low after which there is a steep increase for most p-values. The lowest p-values correspond to samples which have many losses. Since losses also change with a larger probability, losses are even more rare to occur in the entire cohort. As a consequence, the Log LR score becomes even lower for samples with many concordant losses. The highest p-values correspond to samples which consist of more than 50% of gains. Since gains are the most common state, a concordant gain will not contribute that much to the clonality hypothesis resulting in a relatively high Log LR score and consequently high p-value.

In general, it is expected that a larger number of bin changes would yield a relatively higher p-value. Figure 5.14 shows the average p-value of the Log LR method for all percentages x (i.e. for each sample the average of all the average p-values for all percentages x is plotted) taken into consideration against the average of the average number of changes applied to the bins in all percentages x for the two different settings.

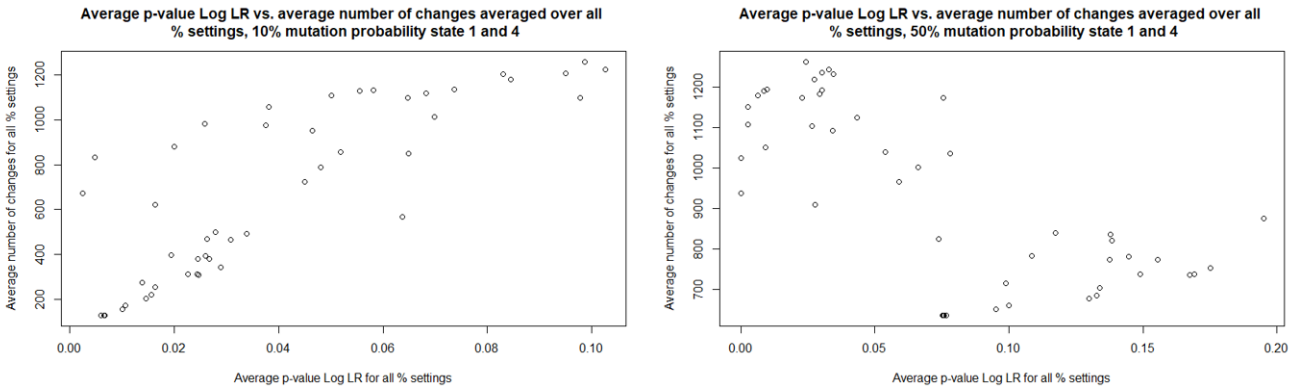


Figure 5.14: Average p-value of the Log LR method averaged over all percentages of bins versus the average number of changes averaged over all percentages of bins of the 50 artificial pairs for the first (left panel) and second (right panel) setting.

The left panel in Figure 5.14, corresponding to the setting where gains and losses change with probability 10%, shows a positive correlation between the average number of changes and the average p-value. Pearson's correlation coefficient is estimated at 0.8309, with a corresponding p-value of $8.183 \cdot 10^{-14}$. The samples with the lowest p-values are the samples that either completely consist of gains or samples which have relatively many losses. Samples whose profiles are solely comprised of gains undergo the smallest number of mutations leading to a higher degree of concordance in the profiles of the artificial pairs and thus a lower p-value. Profiles which contain many losses, on the other hand, are subjected to more state changes in the bins as a significant percentage of the bins are normal states. However, as losses are rare to occur in the data the probability of a double loss is quite small and will with a higher likelihood occur in the clonal setting. When an artificial pair has many concordant losses, the Log LR score will be in favor of the clonality hypothesis resulting in a relatively low p-value.

When the gain and loss states change with probability 50%, a negative correlation seems to be present between the two variables: Pearson's correlation coefficient is equal to -0.7022 with a corresponding p-value of $1.322 \cdot 10^{-8}$. However, looking at the right panel in Figure 5.14 more closely, two groups can be detected by drawing a horizontal line at $y = 900$. The first group (group 1) involves artificial pairs which contain many concordant bins (less than 900 changes on average) but have a high p-value. The second group (group 2), on the other hand, consists of samples which generally undergo many mutations during the simulations (more than 900 changes on average), yet have a relatively low average p-value.

Figure 5.15 displays the histograms of the percentages of bins that are a gain in the samples of group 1 and group 2. The histograms clearly show that the samples in group 1 mainly consists of gains, while the distributions of the samples in group 2 have less gains. When gains and losses are changed with probability 50%, samples with many gains have the highest p-values even though these samples do not undergo many mutations. As stated before, the Log LR method is sensitive to the rarity of an aberrant event. Because of the fact that gains are quite common, a concordant gain will contribute less to the clonality hypothesis than a concordant loss. Moreover, as more mutations are added in comparison to the first setting for samples with many gains, the Log LR score and consequently p-value will increase for these samples.

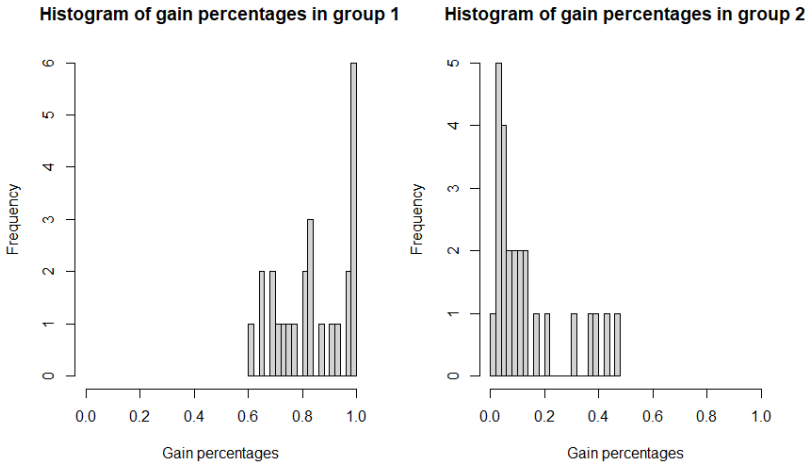


Figure 5.15: Distribution of the percentages of bins that are a gain in group 1 and group 2.

For the Log LR, there seems to be a change point present in the data. Once gains and losses change with a certain probability in the workflow the positive correlation between the average p-value and the average number of changes diminishes. The reason why the positive correlation disappears has to do with the fact that mutations are added to all samples simultaneously: the Log LR uses the probabilities within the cohort to compute the Log LR score for each artificial pair. For each simulation setting and percentage of bins, the state frequencies of the entire cohort differ leading to different Log LR scores. As the positive correlation between the average p-value and average number of state changes is lost

when gains and losses are changed with probability 50%, it may be questioned whether this change probability is realistic.

5.3.3. Results adapted SI

Figure 5.16 shows the evolution of the average p-values outputted by the adapted SI for the two different settings. The left and right panel in Figure 5.16 display very similar results. In both settings, samples which consist of many normal states generally have the highest p-values. For example, the p-value of sample 1331 quickly increases in both cases when only a small percentage of bins is taken into account. This has to do with the fact that sample 1331 consist of more than 99% of normal bins. As the adapted SI only takes into account concordant aberrant events, the adapted SI score of this sample without adding mutations is already relatively low. When mutations are added, the adapted SI score will quickly go towards zero resulting in a very high p-value for this sample. Generally, the larger the number of normal states a sample has, the quicker the p-value increases. Even though both settings show similar results, the p-values of the second setting seem to increase a bit faster than the p-values of the first setting. This makes sense as the change probabilities in the second setting are larger than in the first setting. As a consequence, more changes are applied in the second setting resulting in a smaller adapted SI score and consequently higher p-value. Moreover, for both settings, a larger number of gains implies a smaller number of changes and thus a lower p-value. However, in contrast to the Log LR, samples with many losses have a relatively high p-value. As samples with losses also have many normal states, extra discordant aberrant events are introduced in the artificial pair resulting in a smaller adapted SI score.

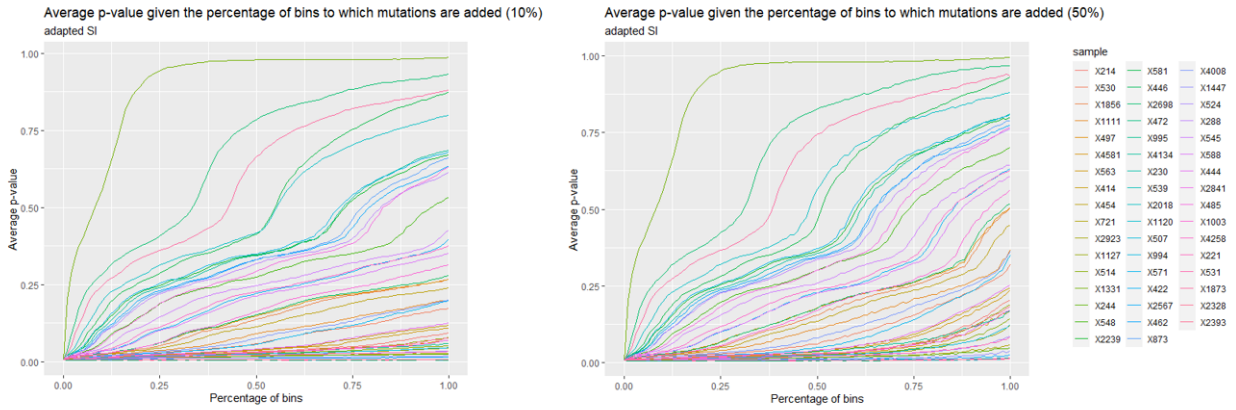


Figure 5.16: Evolution of the average p-values of the 50 artificial pairs for the adapted SI method for the first (left panel) and second (right panel) setting.

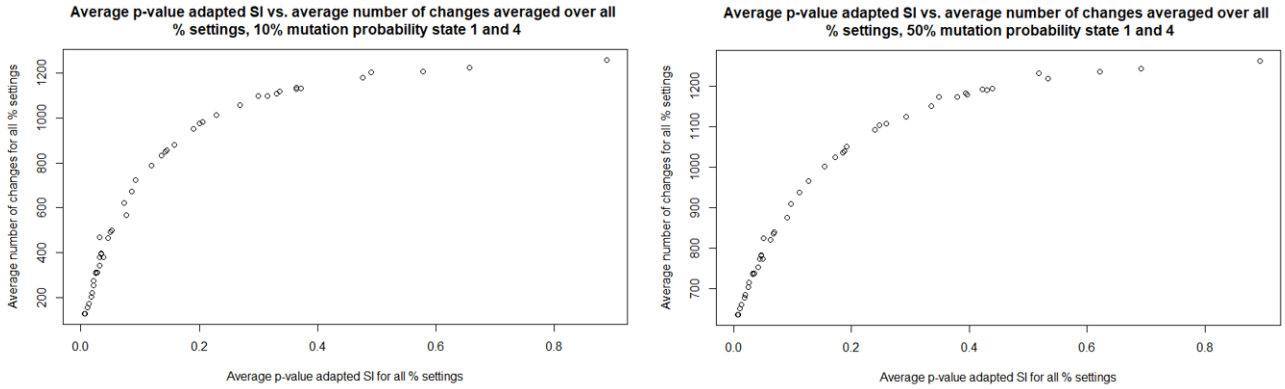


Figure 5.17: Average p-value of the adapted SI method averaged over all percentages of bins versus the average number of changes averaged over all percentages of bins of the 50 artificial pairs for the first (left panel) and second (right panel) setting.

Figure 5.17 displays the relationship between the average p-value of the adapted SI method averaged over all percentages x and the average of the average number of changes applied to the bins in all percentages x for the two different settings. When gains and losses are changed with probability 10%, Pearson's correlation coefficient shows a correlation of 0.8568 with a p-value of $2.076 \cdot 10^{-15}$. In the second setting, Pearson's correlation coefficient is estimated at 0.8969, with a corresponding p-value below $2.2 \cdot 10^{-16}$, which is the lower bound for p-values in R. However, looking at the plots in Figure 5.17 more closely, a clear logarithmic relation can be seen between the average p-value and average number of changes. Until the overall average p-value equals 0.1, a linear relationship can be seen in the plots. These points correspond to the samples consisting of many gains as these generally undergo the least number of changes and thus have the lowest p-values. After this point, the curve becomes less steep and starts taking a logarithmic form.

In contrast to the Log LR, a significant positive correlation was found between the average p-value of the adapted SI and the average number of changes averaged over all percentages of bins in both settings. Moreover, the artificial pairs with the highest p-values correspond to the samples having many normal states. In conclusion, the outcome of the adapted SI seems to be insensitive to the way that mutations are added: samples with many normal states have the highest p-values and samples with many gains the lowest, irrespective of the probability that gains and losses change.

5.3.4. Conclusion

The main objective of this section is to investigate the behavior of the comparison methods by means of simulating artificial clonal pairs. A specific workflow was constructed with which mutations would be added and two different settings of the workflow were examined. For each comparison method and setting, the average p-values and the relationship of the average p-values with the average number of bin changes were explored. Note that the way in which the mutations were added to the samples do not necessarily reflect how clonal secondary tumors arise in reality. Nevertheless, the constructed artificial pairs still give a good overview of the workings of the comparison methods.

For the Log LR method, it was found that samples with many gain and loss states have the lowest p-values when losses and gains are changed with probability 10% due to the fact that gains and losses change with a smaller probability than the other two states. However, when the change probability of gains and losses are increased to 50%, samples with many gains suddenly have the highest p-values. Samples with many losses, on the other hand, still have the lowest p-values on average even though many changes occur in the profiles due to the fact that there are also many normal bins. In general, the Log LR assigns a smaller score to pairs which have many concordant events that are rare to occur. The rarity of an aberrant event thus has a large impact on the Log LR method. Furthermore, the positive relationship between the average p-value and average number of changes, which was present in the first setting, disappears for the second setting.

The adapted SI method has the highest p-values for samples which consist of many normal states regardless of the probability with which gains and losses are changed. Adding mutations to the profiles of these samples gives many discordant aberrant bins which decrease the score quite drastically even when the number of bins taken into account for adding mutations is relatively small. Samples which have many gains, on the other hand, have the lowest p-values since the number of concordant aberrant bins remains quite large. This is also partially due to the fact that samples with many gains do not have many normal states which increase the number of discordant aberrant events within an artificial pair. Unlike the Log LR, a positive correlation between the average p-value and average number of bins remained present. The adapted SI is thus not sensitive to the change probability of gains and losses. However, even though the correlation is positive, the relationship between the two variables does not look entirely linear but more like a logarithm.

In conclusion, it can be stated that there are overlapping characteristics between the two comparison methods, but each comparison method also has its unique features. When the gains and losses change with probability 10%, the average p-value and average number of changes are positively correlated for both comparison methods. This entails that with an increasing number of changes it becomes less likely that the methods will consider the pair of samples as clonal. However, when the change

probability is increased to 50%, the positive correlation disappears for the Log LR, while it remains present for the adapted SI. Samples with many losses have a relatively low p-value for the Log LR as losses are rare to occur, so clonality is quickly assigned to such a pair. On the other hand, the adapted SI assigns a relatively high p-value to samples with many losses as these samples also have relatively many normal states that upon adding mutations, decreases the adapted SI score. Hence, samples with many losses will be called clonal by the Log LR method but not clonal by the adaptive SI method. The reverse is true for samples with many gains. The Log LR assigns relatively high p-values to samples with many gains since gains are quite common to occur so that a concordant gain does not contribute that much to the clonality hypothesis as a concordant loss or CN-LOH event. For the adapted SI, samples with many gains have a relatively low p-value as these samples do not have many normal states.

6

Results fresh frozen pairs

This chapter examines how the segmentation algorithms and comparison methods introduced in Chapters 3 and 4 perform on 23 contralateral breast cancer pairs whose tumor tissue was preserved as fresh frozen material. Section 6.1 gives an overview of the characteristics of the 23 fresh frozen pairs and gives more information about how the data was obtained and preprocessed. Sections 6.2 and 6.3 present the results of the comparison methods where ASCAT and OncoSNP were used as segmentation algorithms respectively. In Section 6.4, the influence of the choice of the independence distribution on the comparison results are investigated. Given the comparison results, Section 6.5, explores the possibility of combining the comparison results to come to a final verdict for each pair. For this, both expert judgment as well as a general approach are suggested. The chapter concludes with Section 6.6 in which a prototype of the final model is described.

6.1. Data description

The data consists of 23 fresh frozen (abbreviated to FF) contralateral breast tumor pairs. It should be noted that there is one patient (encoded as FF-31 by the lab) who has had three tumors. For this patient, two pairs can be formed which are encoded as FF-31A and FF-31B respectively. The total dataset consists of 45 breast tumors.

The surgery dates of the 45 tumors range from 1986 until 2008. Looking at the time interval between the two tumors from a patient, 11 pairs are synchronous meaning that the secondary tumor was discovered less than six months after the first tumor was diagnosed. Of these 11 pairs, seven pairs had the same date of surgery. For the other 12 pairs, which are metachronous, the median time between the diagnosis of the primary and secondary tumor is 861 days (approximately 2.3 years). For each tumor, the percentage of tumor cells present in the fresh frozen material is estimated by the lab. The tumor cell percentage estimates range between 30 and 90%, with a median value of 63%.

The Estrogen Receptor (ER) status (positive or negative) was also determined for each tumor. As explained in Section 1.2, if the ER status of a tumor is positive the tumor cells depend on the hormone estrogen to grow. Two methods were used to assess the ER status: ELISA (enzyme-linked immunoassay) and immunohistochemistry (IHC) which is performed in a pathology lab. If there is a discordance between the outcome of the ELISA and the IHC test, the IHC test result will determine the final ER status of the tumor. For five tumors, the ER status was unknown as both ELISA and IHC results were missing. As a result, the concordance in ER status could only be determined for 18 out of the 23 pairs. Of these 18 pairs, only 2 pairs had a discordant ER status. For the other 16 pairs, 12 pairs were found to be concordant ER positive, while only four pairs were concordant ER negative.

The SNP array used to investigate the DNA of the tumors is the extended Illumina Infinium GSA V3 chip containing 730,059 SNPs. The raw data resulting from the SNP chip is preprocessed using the approach described in Section 5.1.1. After excluding the SNPs on the Y and mitochondrial chromosome, SNPs that have a double genotype and SNPs which have an unknown LRR or BAF for at least

one sample, a total of 695,670 SNPs (95.3% of all the SNPs on the chip) remain and are given as input to the two segmentation algorithms.

For each segmentation algorithm output, copy number two is used as a baseline to determine if a SNP belongs to a gain, loss, CN-LOH or normal state. Once each SNP is assigned to one of the four states, the SNPs are binned together using the distance based bin approach described in Section 4.1. Since the SNP chip is the same as in Chapter 5, the distance based bins which are employed are the same as in Figure 5.1. After removing the bins which consist of less than 10 SNPs, 2828 bins are considered in the comparison methods.

6.2. Results using ASCAT

This section presents the results of the model when ASCAT is being used as the segmentation algorithm for the raw data. In the first subsection, the distribution of the 45 samples are discussed. The second subsection examines the performance of the comparison methods when the segmentation profiles are fitted by ASCAT and moreover shines a light on a problem which may occur when the data is fitted by ASCAT. The last two subsections describe possible solutions which may be used to resolve the problem.

6.2.1. Distribution of the data

The segmented profiles of the 45 samples all have a goodness of fit above 80% (range 80.49-99.41%). Hence, the results of the comparison methods can be deemed as reliable. If, for instance, one of the two samples in a pair has a goodness of fit below 80%, the reliability of the outcome of the comparison methods may be questioned as it is uncertain whether the profile of the sample with the lower goodness of fit is correct.

Figure 6.1 shows the ploidy numbers of the 45 tumor samples estimated by ASCAT. Similar to the 50 sporadic breast cancer patients investigated in Chapter 5, the ploidy numbers have a long right tail. Of the 45 samples, 18 have an estimated ploidy greater than 2.5.

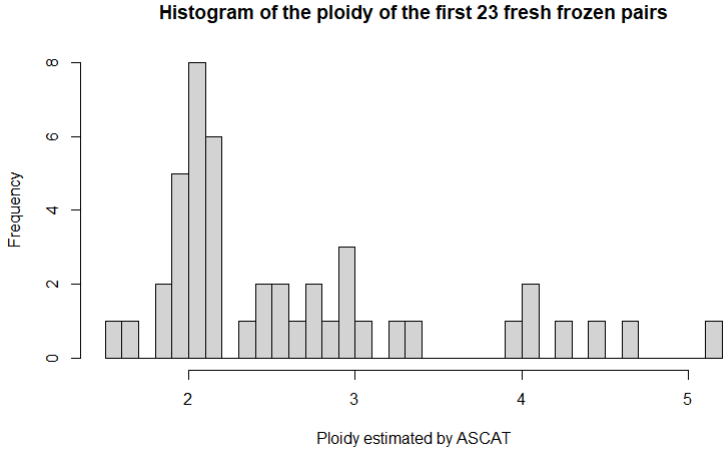


Figure 6.1: Ploidy estimated by ASCAT for the 45 tumor samples.

Since the ploidy is skewed to the right, copy numbers above two are more prevalent than copy numbers below two. Consequently, as copy number two is chosen as a baseline, gains are more common to occur than losses. The dominance of the gain states are clearly reflected in Figure 6.2 which shows the histograms of the percentages of bins that are a loss, normal, CN-LOH and gain state in all samples. There are 5 samples which have profiles that solely consist of gains, i.e. the distributions of these samples are identical. Of these 5 samples, none are coming from the same patient. Moreover, losses and CN-LOH states appear to be equally rare, even though some samples consist of fairly many CN-LOH states.

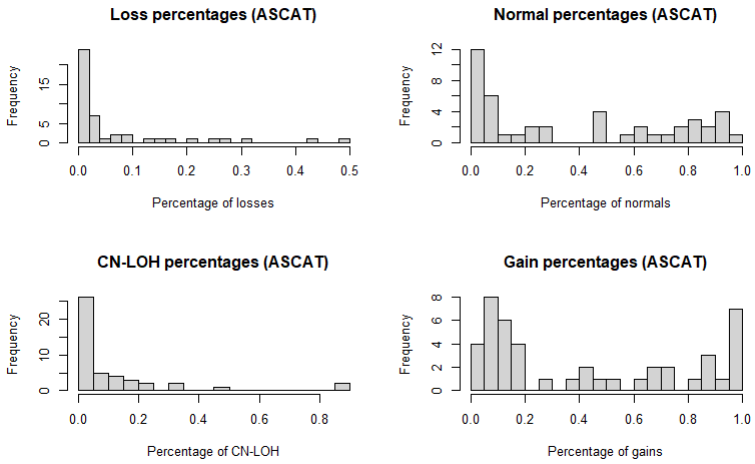


Figure 6.2: Histograms of loss, normal, CN-LOH and gain percentages of the 45 samples fitted by ASCAT.

Next to the goodness of fit values and ploidy, ASCAT also estimates the tumor cell percentage for each sample. Comparing the ASCAT estimates to the lab estimates, which were done by a pathology-trained technician, the lab estimates for the tumor cell percentage were larger than the ASCAT estimate for 27 samples. Moreover, for 20 out of these 27 samples, the difference in the estimate was more than 10%. However, as the lab is also an estimate based on visual inspection of tissue cell slices, it is uncertain which of the two estimates the tumor cell percentage correctly.

6.2.2. Comparison results

As explained in Chapter 4, the significance of the comparison scores is determined by means of constructing a reference distribution assuming independence. Given a significance level α, the null hypothesis assuming independence is rejected if the corresponding p-value of the pair is below the significance level α. In this report, the significance level α is set equal to 0.05. A pair is considered to be on the border of being clonal if the p-value is in between the interval $(0.05, 0.1)$.

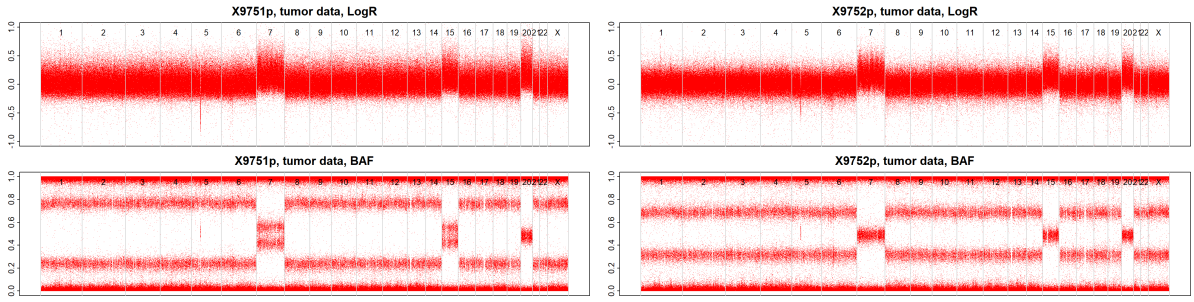
In order to validate the output of the comparison methods, two experts from the Erasmus MC in Rotterdam, who are experienced in interpreting SNP array data, have examined the raw data profiles of the 23 pairs. For each pair, the experts have tried to determine whether the pair is clonal or not based on the patterns of the genomic profiles of the two tumors. One of the characteristics the experts have looked at are the regions having very high and low LRR values. When a region has very low LRR values, a double loss has occurred in this region. If the primary tumor has copy number zero in a certain region, the same copy number zero region must also be present in a clonal tumor as a double loss can not be recovered. Once all the genetic information is lost, it will remain lost. Therefore, if the primary tumor has very low LRR values but the secondary tumor has relatively normal LRR values, it is likely that the two tumors are not of clonal origin. A similar approach is used for very high LRR values which indicate high copy numbers. As a high copy number might be beneficial for a tumor, concordant high peaks in the LRR data of the primary and secondary tumor are expected when the two tumors are of clonal origin. Of the 23 fresh frozen pairs, 6 pairs were labeled as clonal and 17 pairs were labeled as independent by the experts. Although the true clonal status is not known, the expert-calls were used as a surrogate ground-truth in the results below.

Table 6.1 shows the results of the comparison methods when ASCAT is used as segmentation algorithm. If the pair is colored green in the table, the pair was labeled as clonal by the experts. A red pair, on the other hand, implies that the pair is judged as independent by the experts. The fourth column in Table 6.1 shows how many clonal (c) and independent (i) labeled pairs had a p-value higher than 0.1. For example, 3 clonal pairs and 14 independent pairs had a p-value above 0.1 for the Log LR.

Comparison method	Clonal	p-value $\in (0.05, 0.1)$	p-value > 0.1 (c/i)
Log LR	FF-10, FF-20, FF-31A	FF-6, FF-19, FF-35	3/14
SI	FF-20	FF-31A	4/17
Adapted SI	FF-20	FF-10, FF-31A	3/17

Table 6.1: Results of the comparison methods when ASCAT is used as a segmentation algorithm with green and red indicated pairs labeled as clonal/independent by expert-judgment.

The Log LR method detects 3 out of 6 pairs correctly as clonal: FF-10, FF-20 and FF-31A. The raw data profiles of these 3 pairs show a clear clonality pattern. An example can be seen in Figure 6.3.

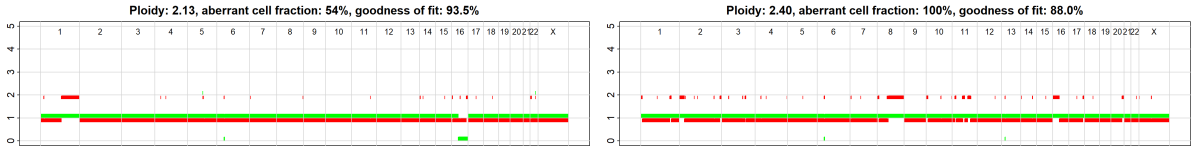


(a) Raw data profile (LRR and BAF values) of the primary tumor of pair FF-20.

(b) Raw data profile (LRR and BAF values) of the secondary tumor of pair FF-20.

Figure 6.3: Raw data plots of the two tumors of pair FF-20.

Each raw data plot consists of two panels: the upper panel corresponds to the LRR values, the lower panel to the BAF values. Pair FF-20 is correctly labeled as clonal by the adapted SI. Pairs FF-10 and FF-31A, on the other hand, only have a borderline significant p-value for the adapted SI. Next to that, there are 3 pairs, FF-6, FF-19 and FF-35, which have borderline significant p-values for the Log LR but are judged as independent by the experts. The probable reason why these samples are labeled as close to clonal for the Log LR may be that the samples within the 3 pairs consist of many concordant normal states. Since the number of concordant bins for these samples is high, the Log LR yields a relatively low p-value for these pairs. However, the number of concordant aberrant bins is relatively low so that these pairs do not have a low p-value for the adapted SI. Figure 6.4 shows the ASCAT profiles of the primary and secondary tumor of pair FF-6.

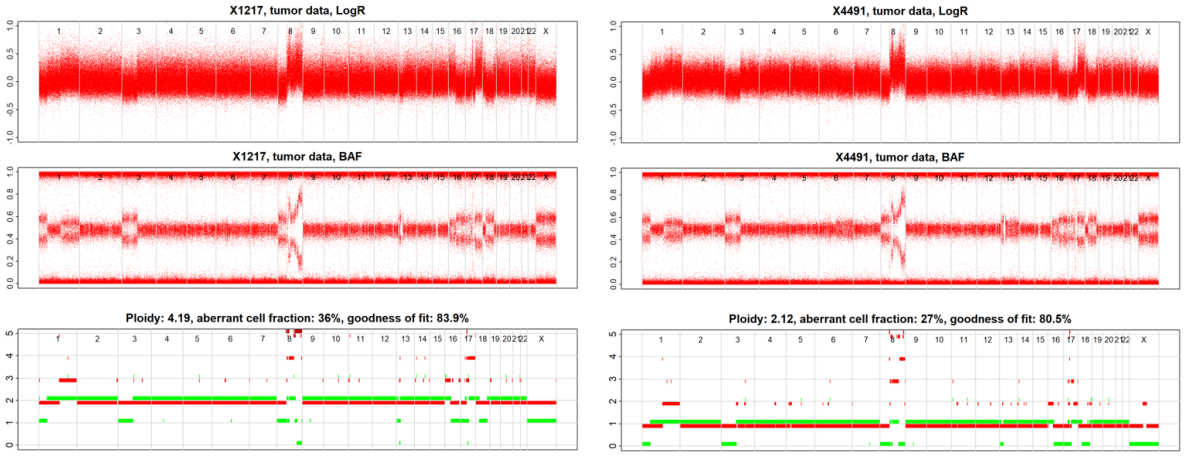


(a) ASCAT profile of the primary tumor of pair FF-6.

(b) ASCAT profile of the secondary tumor of pair FF-6.

Figure 6.4: ASCAT profiles of the primary and secondary tumor of pair FF-6.

The results of the comparison methods where ASCAT is used as segmentation algorithm show that only 3 out of 6 clonal pairs are correctly detected as clonal. Investigating the raw data and corresponding ASCAT profiles of the three clonal pairs that are not detected as clonal by any of the comparison methods more thoroughly it was discovered that ASCAT sometimes incorrectly estimates the ploidy for one of the two samples. As a result, the pair is not classified as clonal by the comparison methods even though the raw data profiles are similar. This occurred to pair FF-25, see Figure 6.5.



(a) Raw data profile (LRR and BAF values) and corresponding ASCAT profile of the primary tumor of pair FF-25.

(b) Raw data profile (LRR and BAF values) and corresponding ASCAT profile of the secondary tumor of pair FF-25.

Figure 6.5: Raw data and ASCAT profiles of the primary and secondary tumor of pair FF-25.

The raw data profiles of pair FF-25 look very similar, but the ploidy estimate of the primary tumor is almost two times larger than the ploidy estimate of the secondary tumor. As a result, the ASCAT profile of the primary tumor will show almost the same pattern as the ASCAT profile of the secondary tumor, but the primary profile has shifted upward with copy number two. This can be seen in Figure 6.5. The ploidy of the secondary tumor (right panel) is around 2, so that the red and green line (corresponding to the maximum and minimum copy number respectively) mainly lie around copy number one. For the primary tumor (left panel), which has an estimated ploidy of 4.19, the red and green line are both at copy number two, but the pattern of the profile is extremely similar to the profile of the secondary tumor. When copy number two is used as a baseline, the primary tumor consists mainly of gains, while the secondary tumor consists of mostly normal states. As a consequence, the comparison methods do not detect that the pair is clonal even though the raw data shows otherwise. The next two subsections provide possible solutions which can overcome this problem.

6.2.3. Baseline corrections

The first solution which can be applied is a baseline correction. Instead of taking copy number two as a baseline for all samples, each sample has a baseline equal to its estimated ploidy. Note that by using the ploidy of each sample as a baseline, the influence of aneuploidy is not taken into account in the output of the segmentation algorithm. Given the output of the raw or rounded copy numbers, the state of a segment is determined based on whether the sum of the raw or rounded copy numbers are within or outside of a defined interval around the estimated ploidy. Three different baseline corrections are considered:

1. The segment is defined as a gain when the sum of the raw copy numbers is 0.5 above the estimated ploidy and a loss when the sum of the raw copy numbers is 0.5 below the estimated ploidy. When the sum of the raw copy numbers is within the interval around the estimated ploidy, the segment is considered to be normal if the rounded copy numbers of both alleles are equal to one another. Otherwise, the segment is defined as CN-LOH.
2. Same as the first baseline correction, but the margin is increased to 0.6.
3. Instead of the raw copy numbers, the rounded copy numbers and a margin of 0.6 around the estimated ploidy is being used similar as in [48]. A gain and loss is defined when the sum of the rounded copy numbers are 0.6 above or below the estimated ploidy. When the sum of the rounded copy numbers falls within the interval defined around the estimated ploidy, a normal state occurs when the rounded copy numbers of both alleles are equal to one another. If this is not the case, the segment is considered to be CN-LOH.

The segmentation output of each sample is re-evaluated using all three baseline corrections. Tables 6.2, 6.3 and 6.4 show the output of the comparison methods when the samples are corrected using the

three baseline corrections. A green pair indicates that the pair is judged as clonal by the experts, while a red pair implies that the pair is seen as independent by the experts. The last column of each table states how many clonal and independent pairs had a p-value higher than 0.1.

Comparison method	Clonal	p-value $\in (0.05, 0.1)$	p-value > 0.1 (c/i)
Log LR	FF-6, FF-19, FF-20, FF-25, FF-35	FF-31A	3/14
SI	FF-2, FF-11, FF-20, FF-25, FF-29, FF-35	FF-27	4/12
Adapted SI	FF-2, FF-11, FF-20, FF-25, FF-27, FF-35	FF-10, FF-29	3/12

Table 6.2: Results of the comparison methods when the first baseline correction is used for ASCAT.

Comparison method	Clonal	p-value $\in (0.05, 0.1)$	p-value > 0.1 (c/i)
Log LR	FF-6, FF-19, FF-20, FF-35	FF-10, FF-25, FF-30, FF-31A	2/13
SI	FF-2, FF-10, FF-11, FF-20, FF-27, FF-31A	FF-25, FF-29	2/13
Adapted SI	FF-2, FF-10, FF-11, FF-20, FF-27, FF-31A	FF-25, FF-29	2/13

Table 6.3: Results of the comparison methods when the second baseline correction is used for ASCAT.

Comparison method	Clonal	p-value $\in (0.05, 0.1)$	p-value > 0.1 (c/i)
Log LR	FF-6, FF-19, FF-20, FF-25, FF-30, FF-35	FF-31A, FF-36	3/12
SI	FF-2, FF-20, FF-25, FF-29, FF-31A	FF-11	3/14
Adapted SI	FF-2, FF-20, FF-25, FF-29, FF-31A	FF-11	3/14

Table 6.4: Results of the comparison methods when the third baseline correction is used for ASCAT.

Using the first and third baseline correction, pair FF-25 is now correctly seen as clonal by the comparison methods. For the second baseline correction, pair FF-25 has a borderline significant p-value. However, when including baseline corrections it can be seen that pairs FF-10 and FF-31A are sometimes not detected as clonal anymore when the ploidy is used as a baseline. The reason why these pairs are not considered clonal anymore has to do with rounding errors. For each sample the ploidy is used as a baseline. The state of each segment is determined by means of looking whether the sum of the raw or rounded copy numbers lie above, below or within a certain interval around the ploidy. As the copy numbers and estimated ploidy are slightly different for each sample, the sum of the copy numbers of both samples may lead to different states. For this, consider a pair that has similar copy numbers and a shared segment, i.e. the boundaries of the two segments are similar. The sum of the copy numbers for one sample may lie slightly below the interval defined around the ploidy, while for the other sample the sum of the copy numbers fall within the interval. As a consequence, the two segments will be assigned to different states even though the boundaries and copy numbers are almost identical. Since the states of the SNPs are binned, a discrepancy in the segment states also leads to a higher number of discordant bins which in result causes a lower comparison score and consequently higher p-value. Lastly, the results after baseline correction also show many pairs labeled as independent now being scored as clonal by the comparison methods.

In conclusion, using the ploidy as a baseline is not the most ideal solution to identify pairs with a similar profile that are erroneously identified as independent due to incorrect estimation of the ploidy as the sensitivity to detect truly clonal pairs decreases.

6.2.4. Correcting the segmentation output

Instead of using a baseline correction, the segmentation output of one of the two samples can be modified if the absolute difference in the ploidy numbers of a pair is greater than a certain threshold. As can be seen in Figure 6.5, the difference in ploidy numbers is approximately equal to 2. Based on this, a correction is applied to the segmentation output of a pair if the difference in ploidy numbers is greater than 1.5. The following correction steps are taken:

- For each pair having an absolute difference in estimated ploidy greater than 1.5, the segmentation output of the sample with the largest ploidy in the pair is corrected while leaving the other sample constant. In the segmentation frame of the sample with the largest ploidy, the following is done:
  - If both the major (red line) and minor (green line) copy number are greater than zero, one is subtracted from both.
  - If the major copy number is greater than zero and the minor copy number is equal to zero, two is subtracted from the major copy number.
  - If both major and minor copy number are equal to zero (homozygous deletion), the copy numbers remain zero.

Once the pairs are corrected, the states are determined for all pairs using copy number two as a baseline. After this, the states are binned and given as input to the comparison methods. Of the 23 pairs, 7 pairs had an absolute ploidy number difference greater than 1.5 and were corrected. Table 6.5 shows the results of the comparison methods.

Comparison method	Clonal	p-value $\in (0.05, 0.1)$	p-value > 0.1 (c/i)
Log LR	FF-6, FF-10, FF-19, FF-20, FF-25, FF-31A, FF-35	FF-30	2/13
SI	FF-2, FF-10, FF-11, FF-20, FF-25, FF-31A	FF-29, FF-32	2/13
Adapted SI	FF-2, FF-10, FF-11, FF-20, FF-25, FF-31A, FF-32	FF-29	2/13

Table 6.5: Results of the comparison methods after correcting the ASCAT segmentation output.

In Table 6.5 it can be seen that all comparison methods detect pairs FF-10, FF-20, FF-25 and FF-31A correctly as clonal. In contrast to the baseline correction approach of Section 6.2.3, the pairs which are judged as clonal are not influenced by rounding errors. However, comparing Table 6.5 with Table 6.1 more pairs are labeled as clonal when the segmentation output is modified. An explanation for this may be that the samples with relatively many gains are corrected downwards. In the dataset, 5 samples which are part of distinct pairs, consist solely of gains. As a consequence, the comparison scores between these samples are maximal as all the states are gains, e.g. the adapted SI is equal to 1. Since there are many observations in the independence distribution pointing towards clonality, the p-values of the pairs increase. When the samples with a larger ploidy are corrected downwards, less extreme scores arise in the independence distribution which in turn yield lower p-values. Even though more false positives are introduced, the intersection of the pairs labeled as clonal by the Log LR and adapted SI do not contain any false positives.

In summary, it can be concluded that correcting the segmentation output works better than the baseline corrections of the previous section since pair FF-25 is now correctly detected as clonal without missing the other clonal pairs. However, two pairs which were judged as clonal by the experts, FF-7 and FF-31B, are still not identified as clonal after correcting the segmentation output. Evaluating the segmentation profiles of the samples in these two pairs, it became apparent that the number of concordant events was too small to be significant. In the remainder of this thesis, the segmentation output of ASCAT will be corrected using the steps described above.

6.3. Results using OncoSNP

This section presents the results of the comparison methods when OncoSNP is used as segmentation algorithm. In the first subsection, the distribution of the data is investigated including ploidy numbers and aberrant cell fraction estimates. The second subsection shows the results of the comparison methods.

6.3.1. Distribution of the data

Section 5.2.2 showed that the tumor heterogeneity mode in OncoSNP did not contribute much to the model output compared to the results of the model incorporating only normal cell contamination. Therefore, in order to reduce the computation time, the samples of the 23 pairs are fitted in OncoSNP using only the normal cell contamination mode.

For each sample, the model with the highest likelihood and rank 3 were chosen to construct the segmentation profile. The model assuming diploidy was the best fitting model for 38 out of 45 samples. One sample has the triploid model as the model with the highest likelihood and for 6 samples the results of the diploid and triploid model converged to the same values. For these samples, the ploidy number and tumor cell percentage estimates were also relatively close to the ASCAT estimates, see Table 6.6.

Sample	Lab TC%	ASCAT TC%	OncoSNP TC%	ASCAT ploidy	OncoSNP ploidy
A3779	63	71	70	2.39	3.1
1507	80	61	60	2.75	2.8
4187	90	66	70	2.84	2.9
4197	63	77	50	2.69	3.1
4210	77	75	80	2.97	3.1
4243	80	66	60	2.56	2.7

Table 6.6: Tumor cell percentage (TC%) and ploidy estimates of the 6 samples that converged in OncoSNP.

As the tumor cell percentages and ploidy numbers estimated by ASCAT and OncoSNP are relatively close to one another for most samples, it can be concluded that the raw data contains sufficient information to derive the tumor cell percentage and ploidy number.

Figure 6.6 shows the histogram of the ploidy numbers estimated by OncoSNP. The histogram again shows a skew to the right, though less extreme than the ASCAT estimates. This has to do with the fact that for most of the samples the OncoSNP model assumes that the sample is diploid. However, of the 45 samples, 15 samples still have a ploidy above 2.5. Hence, gains are more common to occur than losses if copy number two is used as a baseline (see Figure 6.7).

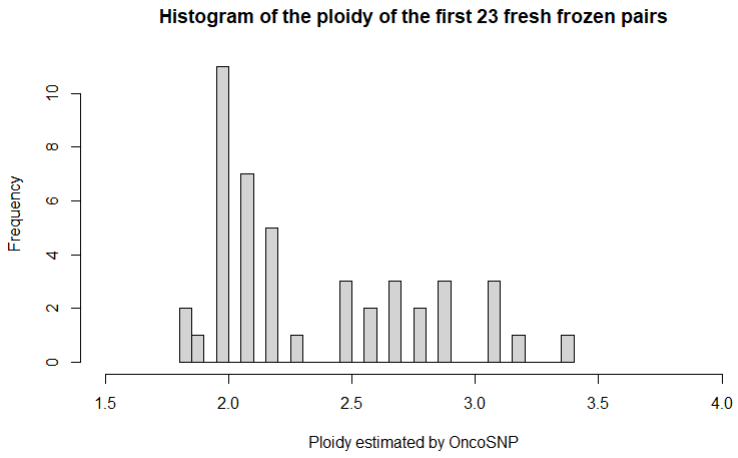


Figure 6.6: Ploidy numbers by OncoSNP for the 45 tumor samples.

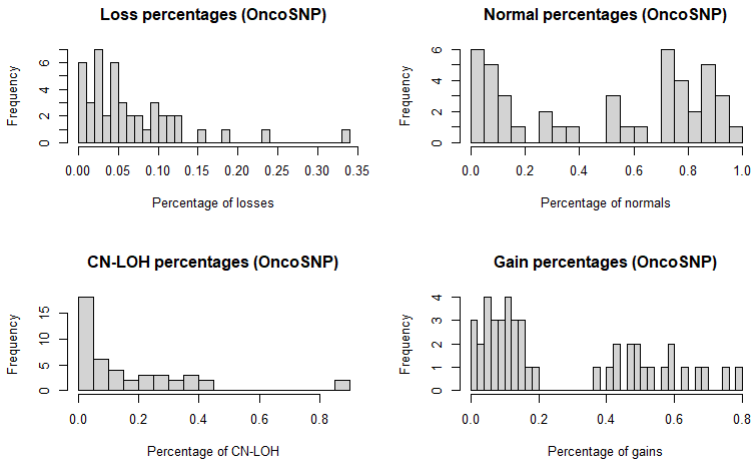


Figure 6.7: Histograms of loss, normal, CN-LOH and gain percentages of the 45 samples fitted by OncoSNP.

Comparing the distribution results of OncoSNP with those of ASCAT (see Figure 6.2), gains are less common to occur when the profiles are fitted using OncoSNP. In contrast to the results of ASCAT, there are no samples that solely consist of gains. Moreover, losses also occur more often in the segmented profiles of OncoSNP. However, gains are still more prevalent than losses.

Next to the ploidy numbers and distribution of the samples, the tumor cell percentage estimates of OncoSNP range between 10 and 90% and is generally lower than the lab estimate: only 7 samples have an OncoSNP tumor cell percentage estimate that is larger than the lab estimate. Of the other 38 samples, 30 have an OncoSNP estimate that is more than 10% smaller than the lab estimate. Next to that, the estimated tumor cell percentages of OncoSNP are also generally smaller than that of ASCAT: only 16 out of 45 samples had a higher tumor cell percentage for OncoSNP than for ASCAT. The mean estimated tumor cell percentage of ASCAT is equal to 60.57, while it is only equal to 45.11 for OncoSNP.

6.3.2. Comparison results

Table 6.7 shows the results of the comparison methods when OncoSNP is used as a segmentation algorithm and $\alpha = 0.05$ is used as significance level. A green/red color indicates that the pair is judged as clonal/independent by the experts. The last column in the table states how many clonal (c) and independent (i) labeled pairs had a p-value higher than 0.1.

Comparison method	Clonal	p-value $\in (0.05, 0.1)$	p-value > 0.1 (c/i)
Log LR	FF-6, FF-19, FF-20, FF-25, FF-31A, FF-36	FF-17	3/13
SI	FF-20, FF-25, FF-27, FF-29, FF-31A	FF-5, FF-10, FF-11	2/13
Adapted SI	FF-10, FF-11, FF-20, FF-25, FF-27, FF-29, FF-31A	FF-5, FF-36	2/12

Table 6.7: Results of the comparison methods when OncoSNP is used as segmentation algorithm.

Pairs FF-20, FF-25 and FF-31A are correctly detected as clonal by all the comparison methods. Next to that, more pairs are wrongly detected as clonal when compared to the ASCAT results of Table 6.1. This can be explained by the fact that gains are less common in the OncoSNP output. As described in Section 6.2.4, the many gains that occur in the ASCAT output causes more scores in the independence distribution to become small which in turn yields higher p-values for the pairs under investigation. Since the segmented profiles of OncoSNP have more balanced distributions, the scores in the independence distribution become less extreme resulting in lower p-values and consequently more pairs which are

detected as clonal by the comparison method.

Pair FF-10, who is judged as clonal by the experts, is only recognized as clonal by the adapted SI. Looking at the ploidy number and tumor percentage content estimates, the tumor percentage estimates are similar, but there is a 0.5 difference in the ploidy number estimate. As a result, the pair will consist of less concordant bins and will not be detected as clonal by the Log LR. This shows that the sensitivity of the comparison methods highly depends on the output of the segmentation algorithm. If the segmentation algorithm makes a mistake for one of the two samples in a pair that have similar genomic profiles, the comparison method will fail to recognize the high degree of clonality that is present within the pair.

Pairs FF-7 and FF-31B, which were judged as clonal by the experts, are again not labeled as clonal by the comparison methods when OncoSNP is used as segmentation algorithm. In order to verify the judgment of the two experts, three additional experts from the Erasmus MC were asked to judge the two discordant pairs in terms of clonality. However, the opinion of the three experts for the two pairs varied as well. Both pairs were labeled as clonal by three out of five experts. One expert stated that a final judgment for pair FF-7 was not possible as the raw data plot of the primary tumor did not provide enough information, i.e. did not show enough aberrant events, to conclude if the secondary tumor is clonal or not. Next to that, one of the two original experts changed the call to independent after re-evaluating pair FF-31B. The two discordant pairs appear to be in a grey area: the clonality status of these pairs can not be reliably determined. As a result, the two ambiguous pairs were excluded for the final evaluation of the comparison methods.

From the findings in this section and Section 6.2, it can be concluded that the performance of the proposed model depends on two factors:

1. The correctness of the output of the segmentation algorithm. If the segmentation algorithm struggles to fit one of the two samples in an apparent clonal pair accurately, the comparison method will not detect the clonal features of the pair.

2. The distribution of the samples in the pairings that are used in the independence distribution. If many samples consist mostly of a particular state, e.g. a gain, more scores in the independence distribution will be relatively small (i.e. pointing towards clonality) yielding higher p-values for the pairs that are examined. The less similar samples from different pairs are, the better the independence distribution will be and consequently the more reliable the p-values will become.

6.4. The composition of pairings in the independence distribution

For each comparison method introduced in Chapter 4, the significance of the comparison scores are determined by means of a permutation method. An independence distribution is constructed for which the comparison scores of independent pairs, consisting of a primary tumor of one pair and secondary tumor of another pair, are computed. When there are n pairs, the independence distribution consists of $n(n-1)$ observations. When there are 23 pairs, the total number of observations is equal to $23 \cdot 22 = 506$. However, in our case, the number of unique independent pairings between primary and secondary tumors which can be made is less than 506 as there are two pairs that correspond to the same patient. As a consequence, the independence distribution consists of 21 double scores and 2 scores which are not independent as they are coming from the same pair. In order to prevent double and dependent observations in the independence distribution, 483 observations are included in the independence distribution. However, the number of observations in the independence distribution can be increased to $2n(n-1)$ if all pairings (primary-primary, secondary-secondary and primary-secondary) are used. This section investigates how including all the pairings in the independence distribution affects the results of the comparison methods.

6.4.1. Downward bias

According to [31], secondary tumors from two different patients should not be paired as this may result in a possible downward biased reference value in the independence distribution. [20] and [23] have shown that secondary tumors frequently possess distinct aberrant events compared to their primary tu-

mors, which are possibly caused by radiation therapy and chemotherapy given between the two events. In other words, if the patient receives radiation therapy or chemotherapy after surgery of the primary tumor, the secondary tumor, if clonally derived from the primary tumor, may encompass additional gain and loss events. If the secondary tumors are paired, the comparison scores in the independence distribution will increase as there are many different non-matching events. As a result, the p-values will be lower than they actually are, i.e. a downward bias is introduced in the p-values of the pairs. Next to the secondary tumors, [31] argues that pairings of independent primary tumors should also not be included in the independence distribution, but a clear reason for this has not been given by [31].

The influence of radiation therapy and chemotherapy on the resulting p-values are investigated using the 23 fresh frozen pairs. Of the 23 patients, 11 patients received some kind of therapy between the primary and secondary tumor: 10 patients received radiation therapy, six patients were given chemotherapy and two patients received anti-hormonal therapy. As almost half of the fresh frozen pairs either received radiation therapy or chemotherapy, it is of interest to see how the resulting p-values are affected by the choice of the independence distribution.

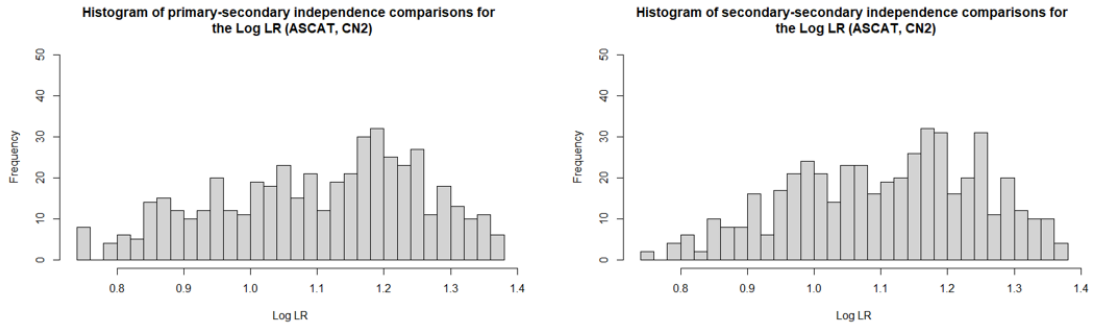


Figure 6.8: Histograms of the independence scores of the primary-secondary (left panel) and secondary-secondary (right panel) pairings for the Log LR, where the profiles are fitted by ASCAT using copy number two as a baseline.

The significance of the scores of the comparison methods are computed using three different independence distributions: the secondary-secondary pairings, the primary-primary pairings and the primary-secondary pairings. Figure 6.8 shows the histograms of the scores of the primary-secondary and secondary-secondary pairings for the Log LR where the profiles are fitted by ASCAT using copy number two as a baseline. Looking at the histograms, smaller Log LR scores are present in the primary-secondary pairings. When the primary-secondary pairings are used, only three pairs are labeled as clonal, while for the secondary-secondary pairings six pairs have a p-value below 0.05. The three extra pairs which are labeled as clonal when the secondary-secondary pairings are all judged as independent by the experts. The fact that more pairs are labeled as clonal is a sign that there is indeed a downward bias present in the p-values when the secondary-secondary pairings are used. Similar results were found for the other comparison method, also in combination with OncoSNP.

However, if all the pairings are used together in the independence distribution (primary-secondary, secondary-secondary and primary-primary), no differences are found in the results of the comparison methods. This has to do with the fact that if the primary-primary pairings are used in the independence distribution, the resulting p-values are higher than the p-values where the primary-secondary pairings are employed as the independence distribution. If all pairings are put together, the primary-primary and secondary-secondary pairings balance each other out yielding similar results.

For this dataset, the primary-primary pairings mitigate the downward bias caused by the secondary-secondary pairings as the results of all the pairings together are similar to the output of the comparison methods when only the primary-secondary pairings are used in the independence distribution. However, the primary-primary pairings may not always balance out the downward bias effect of the secondary-secondary pairings. The mitigating effect occurs for this dataset, but it is unknown whether this generally occurs for all datasets. Therefore, it is advised to only use the comparison scores of the

primary-secondary pairings in the independence distribution.

6.5. Combining the results of the comparison methods

The goal of the model is to determine if a secondary tumor is a metastasis of the primary tumor or a new independent tumor. For this, each comparison method in the model calculates a p-value for each pair. As explained in Section 4.2, the p-value of a pair is derived by means of comparing the comparison score of the pair to the scores in the independence distribution. The p-value is defined as the proportion of observations in the independence distribution which are larger or smaller (depending on the comparison method) than the comparison score. In other words, the p-value of a pair resembles the probability that the two tumors are of independent origin. If the p-value is below the significance level α, the null hypothesis assuming independence is rejected and the pair is considered to be clonal. For the 23 fresh frozen pairs, a significance level of $\alpha = 0.05$ was chosen. Moreover, if the p-value of a pair is within the range $(0.05, 0.1)$ there is evidence that the pair might be clonal. As two comparison methods are employed in the model, each pair gets assigned two p-values that may differ in terms of significance. For instance, a pair may be labeled as clonal by one comparison method, while the other comparison method states that the pair is not clonal.

In the most ideal setting, the oncologist gets to see only one judgment stating that the pair is clonal or not clonal. In order to realize this, the p-values of the comparison methods may be combined into one single p-value reflecting the overall probability that the pair is independent. By combining the p-values, the discordance problem of the comparison methods is solved. Note that the segmentation algorithm is assumed to be fixed when combining the p-values, i.e. the segmentation profiles on which the comparison methods are applied are constructed by either ASCAT or OncoSNP.

Several methods exist in literature that are capable to combine multiple p-values into one summarizing p-value. The most well-known method is Fisher's method [16], which aggregates the p-values of K independent hypothesis tests into a single test statistic χ_F^2 as follows:

$$\chi_F^2 = -2 \sum_{i=1}^{K} \ln p_i,$$

where p_i is the p-value of the i-th hypothesis test and $\ln$ is the natural logarithm. If all the null hypotheses are true, the test statistic χ_F^2 will follow a χ^2 distribution with $2K$ degrees of freedom. The goal of the hypothesis test, according to Fisher, is to discover whether accumulation of information among tests on similar null hypotheses can reject the shared null hypothesis [54]. Next to Fisher's method, there are also other methods that can combine p-values such as the Z-transform test [44] and the weighted Z-method [27, 30]. However, a downside of all the methods mentioned so far is that the p-values are assumed to be independent, i.e. the tests are conducted on different datasets. In this research, each comparison method is applied on the same dataset which makes the p-values dependent. A positive correlation is mostly present between the p-values of the comparison methods: if a pair is similar, the p-values will be low for (most of) the comparison methods. As the p-values are dependent, a different combining function needs to be employed.

[51] gives an overview of several combining functions that can be used to merge dependent p-values into one p-value. In this research the weighted harmonic mean p-value, introduced by [55], will be employed. The weighted harmonic mean p-value, $\mathring{p}$, is defined as follows:

$$\mathring{p} = \frac{\sum_{i=1}^{K} w_i}{\sum_{i=1}^{K} \frac{w_i}{p_i}},$$

where K is the number of p-values, w_i is the weighted of the i-th p-value, p_i, and $\sum_{i=1}^{K} w_i = 1$. If the weights are chosen as $w_i = \frac{1}{K}$ for $i = 1, \ldots, K$, the unweighted harmonic mean p-value arises. The significance of the weighted harmonic mean p-value can be asymptotically determined using the Landau distribution. Given the value of $\mathring{p}$, the corresponding p-value of the weighted harmonic mean

p-value, $p_{\hat{p}}$, can be calculated as:

$$p_{\hat{p}} = \int_{\frac{1}{\hat{p}}}^{\infty} f_{\text{Landau}}\left(x|\log K + 0.874, \frac{\pi}{2}\right) dx,$$

where K is the number of p-values and the density function of the Landau distribution is defined as:

$$f_{\text{Landau}}(x|\mu,\sigma) = \frac{1}{\pi\sigma}\int_0^{\infty} e^{-t\frac{x-\mu}{\sigma}-\frac{2}{\pi}t\log t}\sin(2t)dt.$$

The p-value $p_{\hat{p}}$ becomes exact for large K. In order to use the weighted harmonic mean p-value, weights need to be determined for each comparison method. The next two subsections explain two different ways in which these weights can be established: via expert judgment or by means of using a power curve simulation.

6.5.1. Expert judgment

A first way to determine the weights of the comparison methods is by means of using expert judgment. Given the judgments of the experts, the weight of each comparison method can be determined by means of looking at how many pairs the comparison method has correctly classified. Different scoring functions can be applied for this purpose. For instance, a pair can be assigned score 1 if the comparison method and expert judgment agree (both state that the pair is clonal or both state the pair is not clonal) and -1 if there is disagreement. The total score for a comparison method is equal to the sum of all the individual scores of the pairs. If, for instance, 18 pairs are correctly labeled by a comparison method and 5 are incorrectly labeled, the total score will be equal to $18 - 5 = 13$. An overview of the different scoring functions which were investigated in this thesis can be found in Appendix C.

Given a scoring function and segmentation algorithm, the results of the comparison methods can be evaluated using expert opinion: the larger the total score of a comparison method, the larger the weight becomes in the weighted harmonic mean p-value. In other words, if s_i is the total score for comparison method i and if there are K comparison methods taken into consideration in the analysis, the weight for comparison method i will be equal to:

$$w_i = \frac{s_i}{\sum_{j=1}^{K} s_j}.$$

For example, if the scoring function introduced above is applied on the 23 fresh frozen pairs (i.e. a score of 1 is applied when the pair is correctly classified and -1 otherwise) where the samples are segmented using OncoSNP as segmentation algorithm, a total score of 10 was found for the Log LR and a total score of 13 for the adapted SI. Given these total scores, the weight for each comparison method can be determined as follows:

$$w_{\text{Log LR}} = \frac{10}{10+13} = \frac{10}{23},\ w_{\text{adapted SI}} = \frac{13}{10+13} = \frac{13}{23}.$$

In this case, the p-values of the Log LR get assigned a weight of $\frac{10}{23}$, while the p-values of the adapted SI get a weight equal to $\frac{13}{23}$. When OncoSNP is used to segment the profiles, the adapted SI thus outperforms the Log LR in terms of expert judgment. Note that the two ambiguous pairs, FF-7 and FF-31B, are included in the computation of the total scores. It is also possible to compute the total scores only considering the pairs for which the experts were certain.

Next to determining the weights, a scoring function may also be utilized to discover the best performing comparison method. Note that this corresponds to the situation where one comparison method is assigned weight 1 in the weighted harmonic mean p-value while the other comparison methods are given weight 0. The best performing comparison method is the method with the largest total score. The best performing comparison method for the 23 pairs when OncoSNP was used as segmentation algorithm is the adapted SI.
Expert judgment can thus be used as a means to determine the weights in the weighted harmonic mean

p-value. Given a scoring function, the weights for the comparison methods can be derived by means of normalizing the total scores. However, a downside of using expert judgment is that the weights are derived in a data-driven way: the weights are proportional to the performance of the comparison methods for this dataset consisting of 23 pairs. If a different dataset would have been used, different weights may have been obtained for the comparison methods. The derived weights thus only apply for this dataset and are not valid per se for other datasets. Expert judgment can therefore not be generally applied: expert opinion is required for each new dataset in order to derive the weights. This may become quite time consuming when the number of pairs is large. The next subsection describes a possible general approach that can be used for deriving the weights of the comparison methods.

6.5.2. Power curve simulation

Instead of using expert judgment, which is a data-driven approach, a general method for determining the weights can be constructed using a power curve simulation. [27] and [30] already used the power of each study to determine the weights in the weighted Z-transform method. For the weighted harmonic mean p-value, the weights for the comparison methods can also be derived using the power of the comparison methods. The higher the power of the method, the higher the weight of the comparison method becomes. If P_i is the power of comparison method i and if there are K comparison methods considered, the weight for comparison method i will be equal to:

$$w_i = \frac{P_i}{\sum_{j=1}^{K} P_j}.$$

In order to determine the power of each comparison method, a power curve simulation can be conducted. The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true. The null hypothesis states that the two tumors arose independently of one another, while the alternative hypothesis declares that the two tumors are of clonal origin. For the simulation of the power curve, the set of 50 sporadic breast cancer patients of Chapter 5 will be used where the samples are fitted using ASCAT. Similar to the simulations of Chapter 5, the states of the bins will be changed creating artificial clonal pairs. These artificial clonal pairs will then be used to determine the power of each comparison method. The workflow of the power curve simulation will be as follows:

- For each percentage of bins $x \in (0, 1, 2, \dots, 99, 100)$:
  - For each sample $i \in (1, 2, \dots, 49, 50)$:
    - Mutations are added to sample i 50 times with the other samples being held constant. This gives one artificial clonal pair.
    - Of the remaining 49 samples, one sample is randomly removed yielding 48 samples in total. The remaining 48 samples are then randomly paired giving 24 independent pairs.
    - Given the artificial pair and 24 independent pairs, the comparison method scores are computed for the artificial pair and the corresponding p-value determined by means of constructing an independence distribution.
  - For each percentage of bins x and each sample i, 50 p-values are obtained for each comparison method. In order to determine the power of a comparison method for a given percentage x, the number of instances that the method correctly rejected the null hypothesis of independence can be calculated. For each percentage x, there are $50 \cdot 50 = 2500$ p-values which range from 0 to 1. Given the significance level α, the number of times the p-values are below the significance level can be computed. The significance level in the power curve simulation is chosen equal to 0.05, similar to Sections 6.2 and 6.3. If, for instance, 1800 p-values are below 0.05 for a comparison method when $x = 50$, the power at $x = 50$ will be equal to $\frac{1800}{2500} = 0.72$, i.e. with 72% probability the comparison method correctly detects that the pair is clonal and the null hypothesis of independence will be rejected.
- Plotting all the powers for all percentages x yields the power curve for a comparison method. Note that, as the null hypothesis corresponds to the two tumors being independent and the artificial pairs are simulated under the hypothesis of clonality, the power curve will be a decreasing

function. As more bins are taken into account, more changes will be applied to the bins on average resulting in higher p-values and thus less pairs that are detected as clonal. The total power of each comparison method is the surface of the curve which is in this case defined as the sum of all powers evaluated at all percentages x, i.e. the power P_i of comparison method i is equal to:

$$P_i = \sum_{x=0}^{100} P_i^x,$$

where P_i^x is the power of comparison method i at percentage x.

At first glance, the power curve simulation appears to be a suitable method to derive general weights for the comparison methods. However, a problem arises with this approach as well. In order to derive the power of a comparison method, state changes need to be applied to the bins of the samples. Section 5.3.1 describes an example of how mutations can be added to the samples. Specific assumptions about how a secondary clonal tumor may evolve from its primary tumor need to be made before the power curve simulation can be started. If the way in which state changes are applied in the simulations are not reflecting reality, the resulting weights of the power curve simulation will not be realistic as well. In the power curve simulation approach, there are actually two unknowns that accompany each other: the goal of the simulation is to discover the power of a comparison method, which is unknown, by means of adding mutations to the samples of which it is unknown how these occur in reality. Therefore, the results of the power curve simulation should be interpreted with care. Given a mutation scheme, the results of the simulations only give an indication of the power of the comparison methods if secondary clonal tumors indeed arise as described by the mutation scheme. Hence, as it is unknown how secondary clonal tumors evolve from their primary tumors in reality, a general model for combining the methods can not yet be derived.

Nevertheless, an attempt was made to construct a relatively realistic mutation scheme. This mutation scheme was then given as input to the power curve simulation procedure. The developed scheme is similar to the scheme introduced in Section 5.3.1, with slight adaptations made to the change probabilities of gains and losses. This was done to make the mutation scheme a bit more realistic. In the simulations of Chapter 5, 3 different settings regarding the change probability of a gain or loss were investigated: 10%, 50% and a varying probability, where the change probability is equal to 1 minus the probability that a gain or loss occurs in the bin coming from the Affymetrix dataset. The results of the simulations showed that when the change probability equals 50%, the positive correlation between the number of changes in the bins and the corresponding p-value disappears for the Log LR method. A change probability of 50% for gains and losses seems to be too large when the secondary tumor is of clonal origin. Therefore, the change probability of a gain and loss in the mutation scheme is chosen smaller than 50%. For a gain, the change probability is drawn from a uniform distribution on the interval $(0, 0.3)$, while the change probability for a loss comes from a uniform distribution on the interval $(0, 0.2)$. The change probability of a loss is on average slightly smaller than the change probability of a gain. A loss thus remains with a higher probability a loss than a gain remains a gain. This decision was made based on several factors. First, losses have a finite endpoint: once copy number zero is obtained, there is nothing left. Gains, on the other hand, have an infinite endpoint and can theoretically always go back to copy number two. A gain can thus always be recovered which is not the case for a loss. Moreover, if a loss occurs in the DNA during the cell division process and the loss is not detected by the DNA repair mechanisms, the loss will remain present in the DNA. During the next cell division the loss in the DNA will be regarded as ”normal” and the DNA repair mechanisms are not activated. Next to that, cells with many losses contain less genetic material that needs to be copied and are thus energetically more attractive to copy than cells having more gains.

In summary, the mutation scheme is as follows:

- If the state of the bin is a loss, the change probability is drawn from a uniform distribution on the interval $(0, 0.2)$. The average change probability is equal to 10%. Note that for each bin and each simulation, the change probability differs: a random process is mimicked. If the loss state is changed, it can only go to a CN-LOH state.

- If the state of the bin is normal, the state of the bin changes with probability $1-\mathbb{P}(\text{bin is in state 2})$, where $\mathbb{P}(\text{bin is in state 2})$ is the probability that the bin is normal taken from the Affymetrix dataset introduced in Chapter 5. If the state is changed, it can go to all three other states. The transition probabilities of going to a loss, CN-LOH or gain state are determined by the frequencies of the Affymetrix dataset where rescaling is used so that the probabilities add up to 1.
- If the state of the bin is CN-LOH, the state of the bin changes with probability $1-\mathbb{P}(\text{bin is in state 3})$, where $\mathbb{P}(\text{bin is in state 3})$ is the probability that the bin is CN-LOH taken from the Affymetrix dataset. If the state is changed, it can go to either a gain or a loss state. The transition probabilities of going to a gain or a loss are given by the rescaled frequencies of the Affymetrix dataset.
- If the state of the bin is a gain, the change probability is drawn from a uniform distribution on the interval $(0, 0.3)$. The average change probability is equal to 15%. If the gain state is changed, it can go to a normal or CN-LOH state. The transition probabilities of going to a normal or CN-LOH state are derived from the Affymetrix dataset where rescaling is used.

Note that the decision to make the change probability of a loss smaller than that of a gain is based purely on biological reasoning. Therefore, the results of the power curve simulation should be interpreted with care. The power curve simulation was performed on the same hardware as in Chapter 5 and took approximately 60 hours.

Figure 6.9 shows the power curves of both comparison methods using the mutation scheme described above. For a small number of mutations, the Log LR has a higher power while the adapted SI performs better when more mutations are added to the bins. The two power curves seem to intersect one another at 0.4, i.e. when 40% of bins are taken into account. At this point, the comparison methods have equal power. As mentioned earlier, the total power of a comparison method is defined as the sum of powers evaluated at all percentages of bins. Table 6.8 shows the total power and the resulting weights for each comparison method.

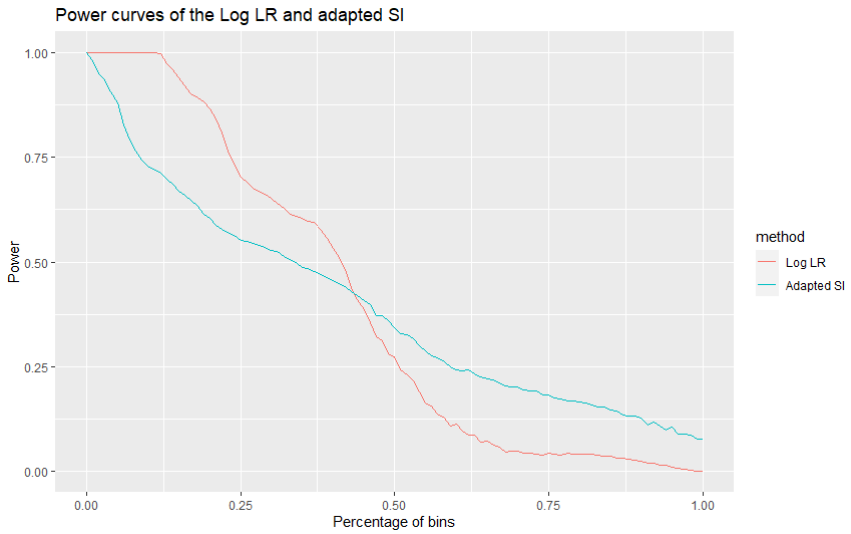


Figure 6.9: Power curve simulation results for the Log LR and adapted SI method.

Comparison method	Log LR	Adapted SI
Total power	40.4764	39.8540
Weight	0.5039	0.4961

Table 6.8: Total power and resulting weights from the power curve simulation.

According to the power curve simulations, the weights should be approximately 50/50. This can also be seen in the plot of the power curve simulations as the surface of both power curves are similar in

size. Given the weights, the p-values of the comparison methods can be combined using the weighted harmonic mean p-value. Table 6.9 shows the pairs with a significant weighted harmonic mean p-value, i.e. the p-value $p_{\mathring{p}}$ of $\mathring{p}$ was below 0.05, when OncoSNP is used as a segmentation algorithm. The coloring of the pairs again reflects whether the pair was judged as clonal (green) or independent (red) by the experts. Note that the two ambiguous pairs, FF-7 and FF-31B, are not considered in the results.

Clonal	p-value $\in (0.05, 0.1)$	p-value > 0.1 (c/i)
FF-6, FF-11, FF-19, FF-20, FF-25, FF-31A, FF-36	FF-20, FF-29	0/12

Table 6.9: Results of the weighted harmonic mean p-value when OncoSNP is used as segmentation algorithm and the weights are derived by means of the power curve simulation.

Comparing Tables 6.7 and 6.9, it can be seen that a pair can be considered significant by the weighted harmonic mean p-value when only one of the two comparison methods has a significant p-value for a pair. For instance, pair FF-11 is considered significant by the weighted harmonic mean p-value, while the Log LR does not label this pair as clonal. However, looking more closely at the individual p-values an undesirable effect seems to be present in the weighted harmonic mean p-value. For example, pair FF-11, which is not considered as clonal by the experts, has a combined p-value of 0.0198. The individual p-values of this pair are 0.21 for the Log LR and 0.01 for the adapted SI. Pair FF-10, which is labeled as clonal by the experts, is not labeled as clonal by the weighted harmonic mean p-value. However, looking at the individual p-values of pair FF-10, the Log LR yields a p-value of 0.13, while the adapted SI states a p-value of 0.039. On average, pair FF-10 has a lower p-value than pair FF-11, but pair FF-11 is labeled as clonal by the weighted harmonic mean p-value. Since the weights for both comparison methods are approximately equal, a very low p-value for one of the two methods results in a very low weighted harmonic mean p-value. Similar results were found for the results when ASCAT is used as a segmentation algorithm. The weighted harmonic mean p-value sharply decreases if one of the two p-values is very low even when the other p-value is quite high. As a consequence, more samples which are judged as independent by the experts are labeled as clonal using the weighted harmonic mean. This effect also works the other way around: pairs which are judged as clonal may not be detected as clonal by the weighted harmonic mean p-value when only one of the two comparison methods has a significant p-value. The suitability of the weighted harmonic mean p-value can therefore be questioned. For this objective, another merging function may prove more suitable such as the weighted mean. A downside, however, is that the significance of the combined p-values can not be explicitly derived for many merging functions including the weighted mean [51].

In conclusion, combining the comparison methods by means of using a merging function for the p-values appeared not successful yet. The weighted harmonic mean p-value is a possible candidate which can be used, but has a downside that it may introduce many false positives as the weighted harmonic mean p-value decreases if one of the two p-values is very low. Similarly, clonal pairs may also be missed when only one of the two p-values is significant. Other merging functions, such as the weighted mean, can be used as an alternative but the significance of the resulting combined p-value can not be derived in many instances. Moreover, many merging functions require a weight for each p-value. In order to be able to apply the weights resulting from a power curve simulation, the best way forward is obtaining more knowledge on the copy numbers changes that occur in a metastasis compared to the primary tumor. That way, reliable mutation schemes can be used to ultimately get better weights for the comparison methods.

6.6. The final model

In the previous section, it became apparent that combining the results of the comparison methods into one p-value is not possible yet. Given this result, this section will describe a prototype of the final model that may be applied in practice. In the first subsection, an overview of the advantages and downsides of ASCAT and OncoSNP are given. The second subsection describes a decision tree which can be used to come to a final conclusion about the clonality status of a pair. The decision tree is applied on the results of the 23 fresh frozen pairs in the third subsection. The last subsection compares the

outcome of the decision tree to the clinical assessments.

6.6.1. ASCAT versus OncoSNP

In Chapter 3, the decision was made to incorporate two segmentation algorithms in the study: OncoSNP, which is a Hidden Markov model that considers all three confounding variables of Section 3.1 and ASCAT, which is available as a package in R and takes into account normal cell contamination and aneuploidy. The output of the segmentation algorithms was binned and given as input to the comparison methods which in turn determined if a pair was clonal or not. In Section 6.5, the segmentation algorithm was assumed to be fixed when combining the p-values of the comparison methods, i.e. either ASCAT or OncoSNP was used to segment the profiles of the samples. Before the decision tree can be designed, it is of importance to compare ASCAT and OncoSNP and determine whether one of the two algorithms is to be preferred or not. Table 6.10 shows the advantages and disadvantages of both segmentation algorithms.

	ASCAT	OncoSNP
Advantages	Easily obtained R package. Detailed aberrant cell fraction and ploidy. Output is 1 model.	Better in detecting losses. Not prone to making errors in ploidy. Applicable for all SNP arrays.
Disadvantages	Prone to making errors in the ploidy. Has troubles detecting losses. Applicable only for specific SNP arrays.	Not easily obtained. Coarse ploidy and aberrant cell fraction. Outputs multiple models.

Table 6.10: Advantages and disadvantages of ASCAT and OncoSNP.

ASCAT is an R package which can be easily downloaded from Github. A new version of ASCAT is released every few months. OncoSNP, on the other hand, is harder to obtain: an e-mail had to be sent to the developer as the download link on the OncoSNP website was removed. OncoSNP is a research project from Oxford University and the developer moved to a different university. As a result, the download links were not available from the website. Next to that, OncoSNP is not actively maintained as the most recent version dates from 2014.

Regarding the applicability of the segmentation algorithms, if normal data is not available ASCAT can only be used for specific SNP arrays. As explained in Section 3.3, when normal data is not available ASCAT can predict the germline genotypes only for certain SNP arrays. If the tumor tissue is evaluated on a SNP array which is not supported by ASCAT, the algorithm can not be run unless normal data is available. OncoSNP can always be applied without normal data, but it is optimized for Illumina SNP arrays.

The output of ASCAT yields a detailed aberrant cell fraction and ploidy estimate of the tumor tissue. The aberrant cell fraction is taken with steps of 1% and the ploidy is rounded to two decimals. The ploidy and aberrant cell fraction estimates of OncoSNP, conversely, are coarse, with the aberrant cell fraction taken in steps of 10% and the ploidy rounded to only one decimal. More detail about the ploidy and aberrant cell fraction is thus given in the output of ASCAT.

Given the aberrant cell fraction and ploidy estimate, ASCAT produces one segmentation profile for a sample. OncoSNP outputs multiple models for a sample: two different baselines (diploid and triploid) and five different ranks are considered in the fitting process. As a consequence, a decision about which of the 10 models to use needs to be made. In this study, the model with the highest likelihood and rank 3 was chosen. It must be noted that the choice of the OncoSNP model influences the performance of the comparison methods. For example, consider a pair that has similar patterns in the raw data profiles. If for one sample in the pair, the likelihoods of the two baseline models are close to one another, it may occur that the model with the slightly lower likelihood is similar to the model with the higher likelihood for the other sample in the pair. As the model with the highest likelihood is chosen, the clonality pattern is not detected by the comparison methods. In the study involving the 23 fresh frozen pairs, this problem was not encountered.

OncoSNP seems to perform slightly better than ASCAT when it comes to detecting losses [39]. This was also visible in the 23 fresh frozen pairs: more gains were encountered when the tumor data was segmented using ASCAT than OncoSNP. Some samples fitted by ASCAT even solely consisted of gains. As a result, the independence distribution will consist of many scores favoring clonality which in turn increase the p-values of the pairs. Hence, less pairs are detected as clonal when ASCAT was used. Furthermore, ASCAT is more prone to incorrectly estimating the ploidy leading to clonal pairs not being recognized even though the two profiles look similar. Nevertheless, both problems can be solved with the correction approach introduced in Section 6.2.4. OncoSNP seems to be less prone to making mistakes in the ploidy estimates, but does make mistakes as well. There was one pair, pair FF-10, which, given the output of OncoSNP, was not identified as clonal by the comparison methods even though the raw data profiles showed otherwise. For this pair, the diploid model was the best fitting model for both samples and the discrepancy was caused by the difference in the estimated ploidy.

In summary, both segmentation algorithms have specific advantages and disadvantages. The objective of the final model is to detect as many true clonal pairs as possible. As both segmentation algorithms have imperfections regarding the fitting of the profiles, using only one of the two algorithms will lead to a less sensitive model. Therefore, in order to increase the performance of the final model, both segmentation algorithms will be applied on the raw data.

6.6.2. Decision tree

This subsection describes how, given the results of the comparison methods, a final verdict about the clonality status of the pair can be obtained by means of a decision tree. As two comparison methods are employed, four p-values will be computed for each pair:

- A p-value for the Log LR method, using ASCAT where the segmentation output is corrected as described in Section 6.2.4.
- A p-value for the Log LR method, using OncoSNP.
- A p-value for the adapted SI method, using ASCAT where the segmentation output is corrected as described in Section 6.2.4.
- A p-value for the adapted SI method, using OncoSNP.

Given the four p-values, a decision tree has been designed which can be used by the oncologist. The decision tree divides the pairs into six different categories and states for each category if additional expert opinion is needed or not:

- Category 1: If three or all four p-values are significant, the pair is with high probability of clonal origin. An extra check is not required for these pairs.
- Category 2: If significant p-values are obtained for the Log LR but not for the adapted SI, the pair has many concordant normal states. For these pairs, the number of matching concordant normal bins is given as well. As not many aberrant events are present on both genomes, a conclusion about whether the pair is of clonal origin or not can not be easily made.
- Category 3: If the p-values are significant for the adapted SI but not for the Log LR, many of the bins that are concordant have an aberrant state. The raw data profiles of these pairs need be manually checked by an expert in order to conclude whether the pair is clonal or not.
- Category 4: If the p-values are significant for only one of the two segmentation algorithms, the other segmentation algorithm may have made an error in the fitting of the segmentation profile for one of the two samples. Extra expert opinion regarding the clonality status is required for these pairs.
- Category 5: If only one of the four p-values is significant, the pair is with high likelihood a false positive. In order to confirm this, an extra check of the raw data profiles needs to be performed by an expert.
- Category 6: If no p-values are significant, the two tumors in the pair are, with high likelihood, independent of one another.

For categories 3, 4 and 5, extra manual checks of the raw data are required in order to conclude if a pair is clonal or not. These manual checks should be done by experts who are experienced in interpreting the data coming from a SNP array. The human in the loop (HITL) thus plays a significant role in the decision tree.

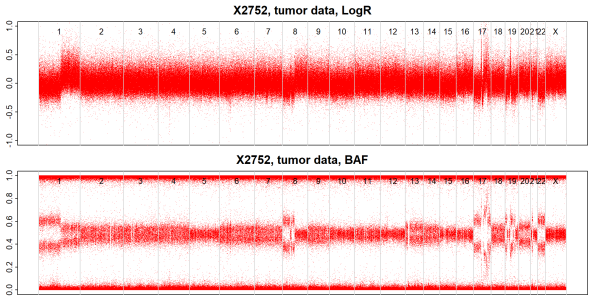
6.6.3. Applying the decision tree on the 23 fresh frozen pairs

Table 6.11 displays how the 23 fresh frozen pairs are divided into the six categories defined by the decision tree. The scores and corresponding p-values of the comparison methods and segmentation algorithms used as input for the decision tree can be found in Appendix D. Note that the color of the pair reflects the judgment of the experts: a green pair implies that the pair is seen as clonal by the experts, while a red pair indicates that the pair is labeled as independent by the experts. The two ambiguous pairs, FF-7 and FF-31B, are colored in black.

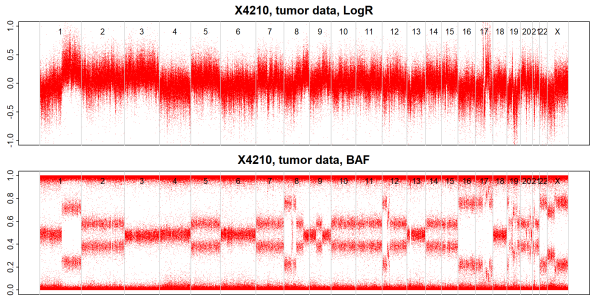
	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6
Pair	FF-10, FF-20, FF-25, FF-31A	FF-6, FF-19	FF-11	-	FF-2, FF-27, FF-29, FF-32, FF-35, FF-36	FF-5, FF-7, FF-13, FF-14, FF-17, FF-18, FF-23, FF-30, FF-31B, FF-33

Table 6.11: Division of the 23 fresh frozen pairs into the six categories defined by the decision tree.

Four pairs are in category 1, all of which were judged as clonal by the experts. The seven pairs in categories 3, 4 and 5, all of which were labeled as independent by the experts, need extra expert opinion in order to derive a conclusion about whether the pair is truly clonal or not. Finally, ten pairs are never labeled as clonal by any of the comparison methods. Of these ten pairs, eight are labeled as independent and two pairs are ambiguous. As explained in Section 6.3, pairs FF-7 and FF-31B belong to a grey area: opinions about the clonality status differed when extra experts were asked to judge the two discordant pairs. The reason why the experts struggled with these pairs has to do with the fact that the raw data plots showed some overlap in the aberrant events, but also plenty of discordant events making it hard to determine if the overlapping events are indeed coming from the primary tumor or if they arose independently. Figure 6.10 shows the raw data plots of pair FF-7 for which the time span between the diagnosis of the primary and secondary tumor is equal to 10 years.



(a) Raw data profile (LRR and BAF values) of the primary tumor of pair FF-7.



(b) Raw data profile (LRR and BAF values) of the secondary tumor of pair FF-7.

Figure 6.10: Raw data plots of the two tumors of pair FF-7.

In Figure 6.10 it can be seen that the aberrant events in the primary tumor, present on chromosomes 1q, 8p, 17 and 19, also occur in the secondary tumor. Compared to the primary tumor, the secondary tumor shows many extra aberrant events. As the time span between the two tumors is 10 years, many extra events can be expected in the case of a metastasis. However, of the four concordant aberrant events, two occur at a so-called hotspot location: chromosomes 1q and 8p are commonly aberrant in breast tumors. This is supported by Figure 5.10. It proved hard to give a concrete judgment for this pair. The reason why pair FF-7 was not detected as clonal by the comparison methods has to do with the fact that many more aberrant events arose in the secondary tumor. As a consequence, the comparison

methods did not detect the four concordant aberrant events as the overall number of concordant bins was small.

In conclusion, it can be stated that the decision tree performs quite well as all the unambiguous clonal pairs (as assessed by the experts) are assigned to category 1. If the 2 ambiguous cases are truly clonal, then further additional branches or alternate criteria for the decision tree are necessary to reliably identify the patients having clonal tumors and thus who would benefit from a change in therapy.

6.6.4. Comparing the decision tree results with the clinical assessments

As mentioned in Section 1.2, the outcome between the molecular and clinical assessment of a pair does not always match. In this section, the degree of concordance between the results of the decision tree and the clinical assessments of the 23 fresh frozen pairs is examined. For this, only the pairs of category 1 are defined as clonal. The two pairs from category 2 were considered as 'missing', i.e. without clonal/independent call and were therefore not included in the comparison analysis. The seven pairs belonging to categories 3 to 5, for which extra expert opinion is needed, are all judged as independent by the experts.

The clinical assessments of the fresh frozen pairs were retrieved from the patient records available at the Erasmus MC. There were three pairs, FF-30, FF-35 and FF-36, for which the clinical data was not available since they have been treated in a different hospital. As a consequence, these pairs will be excluded from the comparison analysis. The two pairs for which the decision tree could not determine the clonality status were clinically evaluated as independent. Of the 18 pairs for which both assessments were available, 15 were regarded as independent and three pairs were assigned as clonal by the clinic. Table 6.12 shows the pairs which were classified as clonal and independent by the decision tree and by the clinic. Note that the color of the pair reflects whether the pair is validated as clonal (green), independent (red) or ambiguous (black) by the experts.

	Decision tree	Clinical assessment
Clonal	FF-10, FF-20, FF-25, FF-31A	FF-7, FF-20, FF-31A
Independent	FF-2, FF-5, FF-7, FF-11, FF-13, FF-14, FF-17, FF-18, FF-23, FF-27, FF-29, FF-30, FF-31B, FF-32, FF-33, FF-35, FF-36	FF-2, FF-5, FF-10, FF-11, FF-13, FF-14, FF-17, FF-18, FF-23, FF-25, FF-27, FF-29, FF-30, FF-31B, FF-32, FF-33, FF-35, FF-36

Table 6.12: Clonal and independent pairs classified by the decision tree and by the clinic.

Comparing the results of the decision tree with the clinical assessments, it can be seen that the decision tree has a higher sensitivity: all clonal pairs are labeled as such by the decision tree, while only two of the four clonal pairs are regarded as a metastasis by the clinic. This shows that clonality in contralateral breast cancer pairs can not be based on clinical assessment alone. A discordance between the clinical assessment and output of the decision tree is present for three out of 18 pairs, which is 16.7%. This is lower than the 35% discordance found in [5]. In order to quantify the degree of concordance between the two methods, Cohen's kappa (κ) coefficient has been computed. Cohen's kappa is defined as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed agreement between the two methods and p_e is the expected agreement between the two methods that would have arisen by chance [11]. If the two methods are in complete agreement, $\kappa = 1$. If the level of agreement between the two methods is equal to the expected agreement that would be expected by chance, $\kappa = 0$. Note that the coefficient can also be negative, implying that the concordance between the two methods is worse than random or there is no effective agreement between the two methods. Table 6.13 shows the number of pairs which were labeled as clonal and independent by the decision tree and the clinic.

	Clonal clinic	Independent clinic	Total
Clonal decision tree	2	2	4
Independent decision tree	1	13	14
Total	3	15	18

Table 6.13: Number of pairs which were labeled as clonal and independent by the decision tree and the clinic.

The observed agreement, p_o, and the agreement occurring by chance, p_e, are equal to:

$$\begin{aligned} p_o &= \frac{2+13}{18} = \frac{5}{6}, \\ p_e &= \frac{2+2}{18} \cdot \frac{2+1}{18} + \frac{1+13}{18} \cdot \frac{2+13}{18} = \frac{37}{54}. \end{aligned}$$

Given these values, Cohen's kappa is equal to:

$$\kappa = \frac{\frac{5}{6} - \frac{37}{54}}{1 - \frac{37}{54}} = \frac{8}{17} \approx 0.4706.$$

This implies that there is a moderate agreement between the results of the decision tree and the clinical assessment.

Even though the decision tree has a higher sensitivity, the clinical assessment showed that pair FF-7, which is an ambiguous pair, is judged as clonal by the treating oncologist. It was described that the patient had multiple distant metastases in the time between the diagnosis of the primary and secondary tumor. As a result, the secondary tumor in the contralateral breast was also reported as a metastasis. As stated in Section 6.6.3, the reason why pair FF-7 is not labeled as clonal by the decision tree has to do with the fact that the secondary tumor shows many more aberrant events than the primary tumor so that the overall degree of concordance in states of the bins is quite small. This is a general shortcoming of the model. Nevertheless, the fact that pair FF-7 is also judged as clonal by the clinic supports the clonality judgments of some of the experts.

In summary, a moderate agreement is present between the decision tree and clinical assessments, with only two pairs labeled as clonal by both methods. Of the three discordant pairs, two clonal pairs were labeled as clonal by the model but not by the clinic and one ambiguous pair was clinically judged as clonal but seen as independent by the model. This shows that it is of importance to always take into account both the clinical results and the DNA profiles of two tumors when determining if a pair is clonal or not. When the molecular features of a pair are not considered, a pair may be labeled as independent, while the secondary tumor shows a very similar copy number profile, thus likely being a metastasis of the primary tumor. Conversely, if only the DNA profiles of the tumors are considered, the model may not be able to detect the pair as clonal as the secondary tumor may display significantly more aberrant events than the primary tumor even though the two tumors share the same aberrant events. Until these hard to judge DNA profiles can be reliably categorized, it is therefore advised to incorporate both the clinical judgments as well as the decision tree results when determining the clonality status of a pair.

7

Summary

Unilateral breast cancer patients have an increased risk to develop a secondary tumor in the contralateral breast. When a secondary tumor arises, it is of importance to determine whether this tumor is a new independent tumor or a metastasis of the first tumor as this is of influence on the therapy plan. In most cases, a second tumor is regarded as a new independent tumor. Only when there is clinical suspicion that the tumor may be a metastasis, clinico-pathological characteristics, such as the histological subtype and hormone receptor status, of the two tumors are used to assess whether the second tumor is clonal. However, using clinico-pathological characteristics to determine clonality is not desirable as the majority of breast cancers have similar characteristics. Moreover, it is possible that some pathological characteristics change in case of a metastasis so that determining clonality based on pathological features leads to a lower sensitivity. This demonstrates that it is of importance to also consider the molecular features when determining if a pair is of clonal origin or independent. Recent studies have developed classification models that can determine the clonality status of a pair based on the DNA profiles of the two tumors. These studies have also shown that the judgment based on the molecular features is not always in concordance with the judgment based on the clinical characteristics.

The aim of this thesis was to develop a more advanced classification model that can determine if a secondary tumor is a metastasis of the primary tumor or a new independent tumor based on the SNP array profiles of the two tumors. In order to quantify the degree of clonal relatedness between the DNA profiles of the two tumors, two steps were taken. In the first step, a segmentation algorithm was applied to the SNP array data to determine the copy number aberrant regions on the genomes. After this, a statistical method, i.e. comparison method, was applied on the resulting segmentation profiles to determine the clonality status of the pair.

In Chapter 3 several segmentation algorithms have been investigated. Given the overview of the segmentation algorithms, ASCAT and OncoSNP appeared to be the most sophisticated algorithms and were therefore chosen for this study. The two chosen algorithms use both the LRR and BAF data as input. Taking both the LRR and BAF data into account is important as it enables us to distinguish normal regions from CN-LOH regions. Moreover, ASCAT and OncoSNP correct for at least two out of three confounding variables such as normal cell contamination and aneuploidy. When a segmentation algorithm does not consider the confounding variables when fitting the raw data, the algorithm may incorrectly define the aberrant regions on the genome which in turn will decrease the sensitivity of the comparison methods.

After the samples are segmented by the segmentation algorithm, the copy numbers of the SNPs are coarsened into four different states: loss, normal, CN-LOH and gain. This is done to increase the sensitivity of the comparison methods. As a sample is diploid in the normal setting, copy number two is used as a baseline. A loss is defined when the copy number is one or smaller, while a gain implies that the copy number is three or higher. The number of SNPs on a SNP array are generally quite large. Therefore, comparing the states SNP by SNP can be quite time consuming. Moreover, the segmentation algorithms may also make errors in determining the boundaries of the aberrant regions due to

noise being present in the raw data. In order to solve both problems, the SNPs are binned together using distance based bins (1 Megabases distance) where no overlap is present between the bins. The state of a bin is equal to the most common state of the SNPs in the bin.

Chapter 4 introduced the two comparison methods which were developed for this study. Both comparison methods have not been used in literature before. The first comparison method that has been invented is the Log Likelihood Ratio method. This comparison method is an adapted and extended version of the Likelihood Ratio method created by [36]. The original LR method can be used to determine if a pair is clonal, but it has two major shortcomings: the method can not distinguish normal from CN-LOH events and moreover only allows one aberrant event, i.e. gain or loss, per chromosome arm. The Log LR method allows all four aberrant states to occur and moreover allows multiple aberrant events per chromosome arm. The second comparison method that has been applied is the adapted Similarity Index method. This comparison method is a modified version of the Similarity Index method invented by [31]. In contrast to the original SI, the adapted SI penalizes certain combinations of states in the primary and secondary tumor more than others. For instance, a gain-loss combination in the primary and secondary tumor is penalized more than a gain-normal combination. Both comparison methods are permutation tests where the null hypothesis states that the two tumors arose independently. Each comparison method computes a score for each pair which reflects the degree of concordance in the states of the bins. Given the comparison scores of the pairs, the p-values of the scores can be determined by means of constructing an independence distribution for which the comparison scores of primary and secondary tumors coming from different pairs are computed. If the p-value falls below the significance level α, the null hypothesis is rejected and the pair is labeled as clonal.

Chapter 5 has investigated the characteristics of the two comparison methods by means of simulating artificial pairs. For this, a dataset of 50 sporadic breast cancer patients has been used. Artificial pairs were constructed by means of applying state changes to the bins of the 50 samples. A specific workflow has been designed for this purpose, where the change probabilities of a gain and loss were varied. An external dataset consisting of 201 patients has been used to determine the transition probabilities of the states. It should be noted that the distribution of the samples in this dataset, i.e. how many gains, losses etc. have occurred, was completely different than the distribution of the 50 breast cancer samples. This had to do with the fact that the 201 samples of the external dataset were segmented using a different segmentation algorithm. However, even though the way in which the state changes were applied were not fully realistic, the simulations still gave an indication of the workings of the comparison methods. The simulation results showed that the Log LR method outputs relatively small p-values for samples which consist of many losses as these states are rare to occur. The adapted SI, on the other hand, assigned relatively high p-values to samples with many losses as these samples consist of many normal states which undergo many changes in the simulations. For samples with many normal states, the adapted SI yielded very high p-values. This had to do with the fact that samples with many normal states undergo many changes in the simulations and moreover do not have many aberrant events before mutations are added. The Log LR method did not yield very high p-values for samples with many normal states as these samples also consist of relatively many loss states which decrease the p-value. The adapted SI mainly assigned low p-values for samples which consist of many gains as these samples undergo the least number of changes. For the Log LR, samples with many gains had relatively high p-values since concordant gains do not contribute much to the clonality hypothesis as gains are already quite common to occur. Comparing the behavior of the Log LR and adapted SI with one another, it can be concluded that the adapted SI mainly focuses on the number of concordant events between the primary and secondary tumor, while the Log LR focuses more on the rarity of a concordant event.

In Chapter 6, the segmentation algorithms and comparison methods were tested on 23 fresh frozen pairs. The DNA of the 23 pairs was analyzed on an Illumina GSA v3 SNP array. In order to validate the performance of the model, two experts from the Erasmus MC in Rotterdam have looked at the raw data of the pairs and stated for each pair whether it is of clonal origin or not. Of the 23 pairs, 6 were judged as clonal and 17 were seen as independent. The significance level α was set equal to 0.05. When the 23 pairs were fitted with ASCAT, it became apparent that a correction is needed for some pairs as ASCAT sometimes erroneously estimates the ploidy number of one of the two samples in a pair. The raw data

profiles of two samples may look similar, but the ploidy numbers are estimated differently for the two samples. This leads to two different ASCAT profiles and the pair is not detected as clonal by the comparison methods. In order to solve this, a correction has been applied to the ASCAT output as follows: if the difference in the estimated ploidy numbers of a pair is larger than 1.5, the sample with the largest ploidy number is corrected downwards. This way of correcting gave satisfactory results as four out of six validated clonal pairs were also classified as clonal by the comparison methods. No correction was needed for the output of OncoSNP. However, the results showed that two pairs which were validated as clonal, pairs FF-7 and FF-31B, were never recognized as clonal by any of the comparison methods for both segmentation algorithms. Three additional experts were asked to judge these two pairs and opinions regarding the clonality status varied. Therefore, these pairs appear to belong to a grey area: the clonality status can not be reliably determined. The performance of the segmentation algorithms and comparison methods were found to depend on two factors: the accuracy of the segmentation algorithms and the distribution of the samples which are used in the independence distribution.

Given one of the two segmentation algorithms, each comparison method outputs a p-value for each pair. In other words, two p-values are obtained for each pair. An attempt was made to combine the results of the comparison methods by means of using the harmonic mean p-value. The harmonic mean p-value requires that each p-value gets assigned a specific weight. Two different approaches were investigated for deriving the weights of the comparison methods: expert judgment and a power curve simulation. However, both approaches did not produce satisfactory results. When expert judgment is used to obtain the weights, the weights only hold for this dataset and can not be generally applied. If a new dataset arises, expert judgment is again needed to derive the weights. The power curve simulation is a general approach, but uses artificial pairs to determine the power of the comparison methods. As long as the way in which the artificial pairs are constructed does not reflect reality, the weights resulting from the power curve simulation are not realistic as well. Moreover, the choice of the harmonic mean p-value may also be questioned as a downward effect is present: if one of the two p-values is small and the other fairly large, the harmonic mean p-value is also small. This may result in more false positives being included in the final output. Hence, the harmonic mean p-value did not seem like an appropriate candidate.

Instead of choosing one segmentation algorithm and combining the results of the comparison methods by means of a merging function, both segmentation algorithms are considered in the final model. This is done to increase the sensitivity: if one segmentation algorithm makes a mistake in the segmentation process, the other algorithm can still be used as a back-up. In order to come to a final verdict for each pair, a decision tree has been designed which can be used by the oncologist. In the decision tree, all four p-values resulting from the two segmentation algorithms and comparison methods are taken into account. Each pair is assigned to one of the six different categories dependent on which p-values are found to be significant. If, for example, all four p-values of a pair are significant, the pair is with high likelihood of clonal origin. For some categories in the decision tree extra expert opinion is required before a final verdict can be made. The decision tree was tested on the 23 fresh frozen pairs and satisfactory results were found. The four pairs for which a concordant clonal judgment was obtained also belong to the clonal category of the decision tree. Next to that, there are two pairs for which the clonality status can not be determined by the decision tree as the primary and secondary tumor both contain too many normal states. As the number of aberrant events is small for both tumors, a conclusion about whether the pair is clonal or not can not be derived. Of the 23 pairs, seven belonged to a category for which extra expert opinion is required. These seven pairs were all judged as independent by the experts. The two ambiguous pairs, FF-7 and FF-31B, belonged to the independent category.

For 18 of the 23 pairs, both the decision tree results and clinical assessments were available. A discordance between the clinical judgment and the decision tree was found in three of the 18 pairs (16.7%), which is lower than the 35% discordance found in [5]. Cohen's kappa coefficient showed that the results of the decision tree were in moderate agreement with the clinical assessments ($\kappa = 0.47$). The decision tree proved to have a higher sensitivity than the clinical evaluation. However, a downside of the decision tree is that the model is not capable to detect clonality when the secondary tumor shows many more aberrant events than the primary tumor even though the aberrant events in the primary tumor are also present in the secondary tumor. This occurred for one of the two ambiguous pairs: FF-7

was labeled as independent by the decision tree, while the clinic regarded the pair as clonal. Therefore, until the ambiguous pairs can be accurately classified, both clinical assessment and molecular features need to be taken into account when determining if a secondary tumor is of clonal origin or not.

8

Conclusion and discussion

This chapter concludes the thesis and gives recommendations for future research.

8.1. Conclusion

The objective of this thesis is to develop a more advanced classification model that can identify whether a contralateral breast cancer pair is of clonal origin or independent based on the DNA profiles of the two tumors. In this research, the DNA profiles were generated by means of SNP arrays. Given this data, two segmentation algorithms and two comparison methods have been investigated. Both comparison methods are permutation methods which test the hypothesis of independence against the clonality hypothesis. A p-value, which reflects the probability that the pair is clonal, is computed by means of comparing the comparison score to the scores in the independence distribution. If the p-value of the comparison score is below a defined significance level, the pair is labeled as clonal by the comparison method.

The final model, where a decision tree is applied on the four p-values to come to a definite conclusion about the clonal status of a pair, demonstrated promising results for the 23 fresh frozen pairs. All of the four pairs unanimously validated as clonal by the experts were assigned to category 1. For seven out of 23 fresh frozen pairs, expert opinion is needed for a final call, two pairs showed too few events to make a call, while the remaining 14 pairs would not need additional input. If the expert calls reflect the true clonal status, four patients would receive a different therapy than was given for their primary tumor. If the current results hold for larger cohorts, it may be that only category 1 pairs would be a candidate for alternate therapy while the oncologist can use standard-of-care criteria to treat the patients with tumors in the other categories.

The decision tree showed to have a higher sensitivity than the clinical assessments of the 23 pairs. Moreover, the model is more advanced than already existing models. Since both the LRR and BAF data are considered in the segmentation process, CN-LOH events can be discriminated from normal regions. Next to that, gain and loss regions are allowed to occur on the same chromosome arm so that a more comprehensive comparison of what has happened on the genome can be performed. Binning the SNPs together smooths out possible errors made by the segmentation algorithm. In summary, the final model appears to be suitable to use in practice. Furthermore, the model is not restricted to contralateral breast cancer pairs: it is also applicable when the secondary tumor arose in a different organ such as the lungs or the brain. However, as the model struggles to detect clonality in the ambiguous pairs, it is advised, until more clarity regarding the clonal status of the ambiguous pairs is obtained, to incorporate both the decision tree results and clinical assessments when determining the clonality status of a pair.

8.2. Recommendations for future research

The model has been tested on 23 fresh frozen pairs, but not yet on FFPE pairs. The performance of the model highly depends on the DNA quality of the samples. The DNA extracted from fresh frozen

material is generally of high quality so that the segmentation algorithms can accurately determine the aberrant copy number regions. As explained in Section 4.1, the DNA quality of FFPE tumor tissues is generally lower than that of fresh frozen tumors due to formalin fixation. Since the DNA quality of FFPE samples is lower, more noise is present in the raw data. As a consequence, the segmentation algorithms may also struggle to determine the copy number aberrant regions. An experiment was conducted with a small set of FFPE samples consisting of three tumors coming from the years 1996, 2003 and 2010 to investigate how the segmentation algorithms perform on FFPE data. The results of the experiment showed that the segmentation algorithms perform decently on the FFPE samples, provided that the DNA of the samples is restored using the correct DNA restoration kit before being analyzed on the SNP array. A complete description of the experiment can be found in Appendix E. Unfortunately, the set of FFPE samples was too small to assess the performance of the comparison methods. It is therefore advised that the final model is also tested on a sufficiently large set of FFPE pairs. In the near future, the Erasmus MC is planning to test the model on 24 FFPE pairs.

When more pairs are evaluated by the model, it has to be decided whether the set of new pairs is to be defined as a new cohort or if the set of pairs is added to the already existing database. In the latter case, more observations will be added to the independence distribution yielding more refined p-values. Note that the probabilities in the Log LR need to be updated when the pairs are added to the existing cohort as the Log LR may otherwise not be defined. For instance, if in the set of 23 pairs the frequency of a loss in a certain bin is zero and if a loss occurs in this specific bin for a new pair, using the frequencies of the 23 pairs would yield a logarithm of zero for this bin, which is equal to minus infinity. As a consequence, the total log likelihood under both hypotheses will be minus infinity resulting in a non-existent Log LR. Therefore, it is of importance to always update the frequencies of the different events when extra pairs are included in the database. A downside of extending the existing database is that the choice of the SNP array could cause problems. ASCAT can only be run without data of matched normal tissue for specific SNP arrays, while OncoSNP is optimized for Illumina SNP arrays. Next to that, some SNP arrays are constituted of less SNP positions. Since the SNPs are binned after segmentation, the number of SNPs in each bin may be considerably less if a different array is used. As explained in Section 4.1, a bin is only taken into account in the comparison methods if the number of SNPs in the bin is at least 10. If a different SNP array is used for the additional pairs, the number of bins on which the comparison is made may decrease. Even though the number of bins may be smaller, the corresponding p-values of the extra pairs can still be computed. For this, the independence distribution needs to be evaluated only on the bins for which the extra pairs have more than 10 SNPs. Nevertheless, as the comparison is done on less bins, the comparison scores may be less reliable. Therefore, if the decision is made to add the new pairs to the already existing cohort, the SNP array should be carefully chosen. In the most ideal case, additional samples are run on the same SNP array as the existing samples in the database. If this is not possible, it is better to regard the set of pairs as a new cohort. Another possibility would be to re-run the existing pairs on the new SNP chip but this can only be done if the tissue samples of the existing pairs are available. More research about what would be the most ideal decision regarding the extension of the model is required.

The possibility to combine the results of the comparison methods has been investigated by means of applying the harmonic mean p-value. For this, the segmentation algorithm was assumed to be fixed, i.e. either ASCAT or OncoSNP was used to segment the raw data profiles. The harmonic mean p-value did not perform as desired since small p-values decreased the harmonic mean p-value leading to more false positives being called by the model. Moreover, in Section 6.6.1 it became clear that the sensitivity of the model is increased when both segmentation algorithms are incorporated in the model. A suggestion for future research is to examine whether other merging functions can be applied on all the four p-values together. In order to apply a merging function, each comparison method needs a different weight. A power curve simulation can be employed to derive the weights for each comparison method. However, the power curve simulation only produces reliable results if the way in which artificial pairs are constructed reflect reality. As long as this is unknown, the power curve simulation results can not be applied. It is therefore desired that more research is conducted on how mutations arise in secondary clonal tumors. How does a metastasis mutate from its primary tumor and at what rate?

Next to that, the current workflow in the power curve simulation makes use of an external dataset

to determine the transition probabilities of going from one state to another. However, the distribution of this external dataset does not match the distribution determined by the segmentation algorithms. This is caused by the fact that the external dataset is segmented using a different segmentation algorithm than the samples to which the mutations are applied. A more realistic power curve simulation result may be obtained if the same segmentation algorithm is applied to both the external dataset as well as the data to which the mutations are applied. More research to what extent this will contribute to a more realistic output is needed.

Instead of combining the p-values of the comparison methods into one p-value, a decision tree has been designed consisting of six categories. Each pair was assigned to one of the six categories dependent on which p-values were found to be significant. The decision tree only took into account when a pair had a p-value below 0.05, but not when the p-value was close to being significant, i.e. $p \in (0.05, 0.1)$. An idea for future research would be to extend the decision tree by also incorporating whether a pair has a p-value that is close to being significant.

Currently, the model does not really incorporate prior knowledge regarding specific mutation regions when computing the comparison scores. The entire genome is treated equally regarding the penalization of discordant events. In breast cancer, specific chromosome arms are so-called hotspots for copy number alterations, i.e. an aberrant event commonly arises on these arms. If, for example, a gain occurs on a certain chromosome arm which is specific for breast cancer then this gain will, in case of a metastasis, remain a gain with a high probability. If this is not the case, the discordant event should be penalized more than discordant events on other chromosome arms. It is advised to examine how prior knowledge can be included in the model.

The human in the loop plays a significant role in the decision tree. For three out of six categories, additional expert opinion is required to come to a final conclusion. Next to that, the performance of the current model depends on both the accuracy of the segmentation algorithms and the distribution of the other samples in the cohort. In order to overcome both problems, it is recommended to explore whether clonality can also be detected by means of artificial intelligence, such as deep learning models. Given the raw data plots of the pairs, deep learning models may be capable to autonomously determine whether a pair is of clonal origin, provided that they are trained on a sufficiently large training set. An advantage of a deep learning model is that the classification of a pair does not depend on the distribution of the samples in the other cohort. Given a decent training set, the model attempts to find similar features in the raw data of the two tumors. Deep learning models have been applied before on SNP array data: [13] has developed a neural network that can detect copy number aberrant regions from the LRR and BAF data. However, the reliability of this neural network can be questioned as it has been trained on only 12 samples so far. A problem that arises when applying a deep learning model for this purpose is the training set. How large should the training set be chosen to ensure that a reliable prediction is obtained? The optimal size may depend on the type of deep learning model, i.e. structure of the model, that is being chosen. Moreover, each pair in the training set should be validated as clonal or independent. The reliability of the labels can be questioned if expert judgment has been used to assess the clonality status of the pairs. Are the opinions of the experts trustworthy or not? More research regarding the applicability of deep learning models and how to obtain an adequate training set is required.

The experts have used the raw SNP array data to label the 23 fresh frozen pairs as either clonal or independent. However, the clonality status of two pairs remained ambiguous: are these pairs clonal or not? This shows that experts can not always validate a contralateral pair based on copy number data alone. For future research it might be an idea to also incorporate additional information, such as the somatic mutations occurring in the tumor driver genes of the two tumors. This might give the experts more confidence in their judgment whether a pair is of clonal origin or not.

The final model considers only the copy number variations that have occurred on the genomes of the two tumors to determine if a pair is clonal or not. In the same way as the experts, the model may be extended by including additional molecular features of the two tumors such as the somatic mutations that appear at specific tumor driver genes. Further research about which tumor driver genes to include is needed.

Both ambiguous pairs, pairs FF-7 and FF-31B, were not detected as clonal by any of the comparison methods and thus assigned to the independence category by the decision tree. The reason why pair FF-7 was not detected as clonal had to do with the fact that the raw data plot of the secondary tumor showed more aberrant events than the primary tumor. However, the four aberrant events on the primary tumor did also occur in the secondary tumor which is a sign that the pair may be clonal. In order to prevent pairs with many more aberrant events in the secondary tumor from falling in the independence category, the model may be improved by using a smaller penalization in the comparison scores for normal events in the primary and aberrant events in the secondary tumor. For instance, the penalization factor for this combination of events can be lowered or not even penalized at all. The adapted SI can be easily modified to accomplish this. The penalization factor may also be dependent on the time between the diagnosis of the primary and secondary tumor: the larger the time span, the smaller the penalization factor. More sample pairs with such profiles are needed to determine the most optimal penalization factor for the adapted SI and to investigate how to properly penalize these events for the Log LR.

The current model coarsens the copy numbers into four different states: loss, normal, CN-LOH and gain. However, a distinction between a single loss and a double loss is not made in the model. If the primary tumor has a double loss in a certain region and the secondary tumor has copy number one in the same region, the model states that this is a clonal event while in reality a double loss can never go back to a single loss. A similar argument can be made for gain events: if the primary tumor has a relatively high copy number, say 10, in a specific region and the secondary tumor has copy number 3 in the same region the model assigns this as a concordant event even though it is unlikely that copy number 10 will go back to copy number 3 in case of a metastasis. Therefore, in order to improve the sensitivity of the model it may be an idea to extend the number of states from four to six: double loss, single loss, normal, CN-LOH, gain and high copy number gain. In order to make a distinction between a gain and high copy number gain, a specific threshold x needs to be determined. If the copy number of a SNP is larger than two but below x, the SNP is assigned as a gain, while for a copy number larger than or equal to x, the SNP is a high copy number gain. The optimal value for x should be further examined. When determining the optimal value for x, the maximum copy numbers that can arise in the segmentation profiles need to be taken into account. In Section 3.4 it can be seen that the maximal copy number which can be assigned by OncoSNP is equal to six. Therefore, if the threshold x is set to seven or larger, a high copy number gain will never occur in the segmentation profiles of OncoSNP. It is therefore advised that the threshold x is set smaller than or equal to six. Next to that, when the model is expanded to six states, the comparison methods also need to be adapted so that certain combinations of events are penalized more than others. For instance, a larger penalization in the adapted SI should be applied to discordant events for which a double loss occurs in the primary tumor as these are a sign that the tumor is not of clonal origin.

Appendices

A

Derivation of the probabilities in the LR

The report has derived the probability of a concordant gain or loss used in the LR method. In this appendix, the probabilities of the other indicators in the likelihood of the LR method are derived.

A.1. Two different events of which one is aberrant and one is normal

Two different events can by definition not be clonal. Therefore the probability of two different events of which one is a normal and the other an aberrant event can only arise independently of one another.

Let's examine the probability that one tumor has a gain at chromosome arm i and the other tumor is normal at the i-th chromosome arm. The other probabilities for this event are derived similarly.

The probability of a gain and normal occurring at chromosome arm i are by the law of total probability:

$$\begin{aligned}\mathbb{P}(\text{gain and normal at chr. arm } i) &= \mathbb{P}(\text{gain and normal at chr. arm } i|\text{no clonal event})\mathbb{P}(\text{no clonal event})\\ &+ \mathbb{P}(\text{gain and normal at chr. arm } i|\text{clonal event})\mathbb{P}(\text{clonal event})\\ &= \mathbb{P}(\text{gain and normal at chr. arm } i|\text{no clonal event})\mathbb{P}(\text{no clonal event}).\end{aligned}$$

The probability of no clonal event is equal to:

$$\begin{aligned}\mathbb{P}(\text{no clonal event on chr. arm } i) &= 1 - \mathbb{P}(\text{clonal event on chr. arm } i)\\ &= 1 - \mathbb{P}(\text{clonal gain on chr. arm } i) - \mathbb{P}(\text{clonal loss on chr. arm } i)\\ &= 1 - cp_{1i} - cp_{2i}.\end{aligned}$$

This probability of a normal and a gain event on chromosome arm i given no clonal event is equal to:

$$\begin{aligned}\mathbb{P}(\text{gain and normal at pos. } i|\text{no clonal event}) &= 2\mathbb{P}(\text{gain at pos } i|\text{no clonal event}) \cdot\\ &\quad \mathbb{P}(\text{normal at pos } i|\text{no clonal event})\\ &= \frac{2(1-c)p_{1i}p_{3i}}{(1-cp_{1i}-cp_{2i})^2}.\end{aligned}$$

Note that the 2 comes from the fact that a gain and normal can occur in two different ways: a gain in the primary and a normal in the secondary or vice versa. This needs to be taken into account in the probabilities as well. Putting everything together, it follows that:

$$\begin{aligned}\mathbb{P}(\text{gain and normal at chromosome arm } i) &= \frac{2(1-c)p_{1i}p_{3i}}{(1-cp_{1i}-cp_{2i})^2}(1-c_{1i}-cp_{2i})\\ &= \frac{2(1-c)p_{1i}p_{3i}}{1-cp_{1i}-cp_{2i}}.\end{aligned}$$

A.2. Two different aberrant events

The only possibility in which two different aberrant events can arise is when one tumor has a gain and the other tumor a loss on chromosome arm i. The probability of a gain and a loss at chromosome arm i is equal to the probability that the two aberrant events arose independently given no clonal event. In other words:

$$\mathbb{P}(\text{gain and loss at chr. arm } i|\text{no clonal event})\mathbb{P}(\text{no clonal event}).$$

The probability of a gain and a loss given no clonal event is equal to:

$$\begin{aligned}\mathbb{P}(\text{gain and a loss at chr. arm } i|\text{no clonal event}) &= 2\mathbb{P}(\text{indep. gain at pos i}|\text{no clonal event})\\ &\quad \mathbb{P}(\text{indep. loss at pos i}|\text{no clonal event})\\ &= \frac{2(1-c)^2 p_{1i}p_{2i}}{(1-cp_{1i}-cp_{2i})^2}.\end{aligned}$$

Note again the factor 2 which is used in the probability as the loss and gain can occur in two different settings as mentioned before. Putting everything together, the total probability equals

$$\begin{aligned}\mathbb{P}(\text{gain and loss at chr. arm } i) &= \frac{2(1-c)^2 p_{1i}p_{2i}}{(1-cp_{1i}-cp_{2i})^2}(1-c_{1i}-cp_{2i})\\ &= \frac{2(1-c)^2 p_{1i}p_{2i}}{1-cp_{1i}-cp_{2i}}.\end{aligned}$$

A.3. Two normal events

By the law of total probability, the probability of a concordant normal event is equal to the probability of a clonal normal event plus the conditional probability that the two normal events arose independently given that there was no initial clonal event.

Note that a concordant normal event is by definition not clonal, as only aberrant events are considered as possible clonal events. Therefore, the probability of a clonal normal event is equal to zero.

Next to that, the probability that two normal events arose independently at chromosome arm i given that there was no clonal event is equal to:

$$\mathbb{P}(\text{concordant normal arose indep. at chr. arm } i|\text{no clonal event})\mathbb{P}(\text{no clonal event})$$

The first probability in the equation above is equal to:

$$\begin{aligned}\mathbb{P}(\text{indep. conc. normal at chr. arm } i|\text{no clonal event}) &= \mathbb{P}(\text{indep. normal}|\text{no clonal event})^2\\ &= \left(\frac{\mathbb{P}(\text{ind. normal at chr. arm } i \cap \text{no clonal event})}{\mathbb{P}(\text{no clonal event})}\right)^2\\ &= \left(\frac{p_{3i}}{1-cp_{1i}-cp_{2i}}\right)^2.\end{aligned}$$

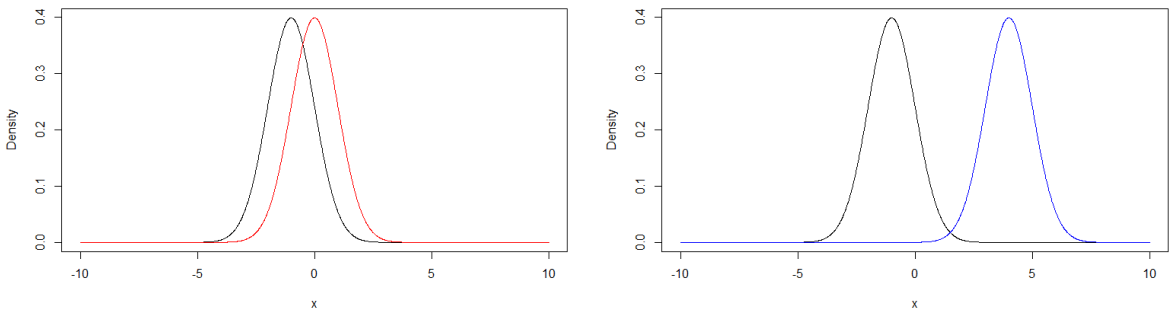
Putting everything together, it follows that:

$$\begin{aligned}\mathbb{P}(\text{concordant normal at chr. arm } i) &= \left(\frac{p_{3i}}{1-cp_{1i}-cp_{2i}}\right)^2(1-c_{1i}-cp_{2i})\\ &= \frac{p_{3i}^2}{1-cp_{1i}-cp_{2i}}.\end{aligned}$$

B

Wasserstein distance

The Wasserstein distance is a metric that describes the distance between two probability distributions P and Q on a given metric space. Figure B.1 shows the probability density functions of three normal random variables. In Figure B.1a, a normal distribution with mean -1 and variance 1 (black line) is plotted together with a normal distribution having mean 0 and variance 1 (red line). Figure B.1b shows a normal distribution with mean -1 and variance 1 (black line) and a normal distribution having mean 4 and variance 1 (blue line). Each probability distribution can be seen as a heap of dirt. The Wasserstein distance computes the distance between two probability distributions by intuitively answering the following question: what is the minimal energy needed to turn one heap into the other? In other words, the metric is the minimum so-called cost of turning one heap into the other which is defined as the amount of earth that needs to be moved times the mean distance with which the earth has to be moved. Therefore, the Wasserstein distance is also known as the Earth Mover's distance.



(a) Two probability density functions: the black line corresponds to a normal distribution with mean 1 and variance 1, the red line to a normal distribution with mean 0 and variance 1.

(b) Two probability density functions: the black line corresponds to a normal distribution with mean 1 and variance 1, the blue line to a normal distribution with mean 4 and variance 1.

Figure B.1: Three probability density functions of normal random variables.

Looking at Figure B.1, the black and the red line plotted in the left panel are located closer to one another than the black and the blue line in the right panel. As a consequence, moving the black heap to the red heap requires less energy than moving the black heap to the blue heap. Hence, the Wasserstein distance between the distributions in Figure B.1a is smaller than the Wasserstein distance between the distributions of Figure B.1b.

Let P and Q be two one-dimensional probability distributions defined on the same metric space. Define $\Gamma(P, Q)$ as the set of all possible couplings (mappings) between P and Q, i.e. $\Gamma(P, Q)$ defines all possibilities with which the masses between P and Q can be moved. The first order Wasserstein distance is

defined as follows [37]:

$$\begin{aligned} W_1(P,Q) &= \inf_{\gamma \in \Gamma(P,Q)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| d\gamma(x,y) \\ &= \int_{\mathbb{R}} |F_P(x) - F_Q(x)| dx \\ &= \int_0^1 |Q_P(x) - Q_Q(x)| dx, \end{aligned}$$

where F is the cumulative distribution function and Q the quantile function of the probability distribution. Note that $Q_Q(x)$ denotes the quantile function of the probability distribution Q.

In general, the p-th order Wasserstein distance is defined as:

$$\begin{aligned} W_p(P,Q) &= \inf_{\gamma \in \Gamma(P,Q)} \left(\int_{\mathbb{R}\times\mathbb{R}} |x-y|^p d\gamma(x,y) \right)^{\frac{1}{p}} \\ &= \left(\int_{\mathbb{R}} |F_P(x) - F_Q(x)|^p dx \right)^{\frac{1}{p}} \\ &= \left(\int_0^1 |Q_P(x) - Q_Q(x)|^p dx \right)^{\frac{1}{p}}. \end{aligned}$$

Note that the distributions P and Q are assumed to be continuous, but the Wasserstein distance can also be defined for discrete distributions. In that case the Wasserstein distance becomes a sum over all the possible values that the distributions P and Q can take:

$$W_p(P,Q) = \left(\sum_{x \in \mathcal{X}} |F_P(x) - F_Q(x)|^p \right)^{\frac{1}{p}},$$

where $\mathcal{X}$ is the union of the values on which P and Q are defined and F the cumulative distribution function. The Wasserstein metric thus measures the distance between two probability distributions.

In order to apply the Wasserstein distance on the output of the segmentation algorithm, the output needs to be converted into a probability distribution. The output for a pair of tumors consists of two vectors of length N (number of bins) existing of the integers 1, 2, 3 and 4, where each number corresponds to one of the four discretized states, i.e.:

$$s^{(k)} = \left(s_1^{(k)}, s_2^{(k)}, \dots, s_N^{(k)} \right), \; k = 1, 2,$$

where $s_i^{(k)}$ is the state at bin i for tumor k, $s_i^{(k)} \in \{1, 2, 3, 4\}$.

Given the vectors $s^{(k)}$ for $k = 1, 2$ the Wasserstein distance can be employed by constructing histograms consisting of N bars where each bar corresponds to a bin and the height of each bar reflects the state of the bin. Figure B.2 shows an example of two sequences consisting of six bins.

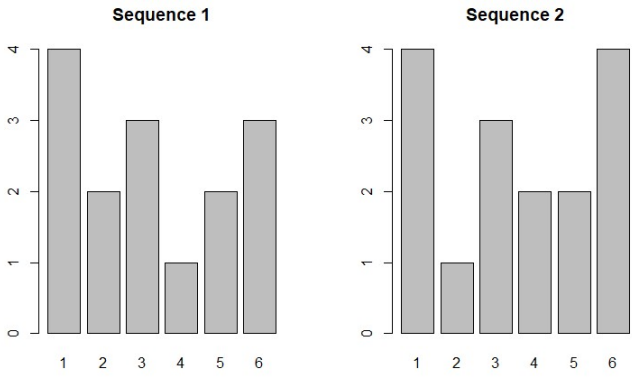


Figure B.2: Two histograms reflecting the states (1,2,3,4) at $N = 6$ bins.

In Figure B.2 it can be seen that the first sequence consists of states $(4, 2, 3, 1, 2, 3)$, while the second sequence equals $(4, 1, 3, 2, 2, 4)$. For each histogram, a cumulative distribution can be generated by taking the cumulative sums of the states at the bins. Given the cumulative distributions, the Wasserstein distance is defined as the sum of the absolute difference in the cumulative sums of the two sequences.

$$W = \sum_{i=1}^{N} |\text{cumsum}_i(s^{(1)}) - \text{cumsum}_i(s^{(2)})|,$$

where cumsum_i is the cumulative sum until element i in the sequence. Note that this definition coincides with the definition of the discrete Wasserstein distance introduced above, where the set $\mathcal{X}$ is equal to the N bins on which the two tumors are compared and the cumulative distributions are the cumulative sums of the states at the bins.

In the example of Figure B.2, the cumulative sum of sequence 1 equals $(4, 6, 9, 10, 12, 15)$ and the cumulative sum of sequence 2 equals $(4, 5, 8, 10, 12, 16)$ so that the Wasserstein distance is equal to:

$$W = |4-4| + |6-5| + |9-8| + |10-10| + |12-12| + |15-16| = |0| + |1| + |1| + |0| + |0| + |-1| = 3.$$

The Wasserstein distance thus computes the discrepancy in the two profiles by means of taking the sum of the absolute difference in the cumulative sum of the two histograms which reflect the states of the bins. It should be noted, however, that not all bins on the genome should be compared simultaneously as this could possibly give a misleading outcome. Instead, the Wasserstein distance needs to be applied on the output per chromosome arm as the copy number changes between chromosome arms are assumed to appear independently from one another. If the Wasserstein distance would be applied on all the bins together, not taking into account the different chromosome arms, a larger Wasserstein distance could arise even though the profiles might be quite similar therefore decreasing the performance of the comparison method. Define

$$W_j = \sum_{i=1}^{N_j} |\text{cumsum}_i(s^{(1)}) - \text{cumsum}_i(s^{(2)})|$$

as the Wasserstein distance for the comparison of two samples on chromosome arm j which consists of N_j bins. The total Wasserstein distance evaluated at all chromosome arms equals:

$$W = \sum_{j=1}^{M} W_j,$$

where M is the number of chromosome arms on which the two tumors are compared.

The Wasserstein distance is applied on this problem by means of transforming the states of the bins

into a cumulative distribution, where the cumulative distribution is constructed by taking the cumulative sum of the states. Note that, as explained in Section 4.4.3, the Wasserstein distance sometimes incorrectly assigns a lower score to a pair that has more differences. This has to do with the fact that the total sum of the states is not the same for each tumor. In other words, the cumulative distribution is not truly a cumulative distribution. In the example in Figure B.2, the total sum of sequence 1 is equal to 15, while the second sequence has a total sum of 16. As the Wasserstein distance demands that each distribution has the same amount of mass, the Wasserstein distance as introduced above is not correctly defined. As a consequence, it incorrectly states that a pair with more differences is more clonal. In order to solve this, a normalization may be applied. For this, the cumulative sums are divided by the total sum of all the states. As a consequence, each vector will add up to 1 after normalization. However, normalizing the states is not a clever idea as the Wasserstein distance can produce misleading outcomes. For example, consider the following two sequences:

$$(2, 4, 4, 4, 2) \text{ and } (1, 2, 2, 2, 1).$$

When the sequences are normalized, i.e. each element is divided by the total sum of the states, the following sequences occur:

$$\left(\frac{2}{16}, \frac{4}{16}, \frac{4}{16}, \frac{4}{16}, \frac{2}{16}\right) \text{ and } \left(\frac{1}{8}, \frac{2}{8}, \frac{2}{8}, \frac{2}{8}, \frac{1}{8}\right).$$

Applying the Wasserstein distance on these two sequences would result in a distance of zero, while the sequences are obviously different from one another. Therefore, normalization is not recommended for the Wasserstein distance.

Since the states do not add up to the same number for each tumor and normalization can not be applied as well, it can be concluded that a true cumulative distribution can not be constructed. Hence, the Wasserstein distance can not be applied for this purpose.

C

Scoring functions

This appendix introduces several scoring functions that may be applied to determine the performance of the comparison methods. For each judgment, the experts also provided a certainty quantification, i.e. how certain are they that their judgment is correct. If, for example, the certainty quantification of a pair is equal to 0.8 and the pair is labeled as clonal, then the expert is 80% certain that this pair is clonal. As each expert provides a certainty quantification for each pair, the certainty quantifications are averaged to come to a final estimate. Five different scoring functions have been investigated:

1. Score 1: equals 1 for a pair if the comparison method and expert judgment agree (both state that the pair is clonal or both state the pair is not clonal) and -1 if there is disagreement.

2. Score 2: is a weighted average of score 1. The certainty quantifications of the two experts are averaged and incorporated in the score as follows: if c_j is the average certainty of the judgment of pair j, then score 2 of pair j, score_j^2 equals: $\text{score}_j^2 = c_j \text{score}_j^1$. For example, if the average certainty for a pair is equal to 0.7 and the pair is incorrectly labeled, score 2 will equal -0.7.

3. Score 3: is similar to score 1, but also considers if a p-value is close to being significant. If a pair is incorrectly classified as not clonal, but the p-value is below 0.1 the score of the pair increases from -1 to -0.5. On the other hand, if a pair is correctly classified as not clonal but the p-value is below 0.1 the score decreases from 1 to 0.5.

4. Score 4: is a weighted average of score 3 in the same way as score 2 where the average certainty quantification is incorporated as well. For example, if a pair with certainty quantification 0.8 is incorrectly labeled as not clonal but the p-value is below 0.1, the score of the pair is equal to $-0.5 \cdot 0.8 = -0.4$.

5. Score 5: is equal to the total number of clonal pairs correctly detected by the comparison method. If the p-value of a clonal pair is below 0.05, a score of 1 is applied. If the p-value of a clonal pair is within the interval $(0.05, 0.1)$, the pair is assigned a score of 0.5. A score of 0 is given to a clonal pair if the p-value is above 0.1. The correctness of the classification of the non-clonal pairs are not considered in score 5.

D

Scores of the 23 fresh frozen pairs

This appendix shows the scores and p-values of the comparison methods for the 23 fresh frozen pairs which were used as input for the decision tree. The coloring of the pairs reflects whether the pair was judged as clonal (green), independent (red) or ambiguous (black) by the experts. If the p-value of the comparison score is below 0.05, the cell is colored green.

Pair	Log LR	p-value Log LR	Adapted SI	p-value Adapted SI
FF-2	0.9352	0.1180	0.6462	0.0228
FF-5	1.1674	0.6128	0.2939	0.2029
FF-6	0.8445	0.0145	0.0757	0.7495
FF-7	1.0666	0.3892	0.3575	0.1408
FF-10	0.8162	0.0021	0.6949	0.0124
FF-11	0.9642	0.1594	0.6318	0.0290
FF-13	1.0273	0.2961	0.2493	0.2588
FF-14	1.1593	0.5983	0.1082	0.6211
FF-17	0.9478	0.1325	0.2150	0.3292
FF-18	1.1803	0.6418	0.2818	0.2153
FF-19	0.8549	0.0269	0.1413	0.5155
FF-20	0.6455	0.0000	0.9954	0.0000
FF-23	1.1607	0.6004	0.1723	0.4286
FF-25	0.7807	0.0000	0.8593	0.0000
FF-27	1.0127	0.2650	0.3214	0.1594
FF-29	1.0472	0.3478	0.4882	0.0642
FF-30	0.9041	0.0807	0.2023	0.3582
FF-31A	0.8002	0.0000	0.8412	0.0041
FF-31B	1.2487	0.8841	0.1285	0.5611
FF-32	0.9621	0.1594	0.6188	0.0311
FF-33	1.2555	0.9068	0.0022	0.9752
FF-35	0.8727	0.0373	0.1233	0.5756
FF-36	0.9354	0.1180	0.3170	0.1636

Table D.1: Scores and p-values of the comparison methods using ASCAT where the segmentation output is corrected as described in Section 6.2.4.

Pair	Log LR	p-value Log LR	Adapted SI	p-value Adapted SI
FF-2	1.1928	0.6687	0.1983	0.3478
FF-5	1.0545	0.3395	0.4280	0.0745
FF-6	0.8647	0.0207	0.1538	0.4617
FF-7	1.1468	0.5445	0.1717	0.4182
FF-10	0.9377	0.1284	0.4775	0.0393
FF-11	0.9815	0.2153	0.5292	0.0104
FF-13	0.9615	0.1698	0.1687	0.4244
FF-14	1.0854	0.4120	0.3766	0.1180
FF-17	0.9193	0.0932	0.1047	0.6211
FF-18	1.2078	0.7122	0.2546	0.2505
FF-19	0.8406	0.0062	0.1549	0.4555
FF-20	0.6621	0.0000	1.0000	0.0000
FF-23	1.0750	0.3892	0.0026	0.9358
FF-25	0.8627	0.0166	0.6836	0.0000
FF-27	1.0142	0.2505	0.4984	0.0248
FF-29	1.0448	0.3209	0.4873	0.0352
FF-30	0.9398	0.1304	0.1002	0.6460
FF-31A	0.7812	0.0000	0.8404	0.0000
FF-31B	1.2429	0.8923	0.1159	0.5797
FF-32	1.1239	0.4824	0.3427	0.1408
FF-33	1.1557	0.5673	0.0033	0.9296
FF-35	0.9620	0.1698	0.1120	0.5921
FF-36	0.8621	0.0166	0.4159	0.0911

Table D.2: Scores and p-values of the comparison methods using OncoSNP.

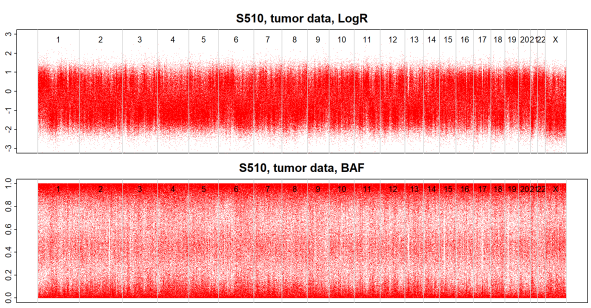
E

Formalin-Fixed Paraffin-Embedded data

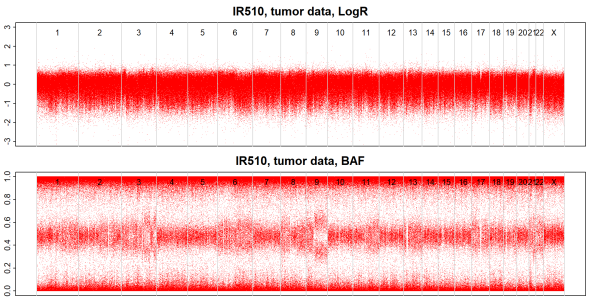
An experiment with a small set of FFPE samples was conducted to discover whether the segmentation algorithms can also be applied on FFPE data. For this, three different tumors coming from the years 1996, 2003 and 2010 were used. The DNA of each tumor was isolated using three different isolation methods which are encoded as 3, 5 and 8 respectively. For instance, tumor 310 corresponds to the tumor from 2010 where the DNA isolation method encoded as 3 was used to isolate the DNA. The FFPE dataset consists of 9 samples in total.

Before the DNA is run on the SNP chip, the quality of the DNA is increased using a DNA restoration kit. It is important to choose the correct DNA restoration kit to ensure that the data is of sufficient quality. The first restoration kit which was applied to the DNA of the samples was the NEBNext restoration kit. This kit unfortunately did not produce desirable results: the LRR and BAF plots still showed too much noise. In other words, the profiles were still too noisy making it impossible for the segmentation algorithms to determine the copy number aberrant regions. In order to solve this, a second, more expensive DNA restoration kit, the Illumina FFPE QC and DNA kit, was used on the DNA of the samples. This restoration kit improved the DNA quality of the samples significantly as the resulting raw data profiles were less noisy and showed more clear patterns. Note that the 9 FFPE samples were analyzed on the Illumina GSA v3 chip for both restoration kits.

Figure E.1 shows the raw data plots of tumor 510 before and after DNA restoration, where the DNA was restored with the Illumina kit. Tumor 510 comes from 2010 and its DNA is isolated using the DNA isolation method which is encoded as 5.



(a) Raw data profile (LRR and BAF values) of tumor 510 (tumor coming from 2010, DNA isolation method encoded as 5) before DNA restoration.



(b) Raw data profile (LRR and BAF values) of tumor 510 (tumor coming from 2010, DNA isolation method encoded as 5) whose DNA was restored using the Illumina FFPE QC and DNA restoration kit.

Figure E.1: Raw data plots of tumor 510 (tumor coming from 2010, DNA isolation method encoded as 5) whose DNA was restored with the NEBNext (left panel) and the Illumina FFPE QC and DNA (right panel) restoration kits.

Comparing the two plots in Figure E.1, it can be seen that after DNA restoration, a clearer pattern becomes visible in the BAF plot and less noise appears to be present in the LRR data. However, even after DNA restoration, the raw data profiles of the FFPE data still show much more noise than the raw

data profiles of fresh frozen samples. Nevertheless, by means of applying the correct DNA restoration kit the amount of noise in the raw data can be significantly reduced so that the segmentation algorithms may still be applied.

Even though the Illumina restoration kit improves the DNA quality, the LRR data does not show clear peaks in Figure E.1b. It may therefore be questioned whether the segmentation algorithms are capable to reliably determine aberrant copy number events from the raw data. Gains are generally discernible from the BAF plots: if a region has more than three bands in the BAF plot, then it is highly likely that a gain has occurred in the region. The main problem for the segmentation algorithms lies in discriminating loss regions from CN-LOH regions [17]. For both events, the BAF values center around 0 and 1, but for a loss region the LRR data lies significantly lower than for a CN-LOH region. If the LRR data shows a flat profile, the segmentation algorithm may incorrectly detect a loss as a CN-LOH region or the other way around.

ASCAT was applied on the 9 Illumina restored FFPE samples. Of the 9 samples, 4 samples had a goodness of fit below 80%. These samples included all three samples from 1996 and sample 310. Note that sample 310 had one of the highest genotype call rates in GenomeStudio. This shows that a high call rate does not necessarily imply that ASCAT can construct a segmentation profile. When the minimal goodness of fit was lowered to 70%, an ASCAT profile could be constructed for all the failed samples. However, the sunrise plots of the samples coming from 1996 did not show clear minima. As a result, the estimated ploidies of the three samples from 1996 were completely different: a ploidy of 1.68 was found for sample 396, a ploidy of 3.82 for sample 596 and sample 896 had a ploidy of 2.49. In the most ideal case, the three ploidy numbers coming from the same tumor are similar. For instance, the estimated ploidy numbers for samples 303, 503 and 803 are 2.13, 2.18 and 2.11 respectively. Note that the estimated ploidy number of sample 310 was completely different than the estimated ploidy numbers of samples 510 and 810. This indicates that the segmentation profiles of samples with a goodness of below 80% are less reliable and should therefore be interpreted with care.

As the LRR data shows a relatively flat profile, it can be questioned whether ASCAT can detect copy number aberrant events from the raw data. In order to examine this, the LRR data was smoothed using a moving average approach. For each chromosome arm, a window containing 25 SNPs was moved over the LRR data and the average of each window computed, i.e. the first window consists of SNPs 1 to 25, the second window of SNPs 2 to 26 etc. Note that the BAF data can not be smoothed using a moving average approach as the structure will be lost. By means of smoothing the LRR data, gain and loss peaks will become more apparent and can be used to determine whether ASCAT can recognize copy number aberrant events from the data. Figure E.2 shows the LRR data of tumor 510 before and after smoothing.

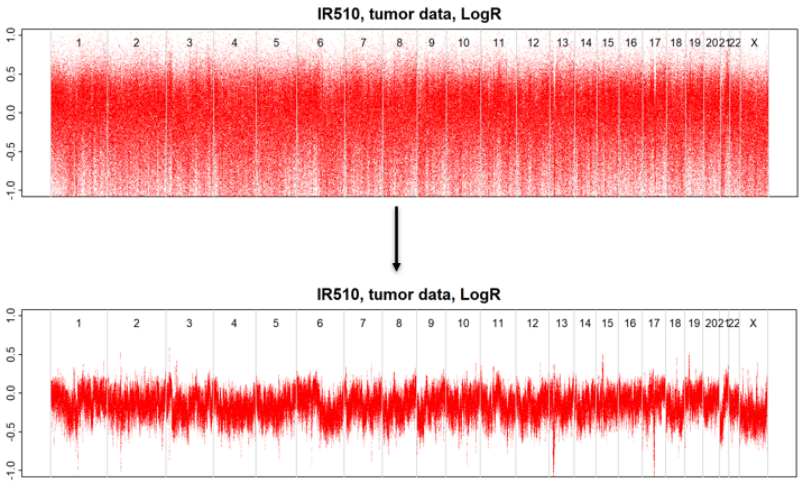


Figure E.2: LRR data of tumor 510 before (top) and after (bottom) smoothing by means of using a moving average window approach.

Given the smoothed LRR data and original BAF data, the performance of ASCAT can be examined for each sample by means of testing whether the algorithm can detect copy number zero and high copy number regions. Copy number zero regions are characterized by very low LRR values and BAF data that can range anywhere from 0 to 1. For example, the smoothed LRR data of tumor 510 in Figure E.2 shows that a double loss occurs at chromosomes 13 and 17. High copy number regions can be detected by high LRR values and multiple bands in the BAF plot. The smoothed LRR data of tumor 510 shows high copy number regions on chromosomes 2, 3, 15 and 19.

The copy number zero and high copy number regions are used as a guideline to determine the performance of ASCAT. For this, only the FFPE samples which have a goodness of fit above 80% are used. Mixed performances were found: for some samples the copy number zero and high copy number regions were almost always detected by ASCAT while for other samples ASCAT sometimes failed to recognize these regions. Figure E.3 shows the smoothed LRR data, original BAF data and the ASCAT profile of sample 810 for which the performance of ASCAT was not that good.

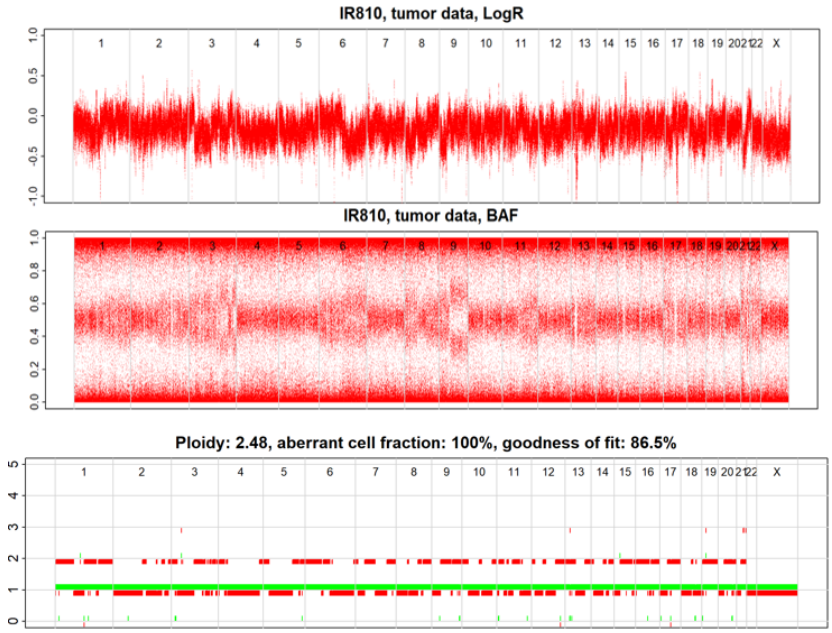


Figure E.3: LRR data (smoothed), BAF data and ASCAT profile of sample 810.

The raw data of tumor 810 shows that there are two clear copy number zero regions: one at chromosome 13 and one at chromosome 17. ASCAT manages to detect the copy number zero region at chromosome 17, but assigns copy number one for the region on chromosome 13. Next to that, the high copy number region on chromosome 15 is not observed by ASCAT. For other samples, on the other hand, ASCAT performed very well. Similar results were obtained for OncoSNP. However, the raw data may be misleading as well: smoothing the LRR gives an indication of where gains and losses have occurred but this is not always the ground truth. Therefore, we should be careful with judging the performances of the segmentation algorithms.

As mentioned in [17], the main problem when segmenting the profiles of FFPE samples lies in the fact that loss and CN-LOH regions are hard to differentiate from one another. In this dataset, both loss and CN-LOH events did not occur that often. As a result, it is hard to tell whether ASCAT or OncoSNP is capable to discern between the two events. Hence, a conclusion about this matter can not be derived. Nevertheless, this fact should be taken into account when the model is being tested on real FFPE pairs.

In order to test the comparison methods, two ”pairs” can be constructed by means of pairing samples coming from the same tumor. For example, tumors 303 and 503 and 510 and 810 can be paired with one another. Moreover, two independent pairs can also be constructed by pairing different tumors

from the same DNA isolation method with one another, e.g. tumor 310 and 396 can be paired. A small test was run with 4 pairs of which two were ”clonal”, i.e. coming from the same tumor, and two were independent. The clonal pairs were found to have the smallest scores for both comparison methods, but the p-values were far from significant using a 0.05 significance level. This has to do with the fact that the sample size is just too small to get a significant p-value. Next to that, the small sample size also causes dependent observations to be introduced in the independence distribution as samples coming from the same patients are compared. For instance, the comparison score of samples 510 and 310 can be included in the independence distribution, even though the two samples are coming from the same tumor. This yields more extreme observations in the independence distribution so that the p-value increases as well.

In conclusion, the segmentation algorithms appear to perform decently on the FFPE samples provided that the correct DNA restoration kit is used. As the dataset does not consist of any pairs and is too small to form artificial pairs, the performance of the comparison methods remains to be tested.

Bibliography

[1] Edward F Attiyeh et al. “Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy”. In: *Genome research* 19.2 (2009), pp. 276–283.

[2] B Banelli et al. “Pathological and molecular characteristics distinguishing contralateral metastatic from new primary breast cancer”. In: *Annals of oncology* 21.6 (2010), pp. 1237–1242.

[3] Colin B Begg et al. “Contralateral breast cancers: Independent cancers or metastases?” In: *International journal of cancer* 142.2 (2018), pp. 347–356.

[4] Aditya Vijay Bhagwate et al. “Bioinformatics and DNA-extraction strategies to reliably detect genetic variants from FFPE breast tissue samples”. In: *BMC genomics* 20.1 (2019), pp. 1–10.

[5] Jana Biermann et al. “Clonal relatedness in tumour pairs of breast cancer patients”. In: *Breast Cancer Research* 20.1 (2018), pp. 1–16.

[6] Marc A Bollet et al. “High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers”. In: *JNCI: Journal of the National Cancer Institute* 100.1 (2008), pp. 48–58.

[7] Breastcancer.org. *Breast Cancer Risk Factors: Genetics*. 2021. URL: `https://www.breastcancer.org/risk/factors/genetics`.

[8] Rong Chen et al. “Receptor conversion in metastatic breast cancer: analysis of 390 cases from a single institution”. In: *Modern Pathology* 33.12 (2020), pp. 2499–2506.

[9] Yue Chen et al. “Epidemiology of contralateral breast cancer”. In: *Cancer Epidemiology and Prevention Biomarkers* 8.10 (1999), pp. 855–861.

[10] Mayo Clinic. *Breast cancer: diagnosis and treatment*. 2021. URL: `https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475`.

[11] Jacob Cohen. “A coefficient of agreement for nominal scales”. In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[12] Stefano Colella et al. “QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data”. In: *Nucleic acids research* 35.6 (2007), pp. 2013–2025.

[13] Hamid Eghbal-Zadeh et al. “DeepSNP: an end-to-end deep neural network with attention-based localization for breakpoint detection in single-nucleotide polymorphism array genomic data”. In: *Journal of Computational Biology* 26.6 (2019), pp. 572–596.

[14] Jacques Ferlay et al. “Global burden of breast cancer”. In: *Breast cancer epidemiology*. Springer, 2010, pp. 1–19.

[15] R.A. Fisher. *The design of experiments. 1935*. Edinburgh: Oliver and Boyd, 1935.

[16] RA Fisher. *Statistical Methods for Research Workers; Oliver, Boyd, eds*. 1932.

[17] AN Hosein et al. “The use of the Illumina FFPE Restoration Protocol to obtain suitable quality DNA for SNP-based CGH–a pilot study”. In: *Hereditary Cancer in Clinical Practice*. Vol. 10. 2. Springer. 2012, pp. 1–1.

[18] Tao Huang et al. “Detection of DNA copy number alterations using penalized least squares regression”. In: *Bioinformatics* 21.20 (2005), pp. 3811–3817.

[19] Philippe Hupé et al. “Analysis of array CGH data: from signal ratio to gain and loss of DNA regions”. In: *Bioinformatics* 20.18 (2004), pp. 3413–3422.

[20] Daisuke Iizuka et al. “DNA copy number aberrations and disruption of the p16Ink4a/Rb pathway in radiation-induced and spontaneous rat mammary carcinomas”. In: *Radiation research* 174.2 (2010), pp. 206–215.

[21] Illumina. ""TOP/BOT" strand and "A/B" allele: A guide to Illumina's method for determining strand and allele for the GoldenGate and Infinium assays". In: *Technical note* (2006).

[22] Illumina. "DNA Copy Number and Loss of Heterozygosity Analysis Algorithms". In: *Technical note* (2017).

[23] Tatsuhiko Imaoka et al. "Radiation-induced mammary carcinogenesis in rodent models: What's different from chemical carcinogenesis?" In: *Journal of radiation research* 50.4 (2009), pp. 281–293.

[24] National Human Genome Research Institute. *Gene*. 2021. URL: https://www.genome.gov/genetics-glossary/Gene.

[25] Christopher I Li et al. "Trends in incidence rates of invasive lobular and ductal breast carcinoma". In: *Jama* 289.11 (2003), pp. 1421–1424.

[26] Esther H Lips et al. "Genomic profiling defines variable clonal relatedness between invasive breast cancer and primary ductal carcinoma in situ". In: *medRxiv* (2021).

[27] Tamás Lipták. "On the combination of independent tests". In: *Magyar Tud Akad Mat Kutato Int Kozl* 3 (1958), pp. 171–197.

[28] Daniel S May and Nancy E Stroup. "The incidence of sarcomas of the breast among women in the United States, 1973–1986". In: *Plastic and reconstructive surgery* 87.1 (1991), p. 193.

[29] Craig H Mermel et al. "GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers". In: *Genome biology* 12.4 (2011), pp. 1–14.

[30] Frederick Mosteller and Robert R Bush. *Selected quantitative techniques*. Addison-Wesley, 1954.

[31] Szilárd Nemes et al. "A diagnostic algorithm to identify paired tumors with clonal origin". In: *Genes, Chromosomes and Cancer* 52.11 (2013), pp. 1007–1016.

[32] Yue S Niu and Heping Zhang. "The screening and ranking algorithm to detect DNA copy number variations". In: *The annals of applied statistics* 6.3 (2012), p. 1306.

[33] Adam B Olshen et al. "Circular binary segmentation for the analysis of array-based DNA copy number data". In: *Biostatistics* 5.4 (2004), pp. 557–572.

[34] World Health Organization. *Cancer*. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/cancer.

[35] Irina Ostrovnaya and Colin B Begg. "Testing clonal relatedness of tumors using array comparative genomic hybridization: a statistical challenge". In: *Clinical Cancer Research* 16.5 (2010), pp. 1358–1367.

[36] Irina Ostrovnaya et al. "A metastasis or a second independent cancer? Evaluating the clonal origin of tumors using array copy number data". In: *Statistics in medicine* 29.15 (2010), pp. 1608–1621.

[37] Victor M Panaretos and Yoav Zemel. "Statistical aspects of Wasserstein distances". In: *Annual review of statistics and its application* 6 (2019), pp. 405–431.

[38] Daniel A Peiffer et al. "High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping". In: *Genome research* 16.9 (2006), pp. 1136–1148.

[39] Adriana Pitea et al. "Copy number aberrations from Affymetrix SNP 6.0 genotyping data—how accurate are commonly used prediction approaches?" In: *Briefings in bioinformatics* 21.1 (2020), pp. 272–281.

[40] Edwin JG Pitman. "Significance tests which may be applied to samples from any populations". In: *Supplement to the Journal of the Royal Statistical Society* 4.1 (1937), pp. 119–130.

[41] Pink Ribbon. *Borstkanker: cijfers en feiten*. 2021. URL: https://www.pinkribbon.nl/over-borstkanker/cijfers-en-feiten.html.

[42] Ganesh N Sharma et al. "Various types and management of breast cancer: an overview". In: *Journal of advanced pharmaceutical technology & research* 1.2 (2010), p. 109.

[43] Sarah Song et al. “qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles”. In: (2012).

[44] Samuel A Stouffer et al. “The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1”. In: (1949).

[45] Wei Sun et al. “Integrated study of copy number states and genotype calls using high-density SNP arrays”. In: *Nucleic acids research* 37.16 (2009), pp. 5365–5377.

[46] Cancer Research UK. *How cancer starts*. 2020. URL: `https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts`.

[47] Mark A Van De Wiel et al. “CGHcall: calling aberrations for array CGH tumor profiles”. In: *Bioinformatics* 23.7 (2007), pp. 892–894.

[48] Peter Van Loo et al. “Allele-specific copy number analysis of tumors”. In: *Proceedings of the National Academy of Sciences* 107.39 (2010), pp. 16910–16915.

[49] ES Venkatraman and Adam B Olshen. “A faster circular binary segmentation algorithm for the analysis of array CGH data”. In: *Bioinformatics* 23.6 (2007), pp. 657–663.

[50] Andrew Viterbi. “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. In: *IEEE transactions on Information Theory* 13.2 (1967), pp. 260–269.

[51] Vladimir Vovk and Ruodu Wang. “Combining e-values and p-values”. In: *arXiv preprint arXiv:1912.06116* 3 (2019).

[52] Frederic M Waldman et al. “Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences”. In: *Journal of the National Cancer Institute* 92.4 (2000), pp. 313–320.

[53] Kai Wang et al. “PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data”. In: *Genome research* 17.11 (2007), pp. 1665–1674.

[54] Michael C Whitlock. “Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach”. In: *Journal of evolutionary biology* 18.5 (2005), pp. 1368–1373.

[55] Daniel J Wilson. “The harmonic mean p-value for combining dependent tests”. In: *Proceedings of the National Academy of Sciences* 116.4 (2019), pp. 1195–1200.

[56] Yi-Ching Yao. “Estimating the number of change-points via Schwarz’criterion”. In: *Statistics & Probability Letters* 6.3 (1988), pp. 181–189.

[57] Christopher Yau et al. “A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data”. In: *Genome biology* 11.9 (2010), pp. 1–15.