Development of a Tool for Copy Number Analysis of Cancer Genomes using High Throughput Sequencing Data

Master's Thesis
Ebrahim Afyounian
Institute of Biosciences and Medical
Technology (BioMediTech)
University of Tampere
June 2015

Acknowledgement

This work was conducted at the Computational Biology group led by Professor Matti Nykter at the Institute of Biosciences and Medical Technology (BioMediTech), University of Tampere. I would like to thank Professor Matti Nykter for giving me the opportunity to conduct this work at the Computational Biology group and his invaluable feedback.

Never would this study have been realized, had it not been for the excellent ideas and feedback I received from my thesis advisor, Matti Annala. Matti Annala mentored me from the conception of the idea of this work until the last stages of finalizing it with his comprehensive knowledge of the field of bioinformatics. I would like to thank Matti Annala for his support.

Finally, I would like to thank my family, especially my second brother Mohammad-Reza and Tiina who have always been a source of motivation and strength to me.

Tampere, June 5, 2015

Ebrahim Afyounian

Master's Thesis

Place:	University of Tampere Institute of Biosciences and Medical Technology (BioMediTech)
Author:	Ebrahim Afyounian
Title:	Development of a Tool for Copy Number Analysis of Cancer Genomes using High Throughput Sequencing Data
Pages:	70
Supervisor:	Matti Annala
Reviewers:	Professor Matti Nykter, Juha Kesseli
Date:	June 2015

Abstract

Genomic copy number alterations (CNA) and loss of heterozygozity (LOH) are two types of genomic instabilities associated with cancer. Acquisition of these genomic instabilities affects the expression level of oncogenes and tumor suppressor genes. Thus, accurate detection of these abnormalities is a crucial step in identifying novel oncogenes and tumor suppressor genes. Whole-genome sequencing of tumor tissues has enabled new opportunities for the detection of such aberrations and the characterization of genomic aberrations in tumor samples.

In this work, a fast tool for the identification of CNAs and copy-neutral LOH in tumor samples using whole-genome sequencing data was developed. The developed tool segments the genome by analyzing the read-depth and B-allele fraction profiles using a double sliding window method. It requires a matched normal sample to correct for biases such as GC-content and mapability and to discriminate somatic from germline events. The developed tool was evaluated on both simulated and real whole-genome sequencing data against competing, state of the art tools to demonstrate its accuracy. The tool, written in the Python programming language, is fast and performs segmentation of a whole genome in less than two minutes.

Table of Contents

List of Abbreviations

BAM. Binary Alignment Map

CBS. Circular Binary Segmentation

CGHub. Cancer Genomics Hub

CNA. Copy Number Alteration

CNV. Copy Number Variation

DNA. Deoxyribonucleic Acid

FDR. False Discovery Ratio

HMM. Hidden Markov Model

HTS. High Throughput Sequencing

IARC. International Agency for Research on Cancer

IGV. Integrative Genomics Viewer

LASSO. Least Absolute Shrinkage eStimatOr

LOH. Loss Of Heterozygozity

NAN. Not A Number

PCR. Polymerase Chain Reaction

PRAD. PRostate ADenocarcinoma

SAM. Sequence Alignment/Map

SNP. Single Nucleotide Polymorphism

TCGA. The Cancer Genome Atlas

TNBC. Triple-Negative Breast Cancer

WES. Whole Exome Sequencing

WGS. Whole Genome Sequencing

List of Figures

List of Tables

1. Introduction

Cancer is a major human health problem and a leading cause of death worldwide. Any attempt towards the prevention or treatment of cancer requires a deep understanding of the processes underlying the initiation, development and dissemination of the cancer. It is widely accepted that cancer is a disease subject to Darwinian evolution. Cancer arises from a single cell by the introduction of a single mutation. The mutation, if selected naturally, is passed to the daughter cells through cell division. As time passes, more and more mutations occur and accumulate. This results in heterogeneous and genetically diverse populations of cancer cells (Nowell 1976, Marusyk, Polyak 2010). Mutations are divided into two groups depending on the cell type in which they occur. If a mutation occurs in reproductive cells (eggs and sperms), it is called a germ-line mutation. Germ-line mutations can pass on to the next generation. In contrast, if a mutation occurs in non-reproductive cells, it is called a somatic mutation. Somatic mutations affect only the individual in which they occur and thus they are not transmitted to future generations.

In the past several years, advances in sequencing technology such as the introduction of high throughput sequencing (HTS) have enabled more in depth study of cancer in terms of detecting genomic alterations. HTS technologies allow determining the exact order of nucleotides present in a given DNA or RNA molecule much more quickly and cheaply than older technologies such as Sanger sequencing. As a result, they facilitate efficient and economic genome-wide production of huge amount of data on DNA, mRNA and epigenetic level. HTS data has been used to identify genomic alterations such as nucleotide substitutions, small insertions and deletions, copy number alterations and chromosomal rearrangements in cancer cells. As a result, the knowledge of the underlying mechanisms of cancer has substantially increased.

Copy number alterations (CNA) and loss of heterozygosity (LOH) events are two types of genomic alterations associated with cancer. Copy number is defined as the number of copies per cell of a particular gene or other DNA sequence. Normally, diploid organisms such as humans have two copies of each autosomal DNA sequence. A CNA is an event where one copy of a genomic region is either gained or lost. CNA denotes somatically acquired copy number changes in the genome. In contrast, copy number variation (CNV) denotes the differences among individuals in the number of copies of a region of the genome. It is now known that the CNAs affect the expression level of genes (Curtis, Shah et al. 2012). Consequently, these differences in expression levels drive differences in clinical behavior of cancer and in susceptibility to treatments (Yakhini, Jurisica 2011).

LOH is an event in which one of two alleles at a heterozygous locus is lost due to segmental aneuploidies or some other mechanisms such as gene conversion, mitotic recombination, and mitotic nondisjunction (Ha, Roth et al. 2012). These events can be identified by examining the heterozygous alleles in normal cells that have become homozygous in the tumor cells. LOH events have been observed in many types of cancer at genomic regions of tumor suppressor genes such as *PTEN*, *RB1* and *TP53*. Genomic allelic losses caused by LOH events can affect the expression of genes. For instance, monoallelic expression can arise as a result of genomic allelic loss via LOH events (Ha, Roth et al. 2012). Monoallelic expression is the expression of a gene from only one of two alleles in a diploid organism. This can occur, for example, when the wild-type allele is replaced by the mutant allele. The altered expression of genes with allelic imbalance due to a LOH event may bring about selective advantages for tumorigenesis and progression. By analyzing a cohort of 23 triple-negative breast cancer (TNBC) patients, Ha et al. (2012) have shown that LOH is a prominent feature of TNBC somatic aberrations and modulates a significant portion of the transcriptome in the form of monoallelic expression (Ha, Roth et al. 2012).

Hence, detection and characterization of events such as CNAs and LOH are crucial in learning more about the fundamental mechanisms of cancer, identification of prospective cancer related genes and eventually in prevention and treatment of cancer.

Traditionally, cytogenetic technologies such as fluorescence in situ hybridization (FISH) were used to identify copy number variations (CNVs). In 1992, Kallioniemi et al. introduced comparative genomic hybridization (CGH) for detecting CNAs in tumor samples (Kallioniemi, Kallioniemi et al. 1992). The introduction of array-based technologies such as arrayCGH (or aCGH) made the genome-wide detection of such aberrations possible. However, these methods were unable to detect copy-neutral LOH events. Single-nucleotide polymorphism (SNP) arrays overcame this limitation by providing the ability to identify both CNAs and LOH (Speicher, Carter 2005). SNP refers to single-base difference in DNA among individuals. However, these methods suffered from certain drawbacks such as hybridization noise, limited genome coverage and low resolution (Zhao, Wang et al. 2013). The introduction of HTS such as whole exome sequencing (WES) and whole genome sequencing (WGS) has overcome certain drawbacks of the previous methods. HTS-based detection approaches have several advantages such as including higher coverage and resolution, more accurate copy numbers estimation and more precise breakpoints detection (Zhao, Wang et al. 2013). These advantages nominate HTS-based detection as an efficient approach for studying cancer genomes.

Although there are several tools available capable of detecting CNAs and LOH events, only few of them utilize WGS data for the detection of such aberrations (Yu, Liu et al. 2014, Mayrhofer, DiLorenzo et al. 2013, Boeva, Popova et al. 2012). In addition, some of these tools require the users to create configuration files and to set several parameters prior to running the analysis where the setting of some parameters might require some prior knowledge about the data. For instance, ControlFREEC (a CNA and LOH detection tool) requires the users to set a value for ploidy of the tumor (Boeva, Popova et al. 2012a). If this value is not known, user is required to run the analysis few times with different values for ploidy. Furthermore, some of the tools’ runtime can take up to few hours.

Thus, the main idea of this thesis is to develop a fast and easy-to-use tool that utilizes WGS data resulting from sequencing of tumor and normal samples in order to detect and segment regions of the genome with CNAs and copy-neutral LOH events. In this thesis, the implementation of this tool is explained.

In chapter 2, the biological and technical background required for understanding this thesis will be explained. The chapter continues with a literature review providing an overview of the methods used to perform somatic copy number analysis and their advantages and disadvantages.

Chapter 3 explains the main research objectives of this research. Furthermore, steps taken to fulfill the main objective of this research will be explained in this chapter.

Chapter 4 explains the materials and methods used to implement the tool. Furthermore, in this chapter, the implementation of a simulator capable of simulating the normal and the cancer genome with events such as deletion, amplification and copy-neutral LOH is explained.

In chapter 5, first, the results from the analysis of some WGS tumor data using the developed tool will be visualized in integrative genomics viewer (IGV). Next, the tool is benchmarked against three other competing, somatic copy number analysis tools and the results of the benchmark will be presented.

In chapter 6, a discussion about this work will be provided, some of the limitations of the developed tool will be discussed, and the future work required to address the limitations will be described. Finally, chapter 7 concludes the thesis.

2. Review of Literature

Cancer is a major human health problem and a leading cause of death worldwide. According to the International Agency for Research on Cancer (IARC), cancer accounted for 8.2 million deaths worldwide in 2012. In the same year, 32.6 million people worldwide were living with cancer (Globocan.iarc.fr 2014). Figure 2.1 illustrates the estimated age-standardized incidence and mortality rates worldwide for both females and males. Furthermore, figure 2.2 illustrates the mortality rates caused by different types of cancer for females and males.

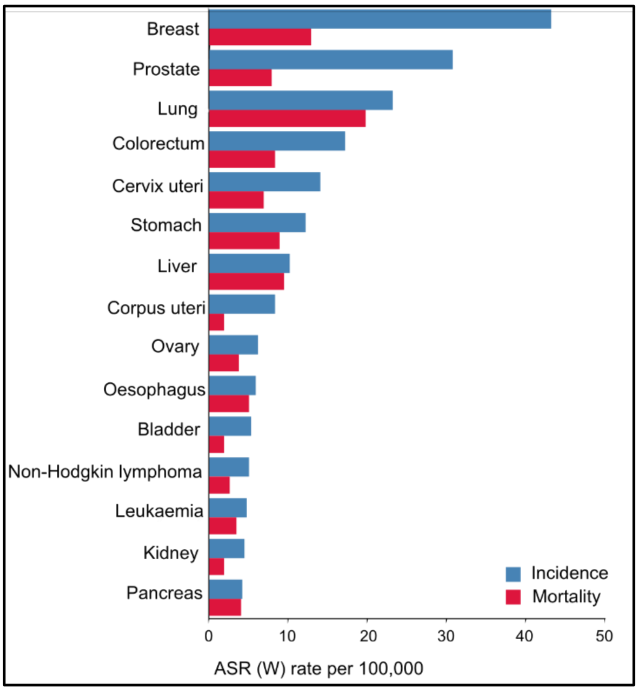

Figure 2.1 *Estimated age-standardized incidence and mortality rates worldwide for both females and males (adapted from globocan.iarc.fr 2014).*

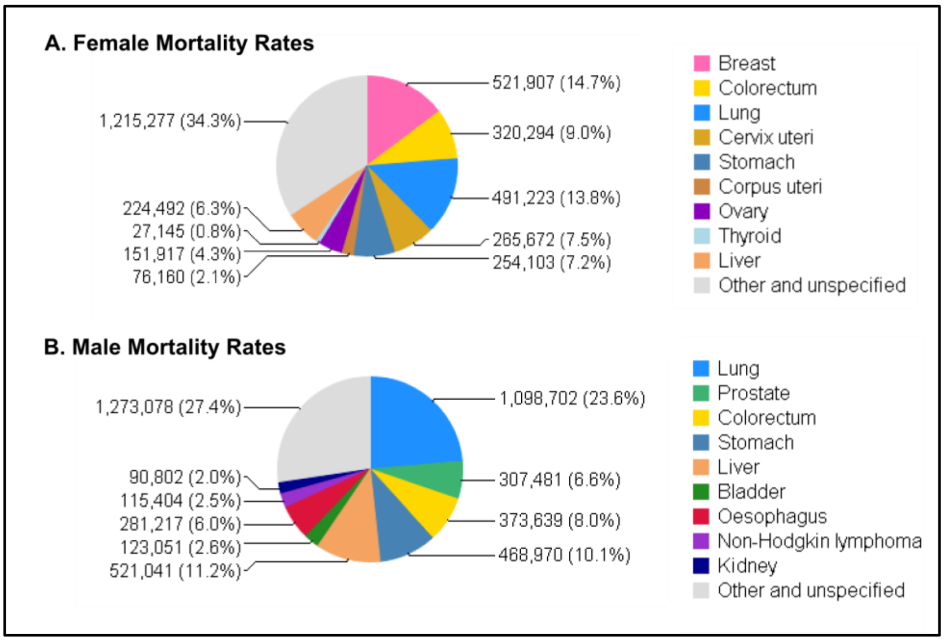

Figure 2.2 *Mortality rates caused by different types of cancer worldwide. (A) Female mortality rates (B) Male mortality rates (adapted from globocan.iarc.fr 2014).*

Over the past few decades, our understanding and knowledge of the underlying processes of cancer has expanded remarkably. It is now known that cancer arises from a single cell. Cancer can be subdivided into a variety of distinct types caused by mutated genes (Martinez, Taylor Parker et al. 2003). In 2000, Hanahan and Weinberg described the rules that govern the transformation of normal human cells into cancer cells as a multistep process. According to them, tumor development is subject to Darwinian evolution where the successive selection of deleterious genetic changes in normal cells and their passage to the daughter cells causes the accumulation of such changes and leads to progressive conversion of normal cells to cancer cells (Hanahan, Weinberg 2000). They have listed six essential alterations in cell physiology that together dictate malignant growth. These six capabilities are shared by most and perhaps all types of human tumors. They call these capabilities "the hallmarks of cancer" where hallmark refers to: "the acquired functional capabilities that allow cancer cells to survive, proliferate, and disseminate" (Hanahan, Weinberg 2000). These six hallmarks are self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. Figure 2.3 illustrates these acquired capabilities of cancer.

In 2011, Hanahan and Weinberg expanded their original model by adding four more capabilities, which are avoidance of immune destruction, deregulation of cellular energetics, genome instability and mutation, and tumor-promoting inflammation (Hanahan, Weinberg 2011). Figure 2.4 illustrates the expanded model of acquired capabilities of cancer.

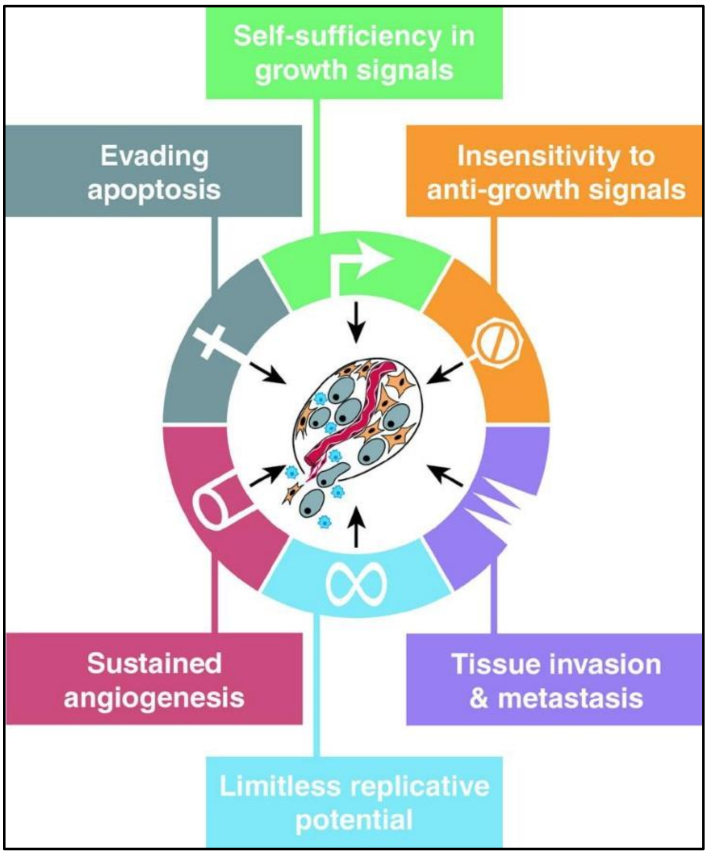
Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
Sustained angiogenesis
Tissue invasion & metastasis
Limitless replicative potential

Figure 2.3 *Acquired capabilities of cancer (adapted from Hanahan and Weinberg, 2000).*

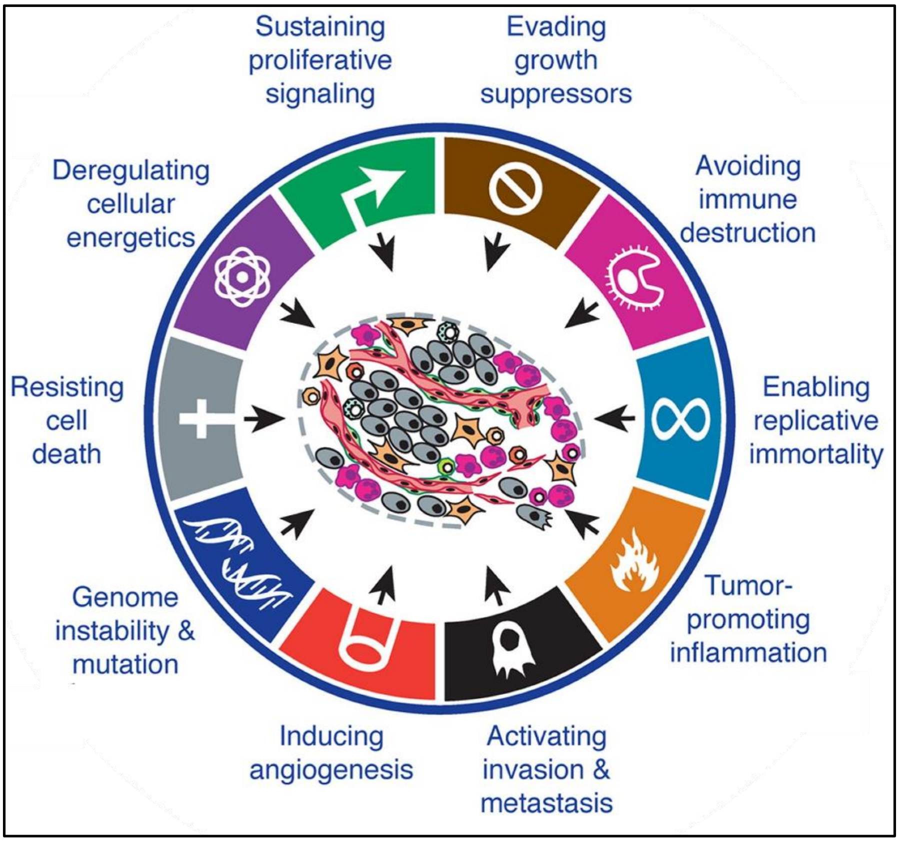
Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

Figure 2.4 *Expanded model of acquired capabilities of cancer (adapted from Hanahan and Weinberg, 2011).*

Of all these hallmarks, genome instability and mutation is the most consistent hallmark found in all cancers (Dancey, Bedard et al. 2012). According to Hanahan and Weinberg, the genome instability is the source of chromosomal rearrangements (Hanahan, Weinberg 2011).

2.1. Chromosomal Rearrangement

Chromosomal rearrangements are present in many cancers. Usually, a chromosomal rearrangement is caused by breakage of DNA double helices in the genome at two different locations. Subsequently, the broken ends are rejoined to produce a new chromosomal arrangement that is different from the original nucleotide sequence (Griffiths, Gelbart et al. 1998). However, not all the chromosomal rearrangements need breakage at two different locations. For instance, *terminal deletion* requires a single breakage followed by the capping of the broken end with a telomere (Moore, Best 2001). Chromosomal rearrangement is subdivided into several classes of events such as deletion, duplication, inversion and translocation. Figure 2.5 illustrates different classes of chromosomal rearrangement.

Deletion, a class of chromosomal rearrangement, is the result of the loss of one region of a chromosome. In a deletion event, there are two chromosome breaks where the intervening segment is lost and subsequently the two ends are rejoined. Deletion itself can be categorized into two different types: hemizygous deletion and homozygous deletion. Hemizygous deletion occurs when a diploid organism loses one allele or portion of an allele from its genome. In contrast, homozygous deletion occurs when both copies of the same allele or the same chromosomal segment of a pair of homologous chromosomes is lost (*Encyclopedic reference of cancer.* 2001). Dong (2001) describes mechanisms through which deletion can inactivate tumor suppressor genes leading to carcinogenesis. Tumor suppressor genes are mainly involved in constraining cells from uncontrolled growth and migration. As an example, hemizygous deletion induces haploinsufficiency that is a condition where the remaining functional copy of a gene (after a hemizygous deletion) is not capable of producing enough gene product to bring about a wild-type condition (Dong 2001). For instance, Cristofano et al. (1998) have demonstrated the causal role of *PTEN* haploinsufficiency in Cowden disease, Lhermitte-Duclos disease and Bannayan-Zonana syndrome (Di Cristofano, Pesce et al. 1998). The deletion event has been observed in several other cancer-related genes such as *RB1*, *CDKN2A/B*, *ARID1A*, *MAP2K4*, *NF1*, *SMAD4*, *BRCA1/2*, *MSH2/6*, *DCC* and *CDH1* (Liu, Morrison et al. 2013).

In duplication, another class of chromosomal rearrangement, an extra copy of a chromosomal region is produced. The duplication event has been observed in several cancer-related genes such as *ERBB2*, *EGFR*, *MYC*, *PIK3CA*, *IGF1R*, *FGFR1/2*, *KRAS*, *CDK4*, *CCND1*, *MDM2*, *MET*, *CDK6* (Liu, Morrison et al. 2013). As an example, the duplication of *ERBB2* oncogene, observed in 20-30% of human breast cancers, leads to the overexpression of its gene product resulting in rapid cell proliferation (Isola, Chu et al. 1999).

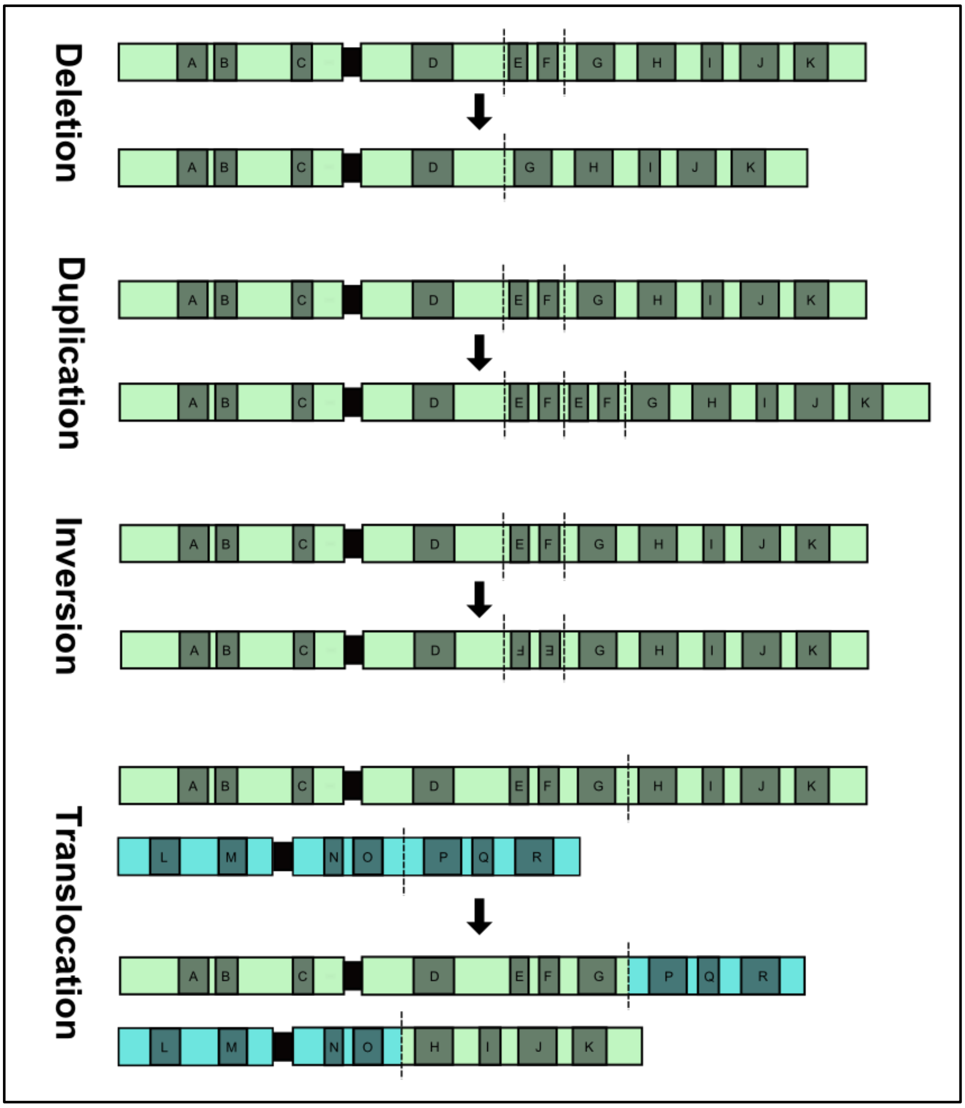

***Figure 2.5** Different classes of chromosomal rearrangement.*

In an inversion event, one chromosome region is inverted. In this class of event, there are two chromosome breaks followed by the inversion of the intervening segment and subsequently the rejoining of the broken ends. For instance, in a subset of parathyroid adenomas, an inversion event in chromosome 11 places the *PTH* gene transcriptional regulatory sequences upstream of *Cyclin D1* gene that results in unregulated *Cyclin D1* overexpression and enhanced cell proliferation (Hemmer, Wasenius et al. 2001).

A translocation event takes place when two nonhomologous chromosomes exchange chromosomal segments. In this class of event, two chromosome breaks take place in two different chromosomes, which is followed by the exchange of segments between two chromosomes and subsequently the broken ends are rejoined (Griffiths, Gelbart et al. 1998). Rowley (2001) describes two mechanisms where translocation plays a role in cancer. In the first mechanism, promoter/enhancer elements of a gene are juxtaposed with the coding region of another gene. For instance, in Burkitt's lymphoma, the promoter of *IGH* gene on chromosome 14 is juxtaposed with the *MYC* gene on chromosome 8 that results in the overexpression of *MYC*. In the second mechanism, translocation results in the juxtaposition of the coding regions of two different genes. For instance, in chronic myeloid leukemia, the coding region of *BCR* gene on chromosome 22 is juxtaposed with the coding region of *ABL* gene on chromosome 9 resulting in a fusion gene capable of activating signaling pathways involved in cell growth and proliferation (Rowley 2001).

Copy number is defined as the number of copies per cell of a particular gene or other DNA sequence (*Encyclopedic dictionary of polymers.* 2011). Chromosomal rearrangement events such as deletion and duplication may change the copy number of a gene. Consequently, the expression level of genes is often correlated with its copy number (Curtis, Shah et al. 2012) - a phenomenon known as *gene dosage effect*. Normally, diploid organisms such as humans have two copies of each autosomal DNA sequence. Somatic copy number changes are known as copy number alterations (CNA). CNAs should be distinguished from copy number variations (CNV). CNA is defined as a sequence that is found at different copy numbers in an individual's germline DNA and in the DNA of a clonal sub-population of cells. On the other hand, CNV is defined as a DNA sequence that is found at different copy numbers in the germline DNA of two different individuals. Normally, an amplification or deletion event of size 50 base pair or greater is defined as CNV (Beroukhim, Mermel et al. 2010).

Loss of heterozygosity (LOH) is an event in which one of two alleles at a heterozygous locus is lost. Subsequently, the lost allele may be discarded or replaced with a duplicated copy of the surviving allele (Weinberg 2013). Copy-neutral LOH occurs if the lost allele is replaced with a duplicated copy of the surviving allele. As a result, the copy number remains unchanged.

Mosén-Ansorena, Aransay and Rodríguez-Ezpeleta (2012) enumerate copy number alterations and loss of heterozygozity (LOH) events as two genomic instabilities associated with cancer (Mosén-Ansorena, Aransay et al. 2012). These events affect the expression of oncogenes and tumor suppressor genes through the mechanisms explained earlier (Pinkel, Segraves et al. 1998, Curtis, Shah et al. 2012, Ha, Roth et al. 2012). Oncogenes and tumor suppressor genes are two important groups of genes with respect to cancer. Oncogenes are normally responsible for cell growth. On the other hand, tumor suppressor genes are responsible for constraining cell proliferation. DNA mutations in oncogenes may result in gain of function and mutations in tumor suppressor genes may result in loss of function. Cancer cells, carrying the oncogenic and tumor suppressor mutations, initiate tumorigenesis and drive tumor progression forward (Hanahan, Weinberg 2000). Hence, accurate detection and characterization of events such as CNAs and LOH are crucial in the discovery of prospective cancer-related genes such as novel oncogenes and tumor suppressor genes. These discoveries are important in the understanding of tumor growth, tumor aggression, and treatment failure and in devising new therapeutic approaches.

2.2. Detection Methods

In the past decades, several methods have been used for the detection of chromosomal aberrations, including CNAs and LOH. These methods have different throughputs, coverages and resolutions. Fluorescence in situ hybridization (FISH), comparative genomic hybridization (CGH), array CGH, single nucleotide polymorphism (SNP) arrays, and HTS-based detection are among the mostly used detection methods.

2.2.1. Fluorescence in Situ Hybridization (FISH)

Weinberg (2013) defines Fluorescence in Situ Hybridization (FISH) as a procedure in which a sequence-specific DNA probe linked to a fluorescent chromophore is annealed to the DNA or RNA of cells that have been immobilized on a microscope slide. As a result, fluorescing spots reveal the presence and often the number of copies of homologous DNA sequences carried by such cells (Weinberg 2013). FISH helps to visualize and map the genetic material in an individual’s cells. It is utilized in order to understand chromosomal abnormalities and other genomic mutations.

2.2.2. Comparative Genomic Hybridization (CGH) and Array CGH (aCGH)

Comparative Genomic Hybridization (CGH) is a molecular cytogenetic method, which was developed by Kallioniemi et al. This method is capable of detecting the relative DNA sequence copy number between genomes that makes it useful for identifying regions of gain or loss of DNA in tumors (Kallioniemi, Kallioniemi et al. 1992).

Weinberg (2013) defines Comparative Genomic Hybridization (CGH) as a procedure in which the copy numbers of a large array of genomic sequences from cells of interest, for instance cancer cells, are compared with the copy numbers of the corresponding sequences in normal reference DNA. This determines whether the various sequences being analyzed are present in increased or decreased copy number in the cells of interest (Weinberg 2013).

Speicher and Carter (2005) explain how CGH works. First, DNA from the test sample and a normal reference sample is extracted. Afterwards, the extracted test and normal samples are labelled differentially; for instance, with green and red color respectively. In the next step, the combined probes are applied to target metaphase chromosomes and compete for complementary hybridization sites. As a result, the amplified regions in the test sample are predominantly green while deleted regions in the test sample are red. Next, digital image analysis is used to quantify the ratios of test to reference fluorescence along the chromosomes. Increased fluorescence ratios denote gains and amplifications. On the other hand, reduced fluorescence ratios denote losses and deletions (Speicher, Carter 2005).

CGH has some limitations. The main limitation of this method is its inability to detect mosaicism, balanced chromosomal translocations, inversions, and whole-genome ploidy changes. Furthermore, the resolution of this method is limited and as a result aberrations smaller than 5-10 Mbp. cannot be detected (Oostlander, Meijer et al. 2004). Array Comparative Genomic hybridization (aCGH) is another method that overcomes the limited resolution of the conventional CGH and it has been mainly in use in cancer research.

In aCGH, metaphase chromosomes are replaced by cloned DNA fragments of size 100-200 kbps that are spotted onto a glass slide (the array). Furthermore, the exact chromosomal location of each fragment is known which provides the possibility to map the aberrant changes directly onto the genomic sequence (Oostlander, Meijer et al. 2004). aCGH follows almost the same procedure as conventional CGH.

In aCGH, the test and normal reference genomes are labelled differentially and hybridized to cloned fragments on the array. Next, with the help of image analysis the intensity differences in the hybridization patterns of both DNAs are calculated. The resulting intensity ratio reflects the copy number aberrations present in the genome (Oostlander, Meijer et al. 2004, Speicher, Carter 2005).

In aCGH, the resolution of the analysis is restricted only by the size of the clones and by the distance between consecutive clones on the array. Furthermore, aCGH can be automated easily for high-throughput applications. Using aCGH, a whole genome analysis can be performed in a single experiment; while using FISH, the same goal is achieved with thousands of independent FISH hybridization (Speicher, Carter 2005, Oostlander, Meijer et al. 2004). Another advantage of the aCGH over FISH is that it does not require targeting a gene of interest. Moreover, aCGH can detect CNA events involving genomic regions as large as whole chromosomes (Moore, Persons et al. 2008). The main limitation of aCGH is its inability to detect aberrations that do not result in copy number changes (Oostlander, Meijer et al. 2004). For instance, in a copy-neutral LOH event, the copy number does not change and this kind of aberration cannot be detected by the aCGH method. However, Single Nucleotide Polymorphism (SNP) arrays are able to overcome such limitation. A SNP is a variation that occurs at a single site in DNA.

2.2.3. Single Nucleotide Polymorphism Arrays

A single nucleotide polymorphism (SNP) is a variation at a single nucleotide position in a DNA sequence among individuals. SNP arrays are high-density arrays based on oligonucleotides that can identify both copy number alterations and LOH at individual nucleotides (Speicher, Carter 2005). One advantage of SNP arrays over aCGH is its ability to measure not only the total intensity at each probe but also the ratio of the intensities between two alleles (Chen, Chang et al. 2013).

To infer a SNP genotype (AA, AB, BB), SNP arrays utilize the principle that each nucleotide base binds to its complementary counterpart (i.e. ‘A’ binds to ‘T’ and ‘C’ binds to ‘G’). In an SNP array experiment, single-stranded DNA fragments hybridize to probes with unique sequences. Each probe composed of *k*-mer oligonucleotides (where k depends on platform) is designed to bind to a specific target DNA fragment. After hybridization, a signal intensity is measured and assigned to each probe and its target depending on the amount of target DNA in the sample and the affinity between target DNA and probe. Finally, these signal intensities are extensively processed and analyzed to infer the SNP genotype (LaFramboise 2009).

In addition, SNP arrays can be used to determine the copy number. The inference of copy number from the SNP array data is based on the observation that the expected probe intensity increases with increased quantity of DNA harboring the region interrogated by the probe. Tools inferring the copy number from the SNP array data, first, summarize the probe-level intensity values at each SNP position and then compare this summary measure against summary measure obtained from a panel of normal samples. This results in the inference of raw copy number, which is a rough measure of the true underlying copy number. These raw copy numbers are subsequently smoothed and segmented to identify segments with CNA. Newer versions of SNP arrays have many thousands of copy number probes solely designed to interrogate copy number variation and as a result, there is no need for the summarization step in inferring the copy number (LaFramboise 2009).

2.2.4. CNA Detection using High Throughput Sequencing Technologies

High throughput sequencing technologies provide the ability to sequence whole genomes or targeted regions of interest rapidly and cheaply. They achieve this by sequencing in parallel massive amounts of short DNA fragments coming from the genome (Liu, Morrison et al. 2013). These technologies are gradually replacing the use of automated Sanger sequencing, which was the dominant method of sequencing for almost two decades.

During the past several years, HTS-based approaches have become the primary means of detecting CNA. HTS-based approaches have several advantages compared with array-based CNA detection approaches. For instance, the estimation of integral copy number from HTS data is more accurate since there is a linear relationship between read-depth and copy number. Furthermore, the identification of the boundaries (breakpoints) of the segments with different copy numbers is more precise since the HTS data covers nearly all nucleotides in the genome. In addition, using HTS data, more alleles can be identified as opposed to the estimation of allele-specific copy number in array-based methods, which is restricted to predefined alleles (Klambauer, Schwarzbauer et al. 2012).

HTS-based CNA detection is performed based on either WGS, WES or targeted sequencing data. WGS is the process of sequencing of an organism’s entire genetic code. However, currently performing a WGS may cost up to a few thousands of euros. In contrast, WES is relatively cheaper since it only sequences the exome (i.e. all the exons or the coding portions of genes) in the genome. However, WES has its own limitations in detecting CNA. For instance, since WES only sequences the coding regions, a hypothetical CNA region extending into the adjacent non-coding region cannot be accurately detected since there is no data coming from the end that falls into the non-coding region.

Alkan et al. (2011) enumerate four different main methods for detecting CNA from HTS data. These are (1) read-pair methods, (2) split-read methods, (3) sequence assembly methods, and (4) read-depth methods (Alkan, Coe et al. 2011).

In the read-pair methods, the mapping information (i.e. span and orientation) of mapped paired-end reads are assessed in order to discover the reads that their span sizes and orientations deviate from what is expected. In principle, this method is capable of detecting most classes of structural variation. For instance, deletion can be detected from read-pairs that map too far apart each other. Some tools that use read-pair method are PEMer (Korbel, Abyzov et al. 2009), VariationHunter (Hormozdiari, Hajirasouliha et al. 2010) and BreakDancer (Chen, Wallis et al. 2009).

Split-read methods utilize the splits in the sequence-reads in order to detect structural variation. If there is a continuous stretch of gaps in the aligned read, it indicates a deletion. If there is a continuous stretch of gaps in the reference where the read is aligned, it indicates an insertion. Even though split-read methods are capable of detecting all classes of structural variation at the single-base pair resolution, they require longer reads. One example of a tool that uses split-read method is Pindel (Ye, Schulz et al. 2009).

Sequence assembly methods are theoretically capable of detecting all classes of structural variation by performing de novo assembly if the reads are long and accurate. However, in practice, a combination of de novo assembly and local assembly is used to construct the genome from scratch. Once the genome is assembled, it can be contrasted against the reference genome to identify the structural variants. One example of a tool that performs de novo assembly and detection of structural variation is Cortex (Iqbal, Caccamo et al. 2012).

Read-depth methods are only capable of detecting two classes of structural variation that are duplications and deletions by examining the increase and decrease in the depth of coverage respectively. Furthermore, read-depth methods are incapable of distinguishing tandem duplications from interspersed duplications (Alkan, Coe et al. 2011). However, read-depth methods can be augmented to identify copy-neutral LOH events by incorporating data about the fraction of alternate allele at heterozygous SNP positions in the method. Some tools that use read-depth method are ControlFREEC (Boeva, Popova et al. 2012b), CNAnorm (Gusnanto, Wood et al. 2012), CLImAT (Yu, Liu et al. 2014), and Patchwork (Mayrhofer, DiLorenzo et al. 2013). Since in this study the aim was to detect only duplications, deletions and copy-neutral LOH, the use of read-depth detection methods seemed appropriate. Because of this choice of method, out of the four detection methods, read-depth methods are explained in more details.

In order to perform CNA detection based on read-depth methods the first step is to map or to align the millions of reads obtained from either WGS or WES to a reference genome. After the reads have been mapped to the reference genome, different read-depth methods use different approaches to perform the CNA detection. However, three main steps are common to many of these methods: (1) raw copy number inference, (2) segmentation and (3) copy number classification (Alkodsi, Louhimo et al. 2015).

2.2.4.1. Raw Copy Number Inference

In the first step, the read-depth (coverage) data is extracted from the aligned reads. Read-depth represents the number of reads that fall into or overlap a local genomic region in the reference genome. Once the read-depth data is available, the raw copy number for each local genomic region can be inferred by calculating the read-depth ratio between tumor and normal samples. The basic assumption behind the use of read-depth in CNA detection is that the average read-depth of a genomic region is proportional to its copy number (Liu, Morrison et al. 2013). In principle, in a diploid genome, the read-depth ratio equal to 1 indicates that there has been no change in the copy number, one copy gain and one copy loss are represented by a ratio of 1.5 and 0.5 respectively (Sathirapongsasuti 2015). Sometimes these ratios are transformed into logratios by taking the logarithm of the calculated ratios. As a result, the expected read-depth logratio for a region with no copy number change is equal to 0. However, some sources of bias such as *mapability* and *GC-content* influence the number of reads aligned to a given region in the genome. Furthermore, normal cell contamination also affects the estimation of raw copy number (Alkodsi, Louhimo et al. 2015). Normal cell contamination refers to the fact that tumor biopsy samples are mixtures of normal and cancer cells. Oesper et al. (2013) define the tumor purity – a term closely related to normal cell contamination

- of a sample as “the fraction of cells in the sample that are cancerous, and not normal cells”. As a result, high normal cell contamination or low tumor purity decreases the power to detect copy number aberrations in the cancer genome by shifting the read-depth ratios and allele fraction values away from the expected values (Oesper, Mahmoody et al. 2013).

Liu et al. (2013) define mapability as “the probability for a region in the reference genome that a read originating from it is unambiguously mapped back to it.” Mapability bias arises from the fact that some of the short reads originating from the identical or highly similar regions in the genome cannot be mapped uniquely to the reference genome. During the read alignment, these reads known as *multi-reads* are either discarded or mapped to a random position out of all equally good match positions. As a results, regions in the genome with low mapability show lower mapped read-depth.

GC-content bias arises because the read-depth of the regions in the genome with high or low GC-content is lower than regions with medium GC-content. It is believed that the GC-content bias is caused mainly by the polymerase chain reaction (PCR) amplification step (Liu, Morrison et al. 2013, Alkodsi, Louhimo et al. 2015).

In order to correct for such biases different tools uses different approaches. Tools that require a matched normal sample assume that these biases largely and implicitly are corrected for when the read-depth ratio of the tumor and normal sample is being calculated. Other tools that do not require a matched normal sample, for instance ControlFREEC, perform mapability bias correction and CG-content normalization.

Even though read-depth methods can detect CNAs solely using read-depth ratios, another source of information, *B-allele fraction* (BAF), can improve the CNA detection results. The human genome is diploid and so normally has two copies of its autosomal chromosomes. A locus in the normal human genome can have one of four possible genotypes (AA, AB, BA, and BB). Let ‘A’ represent the allele that has the same nucleotide as the reference genome (the reference allele) and ‘B’ represent the alternate allele. BAF is the estimation of allelic fraction at a SNP locus and it is calculated as:

$$BAF = \frac{b}{(a+b)}$$

where a and b are the copy numbers of A and B allele respectively (Liu, Morrison et al. 2013). Using this formula, the BAFs for the four possible genotypes, AA, AB, BA and BB are 0, 0.5, 0.5 and 1 respectively. In a normal genome, at heterozygous SNP loci BAF values are equal to 0.5. A CNA event can shift these values away from 0.5. Thus, BAF values at heterozygous SNP loci are informative in a sense that any deviation from 0.5 may point to a CNA event. For instance, the use of BAF can aid in the identification of copy-neutral LOH events normally not detectable solely by the use of read-depth ratios. A copy-neutral LOH happens when in a given region of a diploid genome, one copy of the genome is deleted and the other copy is doubled. Alternatively, in a given region of a diploid genome, one copy of the genome can be doubled and later the other copy is lost. Another mechanism through which copy-neutral LOH can occur is *gene conversion* also known as acquired uniparental disomy (aUPD). Gene conversion is a process through which genetic material from an intact region of genome is transferred to a region containing double-strand breaks. Gene conversion can occur between sister chromatids, homologous chromosomes, and even between homologous sequences on the same chromatid or on different chromosomes (Chen, Cooper et al. 2007). As a result, the DNA sequences become homozygous after the gene conversion event. Even though a CNA event has happened, the read-depth ratio remains equal to one suggesting that no event has happened but BAF values for heterozygous SNP loci is no longer equal to 0.5 pointing to the occurrence of a CNA event. It is important to note that, BAF values at heterozygous loci cannot detect all of the CNA events by themselves. For instance, there might be a CNA event where both copies of the chromosome are doubled; the genotype at the heterozygous SNP loci is AABB, however, the BAF value remains equal to 0.5. Thus, a combination of read-depth ratios and BAF can improve CNA detection results.

2.2.4.2. Segmentation

The second step is to segment the inferred raw copy numbers. Segmentation refers to a process where a chromosome is divided into segments of similar copy number based on the inferred raw copy numbers. In order to create the segments, change-points in the raw copy number should be identified. There are different methods developed to perform the segmentation. For instance, in circular binary segmentation (CBS), the method finds the segments in a chromosome with the criteria of minimizing the within-segment variance and maximizing the between-segments variance (Olshen, Venkatraman et al. 2004, Liu, Morrison et al. 2013). CBS was originally developed for aCGH data. To find a change-point, CBS uses a likelihood ratio test statistic to test whether a segment can be divided into two segments having different means. Conceptually, a segment is spliced at the two ends to form a circle. Then, the likelihood ratio test statistic for testing the hypothesis that the arc from $i + 1$ to j and its complement have different means is given by:

$$Z_{ij} = \{1/(j - i) + 1/(n - j + i)\}^{-1/2}\{(S_j - S_i)/(j - i) - (S_n - S_j + S_i)/(n - j + i)\}.$$

where S_i and S_j are the partial sums of the logratios of intensities (i.e. $X_1, \dots, X_n$) indexed by the location of the n markers being studied in aCGH data (i.e. $S_i = X_1 + \cdots + X_i, 1 \le i \le n$ and $S_j = X_1 + \cdots + X_j, 1 \le j \le n$). In addition, the statistic $Z_C = max_{1 \le i < j \le n} |Z_{ij}|$ is calculated. A change is called if this statistic exceeds a threshold level based on the null distribution of Z_C computed using Monte Carlo simulation or approximation. After the rejection of the null hypothesis (i.e. there is no change in mean in a segment), the change-point(s) is (are) estimated to be i (and j) such that $Z_C = |Z_{ij}|$. Finally, this procedure is applied recursively to identify all the changes (Olshen, Venkatraman et al. 2004). CBS has been shown to be one of the most accurate approaches in the detection of CNAs. However, this approach is slow and requires $O(n^2)$ computations. In contrast, hidden Markov model (HMM), other way to perform segmentation, only requires $O(n)$ computations (Crisan, Xiang 2009).

HMMs are probabilistic models designed to determine a hidden (i.e. unknown) sequence of states based on a sequence of observations (Seiser, Innocenti 2015). A change from one state to another state (where the new state of a sequence is only dependent on its previous state) is described by a matrix of *transition probabilities*. The transition probabilities describe the expected linear order of states (Eddy 2004). Furthermore, each step has an *emission probability*. Emission probabilities model each of the observations as a function of a particular hidden state. To identify the most likely hidden sequence based on the model, HMMs first optimize model parameters (including the emission and transition probabilities) to best describe the sequence of observations by learning from training data. After parameter optimization, HMMs determine the most likely sequence of hidden states using a dynamic programming approach (Seiser, Innocenti 2015).

In HMM-based CNA detection method, the observed read-depth ratios and the BAF values represent residues emitted from the hidden copy number state at each locus. Each possible copy number state has specific emission probabilities that model the read-depth ratio and BAF value composition for that state. In addition, copy number states are dependent on adjacent loci. The transition probabilities describe the probabilities associated with either remaining in the current copy number state or transitioning to a new copy number state at the next locus. The emission and transition probabilities and other parameters are optimized based on the observed sequence of read-depth ratios and BAF values and subsequently the optimized parameters are used to determine the most likely sequence of hidden copy number states. Once the copy number state at each locus is determined, loci with the same state are combined to create a segment (Seiser, Innocenti 2015, Liu, Morrison et al. 2013).

In order to perform the segmentation, some tools use the methods explained above. For instance, Patchwork uses CBS and CLImAT develops its own HMM. However, some tools employ different approaches than described above. For instance, ControlFREEC has developed its own segmentation method by using least absolute shrinkage estimator (LASSO) regression (Alkodsi, Louhimo et al. 2015, Liu, Morrison et al. 2013, Boeva, Popova et al. 2012).

2.2.4.3. Copy Number Classification

The third step is to classify and annotate different segments into different copy number states. Some tools assign an integral copy number to each segment and some other tools only calculate and report the copy number ratio for each segment by finding either the mean or the median copy number of the segment. For instance, ControlFREEC assigns an integral copy number to each segment. Furthermore, it annotates the segments as gain or loss. If the analysis is done using a paired normal sample, ControlFREEC annotates the copy number variation segments either as somatic or germline. Tools using HMM normally combine steps two and three because in step two the state (integral copy number) is already inferred.

3. Research Objectives

The main objective of this research is to develop a tool capable of detecting copy number alterations and copy-neutral LOH events in tumor genomes. The tool should be able to detect accurately genomic aberrations such as deletion, amplification and copy-neutral LOH events and distinguish regions in the cancer genome that have acquired such aberrations from normal regions by annotating them accordingly. Furthermore, the tool should be easy-to-use and fast in performing the analysis.

The main objective can be broken into the following steps:

1. Developing and implementing an algorithm for copy number analysis
2. Implementing a simulator capable of simulating read-depth and BAF data
3. Benchmarking the algorithm with simulated and real data

4. Material and Methods

This chapter explains the implementation of the tool and the materials and methods that were used in order to implement it. Furthermore, in this chapter, the implementation of a simulator capable of simulating the cancer genome with events such as deletion, amplification and copy-neutral LOH is explained.

4.1. Mapping the Reads

HTS technologies produce millions of ‘reads’ that are the result of sequencing of the fragmented DNA. These reads need to be ‘mapped’ or ‘aligned’ to the reference genome by determining their position on the reference genome (Trapnell, Salzberg 2009). Aligning the reads to a reference genome is the first step in many high throughput sequencing data analyses. Since there are millions of reads to be aligned to the reference genome, one of the important challenges is the speed of the alignment. The other challenge is the accuracy of the alignment. There are several tools available capable of performing the mapping of the reads such as Bowtie (Langmead, Trapnell et al. 2009), BWA (Li, Durbin 2009), MAQ (Li, Ruan et al. 2008), etc. Each of these tools employs different strategies. For instance, Bowtie uses Burrows-Wheeler transform to transform and index the reference genome. The result of the transformation is then stored in the computer’s memory. Then, Bowtie aligns a read one character at a time to the Burrows-Wheeler-transformed genome up until the whole read is aligned. Bowtie also utilizes a backtracking algorithm that permits mismatches (Trapnell, Salzberg 2009, Langmead, Trapnell et al. 2009).

4.2. SAM/BAM File Format

The sequence alignment/map (SAM) format is a widely used and generic alignment format for the storage of the aligned reads against the reference sequences. SAM format is capable of supporting short and long reads produced by different sequencing platforms (Li, Handsaker et al. 2009).

The SAM format is comprised of two sections and several tab-delimited fields. The two sections are the *header* and the *alignment* section. The alignment section is comprised of several lines and each line has 11 mandatory and some other optional fields (Li, Handsaker et al. 2009). Table 4.1 describes each of the mandatory fields in the SAM format.

***Table 4.1** Mandatory fields in SAM format (adapted from Li, Handsaker et al. 2009).*

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe (‘=’ if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

The binary alignment/map (BAM) is the binary, compressed representation of SAM that makes it smaller in size. In addition, retrieval of the alignments that overlap a given chromosomal region is faster by performing linear indexing on a position-sorted BAM file. This capability relaxes the need to load the entire file into memory and furthermore, provides the ability of stream-based processing on a specific genomic region (Li, Handsaker et al. 2009).

4.3. Samtools

Samtools is a software package capable of parsing and manipulating alignments in the SAM/BAM format. Samtools provides its users with several abilities such as (1) Sorting, indexing and merging the alignments, (2) Removing PCR duplicates, (3) Generating per-position information in pileup format, (4) Calling SNPs and short indel variants, and (5) Representing alignments in a text-based viewer (Li, Handsaker et al. 2009).

4.4. Wiggle File Format

The wiggle (WIG) format is used for display of dense, continuous data such as GC percent, probability scores, and transcriptome data. This format is composed of one wiggle-track definition line and two other types of lines: declaration lines and data lines. Furthermore, Wiggle format provides two formatting options: *fixedStep* and *variableStep*.

FixedStep formatting is used for data with regular intervals between new data values. This format is best used for genome-wide datasets containing millions of data points. The declaration line starts with the word *fixedStep* and includes specifications for chromosome (*chrom* parameter), start coordinate (*start* parameter), and step size (*step* parameter). This line harbors one optional parameter, the *span* parameter with default value equal to 1. The *span* parameter declares the number of bases that each data value should cover and allows data composed of contiguous runs of bases with the same data value to be specified more succinctly.

VariableStep formatting is used for data with irregular intervals between new data values. This format, like fixedStep, is best used for genome-wide data set comprising of millions of data points. The declaration line starts with the word *variableStep* and is followed by a specification for a chromosome (*chrom* parameter) and the optional *span* parameter having the same meaning as explained earlier (genome.ucsc.edu 2015, genomewiki.ucsc.edu 2015, ensembl.org 2015).

In this work, the read-depth data is stored in files with WIG format.

4.5. SEG File Format

SEG file (segmented data) is a tab-delimited text file that lists loci and their associated numeric values. The first row contains column headings and each subsequent row contains a locus and an associated numeric value (broadinstitute.org 2015). A typical SEG file has at least 5 columns which are: *ID*, *chrom* (indicating the chromosome), *loc.start* (indicating the start location of the locus), *loc.end* (indicating the end location of the locus) and finally one or more columns assigning a numeric value to the specified locus.

This format is used by visualization tools such as the Integrative Genomics Viewer (IGV) for the visualization of the results stored in a file with the SEG format. IGV reads the first four columns as track name, chromosome, start location, and end location. If there are more than 5 columns in the

SEG file, IGV reads the last column as the numeric value for that locus while ignoring all other columns (broadinstitute.org 2015).

In this work, the developed tool outputs the results of copy number analysis and segmentation in the SEG file format. This provides the capability of visualizing the results in tools such as IGV.

4.6. Integrative Genomic Viewer (IGV)

The Integrative Genomics Viewer (IGV) is a high-performance and lightweight visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based data, next-generation sequencing data, and genomic annotations such as aligned sequence reads, mutations, copy number, RNA interference screens, gene expression, methylation and genomic annotations (Robinson, Thorvaldsdóttir et al. 2011).

One advantage of using IGV over other available genome browsers such as MapView, EagleView, etc. is that it supports visualizing and viewing of different supported data types together and to use the sample metadata to group, sort and filter them (Robinson, Thorvaldsdóttir et al. 2011).

IGV provides its users with the ability to zoom in, zoom out and pan across the genome from the whole genome to a single base pair. These, along with other capabilities such as for instance sorting the reads in a BAM file by base, quality or strand, or as for another example the color coding of paired-end reads provides its users with the ability to visually inspect the data and performing further validations (Robinson, Thorvaldsdóttir et al. 2011).

Since IGV supports all file formats used in this work, and the fact that it is freely available and easy-to-use, it was chosen as the genome browser of choice.

4.7. Python Programming Language

Python is a free, portable and general-purpose high-level programming language. In recent years, the popularity of Python programming language has increased among bioinformaticians. In this work, Python programming language was used as the implementation language for our software.

4.8. The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort started in 2006 as a three-year pilot project and continued afterwards in order to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing (The Cancer Genome Atlas 2015).

The main goal of TCGA is to improve the ability to diagnose, treat and prevent cancer by providing an atlas of genomic changes caused by each type of cancer. TCGA has collected more than 30 tumor types. For each tumor type, TCGA has collected and examining up to 500 samples in order to provide the statistical power needed to produce a comprehensive genomic profile of each cancer type.

Furthermore, TCGA provides an infrastructure for making cancer-related data publicly available to the researchers all around the globe. It is believed that this would enable the cancer research community to make and validate important discoveries (The Cancer Genome Atlas 2015). The collected cancer-related data is available for download at either TCGA’s Data Portal (Tcga-data.nci.nih.gov 2015) or at Cancer Genomics Hub (CGHub) (cghub.ucsc.edu 2015). TCGA’s Data Portal hosts clinical information, genomic characterization data, and high-level sequence analysis of the tumor genomes. On the other hand, lower levels of sequence data such as alignments stored in BAM files are available at CGHub.

4.9. Data

In this work, the BAM files storing the alignments of a WGS data for paired normal and tumor samples of 6 individuals diagnosed with prostate adenocarcinoma (PRAD) was selected and downloaded from the CGHub in order to run the copy number analysis tool developed in this thesis over the downloaded data. Furthermore, other available copy number analysis tools were run over the same data in order to benchmark the performance of the developed tool in this work. Table 4.2 represents the filter used to search for the eligible samples in CGHub.

4.10. Input Files

The tool requires three input files. The first two files contain the read-depths for the normal and tumor samples. These files have a wiggle (WIG) format. The third file, which is a tab-separated file with TSV format, contains the BAF values at heterozygous SNPs. These files are obtained by running the tools available in the *pypette* package over the BAM files. Pypette is a set of utilities for the analysis of high throughput sequencing data available at: https://github.com/annalam/pypette (Appendix C represents the set of commands used to extract the read-depth and BAF data from the BAM files).

4.10.1. Extraction of Read-depth and Heterozygous B-Allele Fraction Data

Figure 4.1 illustrates how the read-depth (coverage) is calculated from the BAM files. A window of specific size is slid over the mapped reads and the number of reads which fall within the window or have an overlap with the window determines the read-depth at the window.

Table 4.2 *Filters used for searching for samples in the CGHub.*

Disease:	Prostate adenocarcinoma (PRAD)
Study:	TCGA
Platform:	Illumina
Library Type:	WGS

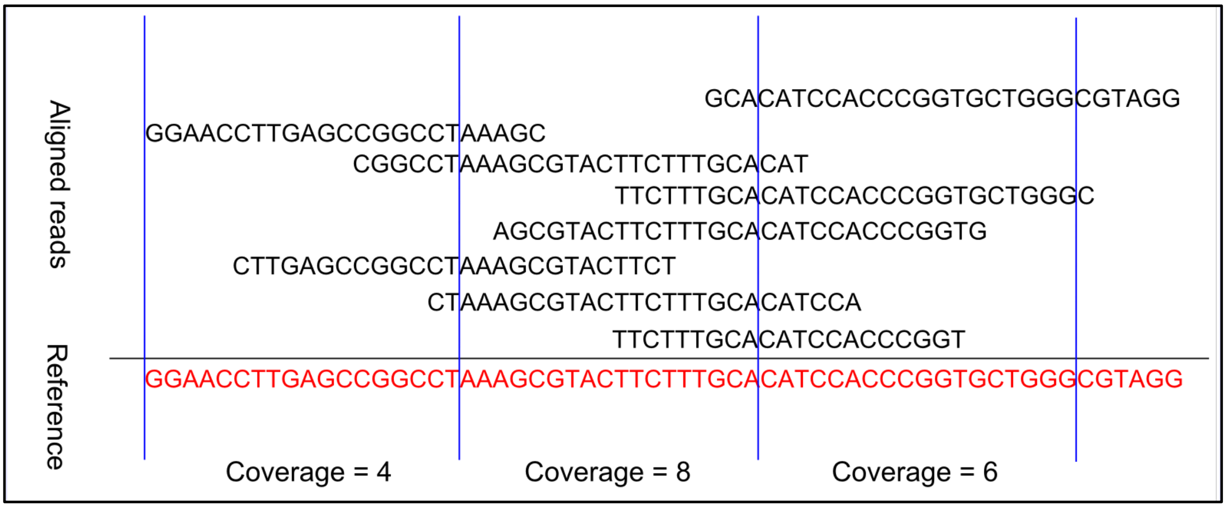
Aligned reads
GCACATCCACCCGGTGCTGGGCGTAGG
GGAACCTTGAGCCGGCCTAAAGC
CGGCCTAAAGCGTACTTCTTTGCACAT
TTCTTTGCACATCCACCCGGTGCTGGGC
AGCGTACTTCTTTGCACATCCACCCGGTG
CTTGAGCCGGCCTAAAGCGTACTTCT
CTAAAGCGTACTTCTTTGCACATCCA
TTCTTTGCACATCCACCCGGT
Reference
GGAACCTTGAGCCGGCCTAAAGCGTACTTCTTTGCACATCCACCCGGTGCTGGGCGTAGG
Coverage = 4
Coverage = 8
Coverage = 6

Figure 4.1 *Calculation of read-depth (coverage) from a BAM file.*

The B-allele fraction is defined as the number of alternate alleles (B-allele) at a specific position in the genome divided by the total number of aligned reads at that position:

$$\text{B\_allele Fraction} = \frac{Number\ of\ alternate\ alleles}{Total\ Number\ of\ reads}$$

The B-allele fraction at heterozygous SNPs is 0.5 since half of the reads come from the reference allele and the other half from the alternate allele.

4.11. Pipeline of the Tool

Figure 4.2 illustrates the pipeline of the tool. First, the read-depths from the normal and tumor sample are used to calculate the read-depth logratios. Furthermore, the B-allele fraction data is smoothed. Then, the read-depth logratios and the smoothed BAF are used for segmenting the genome by using a double sliding window method. Finally, the tool outputs the result of segmentation as a SEG file. In addition, another tool was developed to extract copy-neutral LOH events from the output of the developed tool. In what follows, each of these steps is explained in detail.

4.11.1. Calculation of the Logratios

Figure 4.3 illustrates how the read-depth logratios are calculated. Furthermore, this figure illustrates the visualization of the normal and tumor samples coverages and the resulting read-depth logratios in IGV. It should be mentioned that in the calculation of the read-depth logratios, regions with read-depth below a certain point are discarded (i.e. their value is replaced by not a number (NAN) value) since they are believed to be non-informative. This threshold is given to the tool as an argument – *min_read*. By default, *min_read* is set to be minimum of 50 reads.

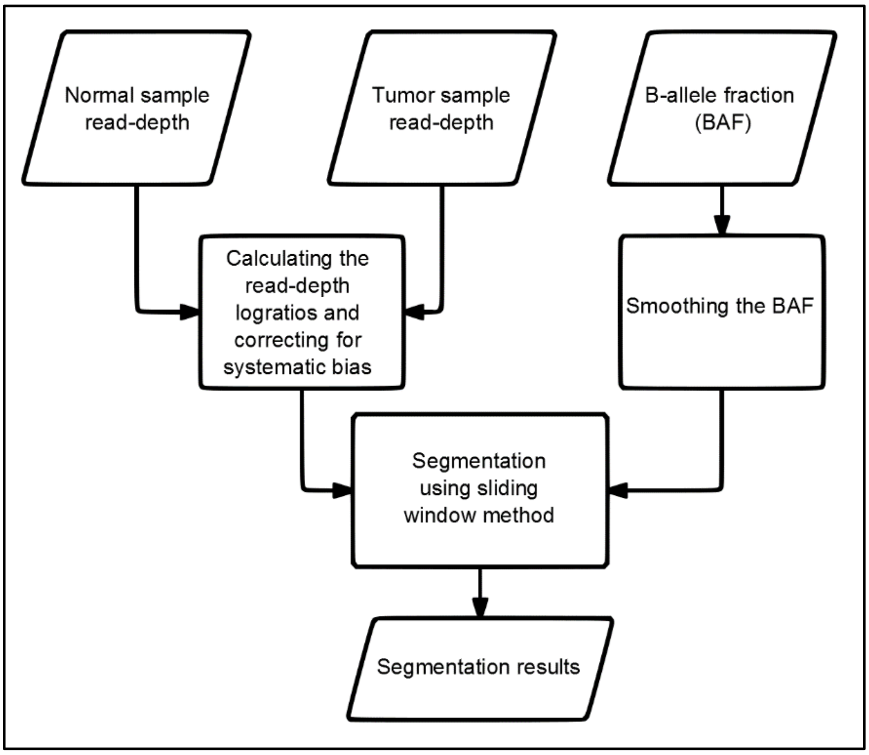
Normal sample read-depth
Tumor sample read-depth
B-allele fraction (BAF)
Calculating the read-depth logratios and correcting for systematic bias
Smoothing the BAF
Segmentation using sliding window method
Segmentation results

Figure 4.2 *Pipeline of the tool.*

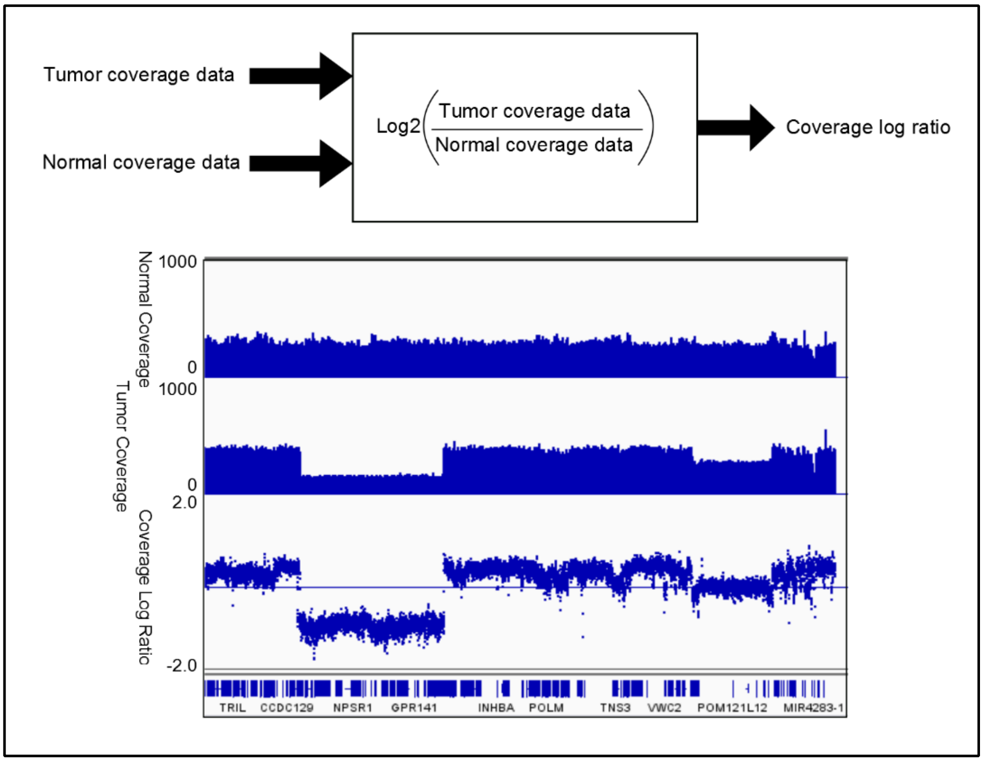

Figure 4.3 *Calculation of the read-depth (coverage) logratio and its visualization in IGV.*

As it can be seen from figure 4.3, read-depth logratios can specify the regions that are harboring aberrations. If a region in genome is not aberrant, the read-depth ratio of tumor read-depth to normal read-depth is close to one and therefore, the logarithm of this ratio is close to zero. However, if a region has undergone amplification, the read-depth ratio of tumor read-depth to normal read-depth is greater than one and as a result, the logarithm of this ratio is greater than 1. In addition, if a region has undergone deletion, the read-depth ratio of tumor read-depth to normal read-depth is less than one and resulting in a negative log ratio. However, if a region in genome has undergone a copy-neutral LOH, it cannot be detected via the read-depth logratio since the read-depth logratio for this region is close to zero. The B-allele fraction is used to identify such aberrations in the genome.

4.11.2. Correction for GC-content and Mapability Bias

As explained earlier biases such as mapability and GC-content complicate the inference of the raw copy number. Low mapability regions in the genome show lower mapped read-depth and the read-depth of the regions in the genome with high or low GC-content is lower than regions with medium GC-content. This requires the need to correct for such biases. In this work, it is assumed that these biases are largely and implicitly corrected for when the read-depth ratio of the tumor and normal sample is being calculated because of using a matched normal sample. The idea behind this assumption is that these biases affect both the normal and the tumor samples in a similar way (Klambauer, Schwarzbauer et al. 2012).

4.11.3. Mirroring and Smoothing of B-allele Fractions

As explained earlier, the read-depth logratios are not alone capable of detecting the copy-neutral LOH events and the use of BAF values at the heterozygous SNPs helps in the identification of such aberrations. The expected value for BAF at a heterozygous SNP in a diploid organism is equal to 0.5

since half of the reads are the reference allele and the other half are the alternate allele. Figure 4.4 illustrates how the BAF values at heterozygous SNPs shift away from the expected value of 0.5 at regions with different aberrations.

As it can be seen from the BAF track in figure 4.4, BAF values are symmetrically positioned around 0.5. In order to make the analysis of BAF data easier, the BAF values are mirrored or flipped around the 0.5 axis. This process is known as mirroring. As a result, the flipped BAF value for a SNP at a normal region in a diploid organism is equal to 0. The flipped BAF track in figure 4.4 illustrates the result of mirroring. Simultaneous with mirroring, through a process known as smoothing, the noise in the BAF data is reduced by applying a median filter to it. The median filter filters the data by running through the data points one by one and replacing each of the data points with the median of the data points that fall into a window of specified size centered on the current data point. To perform the simultaneous mirroring and smoothing, first two different ways of smoothing and mirroring are considered depending on whether a BAF value is close to 0.5 or it is close to either 0 or 1. If the BAF value is close to 0.5, the best way of simultaneous mirroring and smoothing is to calculate:

$$|0.5 - median(x)|.$$

However, if the BAF value is close to either 0 or 1, the best way of simultaneous mirroring and smoothing is to calculate:

$$median(|0.5 - x|).$$

To find out how close a BAF value is to 0.5 (i.e. a heterozygous SNP has a BAF value equal to 0.5) the *heterozygosity* measure is calculated in the following way:

$$heterozygosity \ = H = 1 - 2 * |0.5 - x|.$$

where x represents a BAF value. If a BAF value is close to 0.5, the heterozygozity value is close to 1 and if a BAF value is close to either 0 or 1, the heterozygozity value is close to 0. Using the following linear interpolation formula, BAF values are simultaneously mirrored and smoothed:

$$H * |0.5 - median(x)| + (1 - H) * median(|0.5 - x|).$$

where H is the heterozygosity measure and x is a BAF value. Figures 4.5 to 4.7 illustrate the read-depth logratio and the BAF profiles before mirroring at regions with different events such as hemizygous deletion, homozygous deletion and copy-neutral LOH.

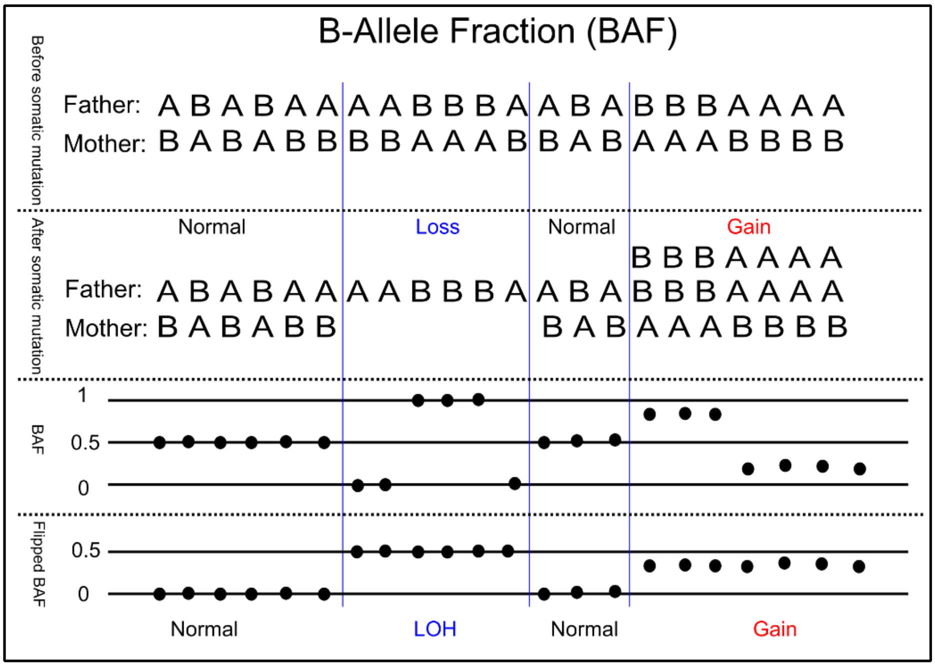
B-Allele Fraction (BAF)
Before somatic mutation
Father: A B A B A A A A B B B A A B A B B B A A A A
Mother: B A B A B B B B A A A B B A B A A A B B B B
After somatic mutation
Normal
Loss
Normal
Gain
B B B A A A A
Father: A B A B A A A A B B B A A B A B B B A A A A
Mother: B A B A B B B A B A A A B B B B
BAF
1
0.5
0
Flipped BAF
0.5
0
Normal
LOH
Normal
Gain

Figure 4.4 *Illustration of BAF and flipped BAF values at regions with different aberrations.*

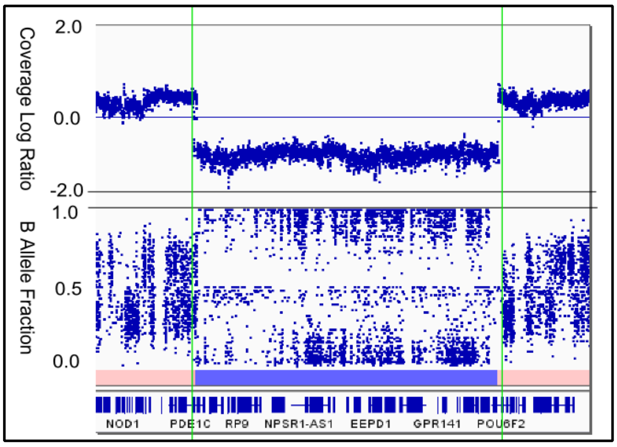

Figure 4.5 *Illustration of a hemizygous deletion event.*

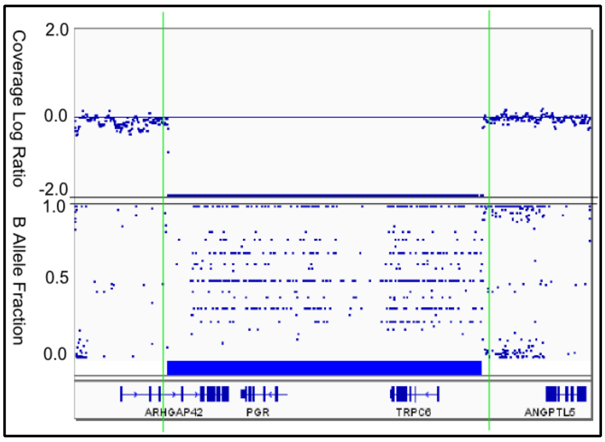

Figure 4.6 *Illustration of a homozygous deletion event.*

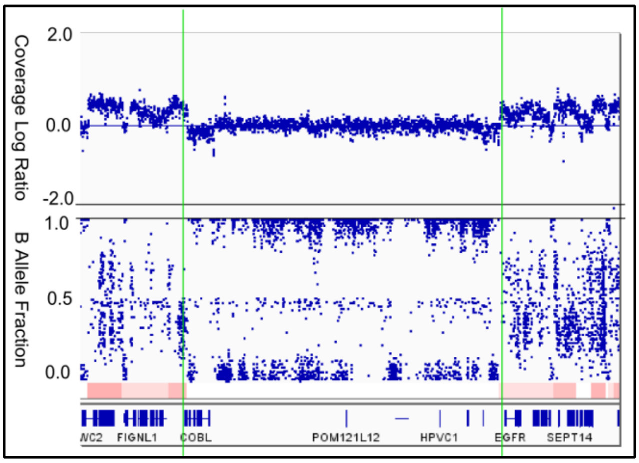

Figure 4.7 *Illustration of a copy-neutral LOH event.*

4.12. Double Sliding Window Method

After the calculation of the read-depth logratios and calculation of smoothed BAF values, the data is ready for segmentation. In order to segment the genome based on events such as deletion or duplication a double sliding window method was used. Here, first, the method is explained in case of segmenting the genome based on only the read-depth logratios values and then, the strategy for the segmentation of the genome based on both read-depth logratios and BAF values will be explained.

Two windows of the same size where the end of the first window touches the beginning of the second window is slid over read-depth logratio values and the mean of these values for each of the windows is calculated. Then the absolute mean difference for logratios for the two consecutive windows is calculated. If the absolute mean difference of the logratios is above a certain threshold then the algorithm places a breakpoint at the position where the two windows touch each other. The breakpoint has a value that corresponds to the absolute mean difference of the logratios within the windows. Since both windows are slid over the logratio values one step at a time, we will end up with a cluster of breakpoints with different values. The next step is to choose the breakpoint with the maximum value and to discard the remaining breakpoints in the cluster. Figure 4.8 illustrates the idea of double sliding window for the logratio data. Figure 4.8 also illustrates the cluster of breakpoints at a specific locus of the genome where the breakpoint with the maximum value is highlighted in red. If for the placement of a breakpoint, a decision is made solely based on a read-depth logratio threshold, the decision boundary is determined only by one value. Figure 4.9 illustrates the decision boundary based on one value.

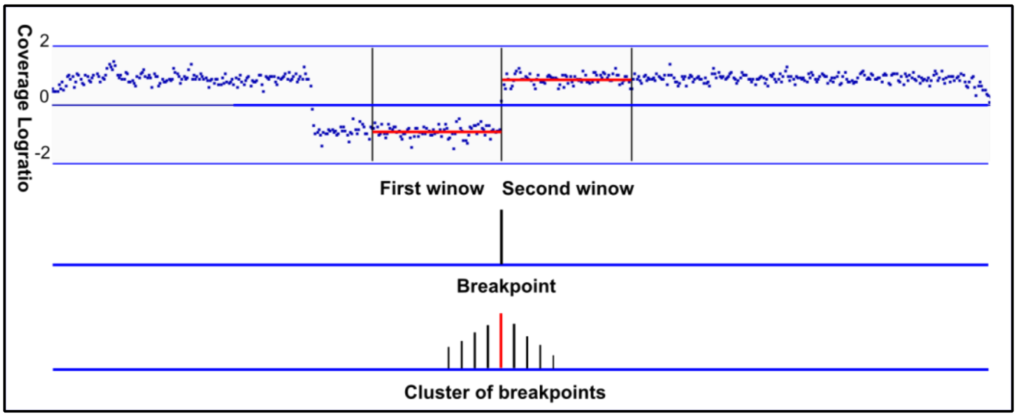

Figure 4.8 *Illustration of the double sliding window method functioning.*

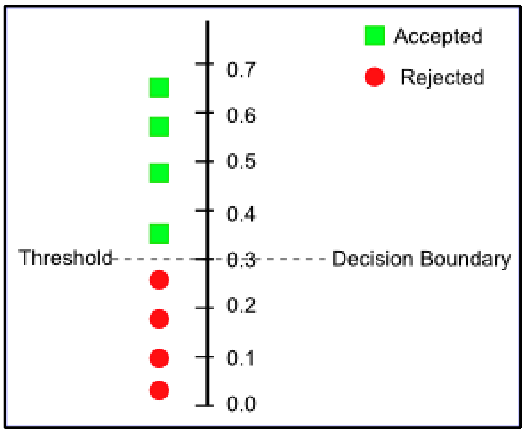

Figure 4.9 *Decision boundary based on one value.*

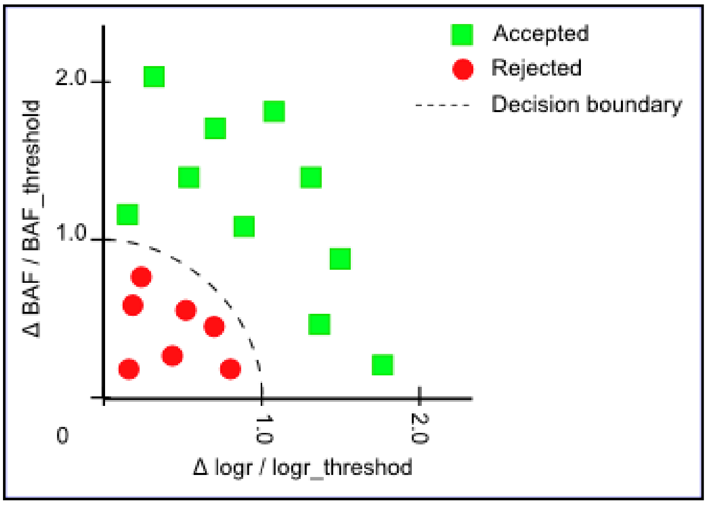

Figure 4.10 Decision boundary based on two values.

However, the developed tool uses two thresholds (i.e. read-depth logratio threshold and the BAF threshold) in order to decide whether a breakpoint should be placed at the position in the genome where two consecutive windows touch each other. To do this, first a compound score is calculated for each position in the genome using the absolute mean difference of the read-depth logratio values and the absolute mean difference of the BAF values. If the compound score is greater than 1, then a breakpoint is placed where the two windows touch each other. The following formula defines how the compound score is calculated.

$$compund\ score = \frac{abs\big(mean(logr_{win\_1}) - mean(logr_{win\_2})\big)^2}{logratio\ threshold} + \frac{abs\big(mean(BAF_{win\_1}) - mean(BAF_{win\_2})\big)^2}{BAF\ threshold}$$

Where BAF_{win_i} represents the BAF values in the i^{th} window and $logr_{win\_i}$ represents the read-depth logratio values in the i^{th} window.

In this case, two values determine the decision boundary. Figure 4.10 illustrates the decision boundary based on these two thresholds.

4.13. Multiple Window Sizes and Thresholds and Determination of Consensus Breakpoints

It is possible some breakpoints will not be detected with only one window of a specific size. Thus, the tool uses multiple windows with different sizes and different thresholds. The size of each new window is 1.5 times larger than the previous one. The read-depth logratio threshold and BAF threshold are calculated for each of the new windows based the previous thresholds and a standard score in the following way. The calculated mean of the read-depth logratios within a window has a probability distribution. Assuming that the observed read-depth logratios within a window of size *n* are observations of a random sample from a normal (i.e. $N(\mu, \sigma^2)$) population, the probability distribution of the mean is also normal (i.e. $N(\mu,\ {\sigma^2}/{n})$). This is based on the theorem stating that the linear combination of mutually independent normal random variables follows the normal distribution (Flury 1997). Increasing the window size, increases the sample size and consequently decreases the variance of the probability distribution of the mean. As a result, it can be determined how much the standard deviation of the mean decreases when the window size becomes 1.5 times larger than the previous window size (i.e. $\sigma_{new\_win} = \frac{1}{\sqrt[2]{3/2}} * \sigma_{old\_win}$). The decrease in the standard deviation decreases the threshold used to call whether a change in the mean has occurred. The threshold specifies how many standard deviations away from the mean should be considered as a change. Therefore, the threshold can be thought of as the *standard score* (or *z-score*) multiplied by standard deviation:

$$threshold = z_{score} * \sigma.$$

where z_score is the standard score and σ is the standard deviation. Having calculated the new standard deviation for the larger window, the new threshold can also be calculated as follows:

$$threshold_{new\_win} = z_{score} * \sigma_{new_{win}} = z_{score} * \frac{1}{\sqrt[2]{3/2}} * \sigma_{old_{win}}.$$

After finding the breakpoints for each of the windows, all the breakpoints detected by the smallest window are accepted to be the consensus breakpoints. Then, breakpoints detected by the second window that is bigger than the first window and smaller than the remaining windows are added to the list of consensus breakpoints. However, for each of the breakpoints detected by the second window it is checked whether an already detected breakpoint falls inside the vicinity of this breakpoint. The vicinity is defined as one current window size before and one after the breakpoint. If there is already a breakpoint in the vicinity of the current breakpoint, the breakpoint is discarded and if not, the breakpoint is added to the list of consensus breakpoints. This procedure is continued until all the detected breakpoints by all the windows are checked. Figure 4.11 illustrates how consensus is found for two different windows.

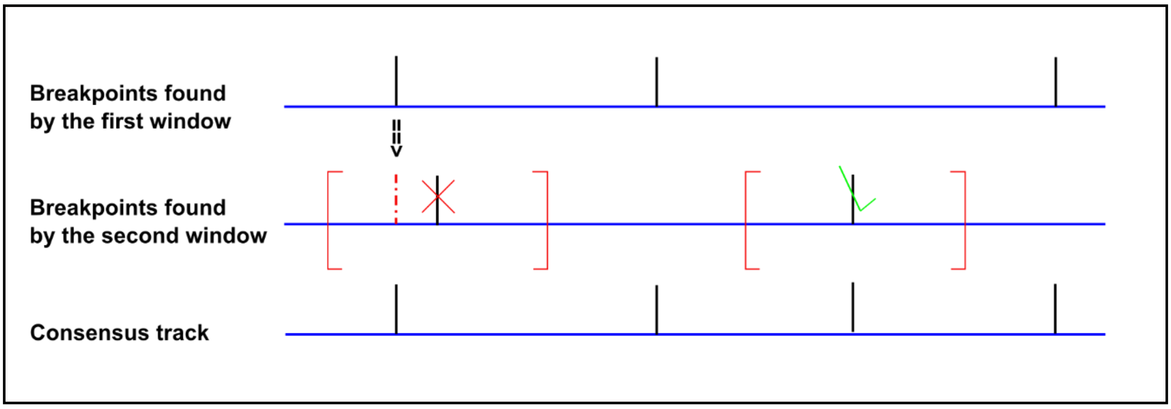
Breakpoints found by the first window
Breakpoints found by the second window
Consensus track

Figure 4.11 *Determination of consensus breakpoints.*

4.14. Segmentation and Merging the Segments

Once the consensus breakpoints are identified, the segmentation is performed. Two consecutive breakpoints constitute a segment. For each segment, the average read-depth logratio and average BAF is calculated. However, if the absolute difference of the average logratios of two consecutive segments is below certain read-depth logratio merging threshold and the absolute difference of the average BAF of two consecutive segments is below a certain BAF merging threshold, these two consecutive segments are merged into one segment. Subsequently, the new average logratio and average BAF for the new segment are calculated. The results are written into a SEG file so that it can be visualized in IGV. The SEG file contains the following fields: Sample ID, chromosome, segment start point, segment endpoint, segment coverage logratio mean and segment BAF mean. Figure 4.12 illustrates how the segmentation and merging works.

4.15. Extraction of Copy-neutral LOH Events from SEG Files

As explained earlier the resulting SEG file contains several fields. However, IGV is not capable of visualizing the results if the SEG file contains both the segment read-depth logratio mean and segment BAF mean. Thus, in order to be able to visualize the results in IGV, BAF mean field should be excluded from the SEG file and the visualization is done solely based on the read-depth logratio. This results in a situation where the segments with the copy-neutral LOH events are represented as normal and this regions cannot be detected by eye solely by looking at the visualized results in IGV. Therefore, a piece of code was written in order to extract copy-neutral LOH segments from the original SEG file. To do this, each segment in the SEG file is checked against two thresholds. If the read-depth logratio mean of each segment is below 0.1 and the BAF mean of the same segment is above 0.35, then the segment is considered to harbor a copy-neutral LOH event.

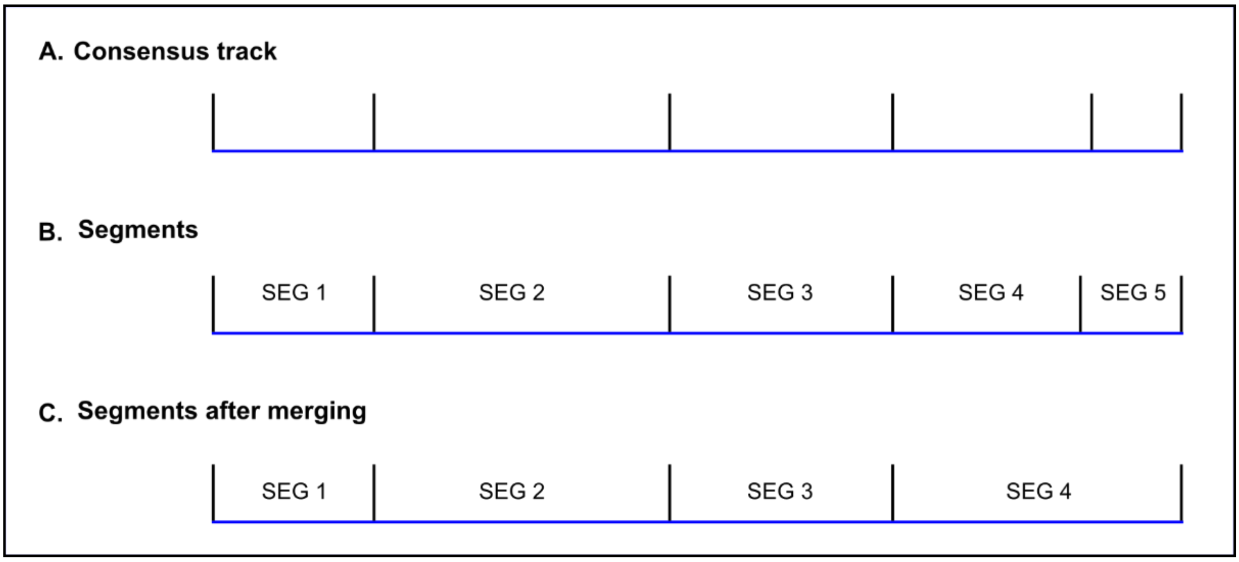

Figure 4.12 *Segmentation and merging of consensus breakpoints. (A) Consensus breakpoints (B) Segmentation based on breakpoints. (C) SEG 4 and SEG 5 from panel B are merged to make a new segment.*

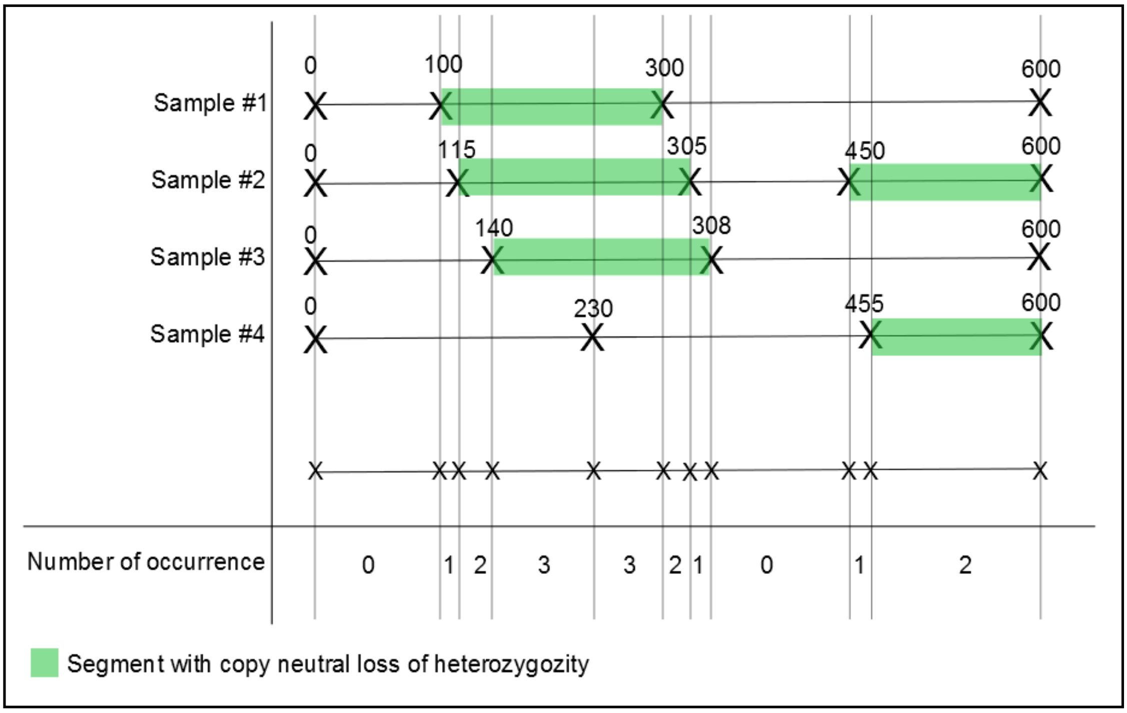

Figure 4.13 Extraction and report of regions with recurrent copy-neutral LOH across multiple samples.

4.16. Extraction of Regions with Recurrent Copy-neutral LOH across Multiple Samples

Regions with copy-neutral LOH may harbor genes with driver mutations (Barresi, Romano et al. 2010). Thus detecting regions with recurrent copy-neutral LOH across multiple samples may be a good starting point for the detection of genes with driver mutation. Therefore, a small tool was developed to find these regions. The tool reports regions in the genome with copy-neutral LOH event and frequency of occurrence in that region across multiple samples. The results are reported chromosome-wise and sorted descending based on frequency of occurrence. Figure 4.13 illustrates how the regions with recurrent copy-neutral LOH are identified and reported.

4.17. Simulator

In order to evaluate the algorithm, a simulator capable of simulating the read-depth for both normal and tumor samples and BAF based on events such as deletion, amplification and copy-neutral LOH was developed. Simulator takes as input a normal sample read-depth data. The normal sample read-depth data harbors noise and spatial correlation. By spatial correlation, it is meant that the values of the read-depths of two adjacent positions in the genome are correlated. As a result, the read-depth values of two adjacent positions tend to be close together and normally it is not expected to observe two adjacent positions in the genome having a huge difference in the read-depth. Figure 4.14 illustrates the idea of spatial correlation. If the read-depth values in a read-depth file are iterated over, each x on the X-axis represents the read-depth value seen at the current position of the read-depth file. In addition, each y on the Y-axis represents the read-depth value seen at the next position in the read-depth file. Each point (x, y) on the XY plane is associated with a value z on the Z-axis. This value denotes how many times a specific (x, y) combination has occurred in the read-depth file. For instance, (400, 100, 0) denotes that no instances has been observed where the read-depth jumps from 400 at the current position to the read-depth of 100 at the next position. As another example, (250, 254, 115) denotes that 115 times a change of read-depth from 250 reads at the current position to 254 reads at the next position has been occurred. As it can be seen from figure 4.14, the values are clustered near 0 and near 250.

The simulator, first, learns a model that takes into account the spatial correlation of adjacent positions and the noise present in the normal sample read-depth data. The model is used to simulate the read-depth for each position in the genome harboring both noise and spatial correlation. Next, the noise is removed using a median filter. The result is the simulated read-depth with only spatial correlation of adjacent positions taken into account. This result is used to construct the normal and tumor read-depths.

The normal read-depth is constructed by adding independent Poisson noise to the simulated read-depth at each position of the genome. To construct the tumor read-depth, first, two copy number tracks are constructed since chromosomes come in pairs (one for the maternal and one for the paternal). These two copy number tracks together represent the copy number status at each position of the genome. In the beginning these tracks represent a normal copy number (normal copy number = 2). Then, several regions are randomly constructed where their copy numbers deviate from normal. These regions represent events such as deletion, amplification and copy-neutral LOH. These regions are incorporated in the copy number track at random positions. Using both copy number track and the simulated read-depth, the read-depth at each position of the genome for tumor sample is calculated. Finally, the tumor read-depth is constructed by adding independent Poisson noise to each position of the genome. Figure 4.15 illustrates the pipeline of the simulator.

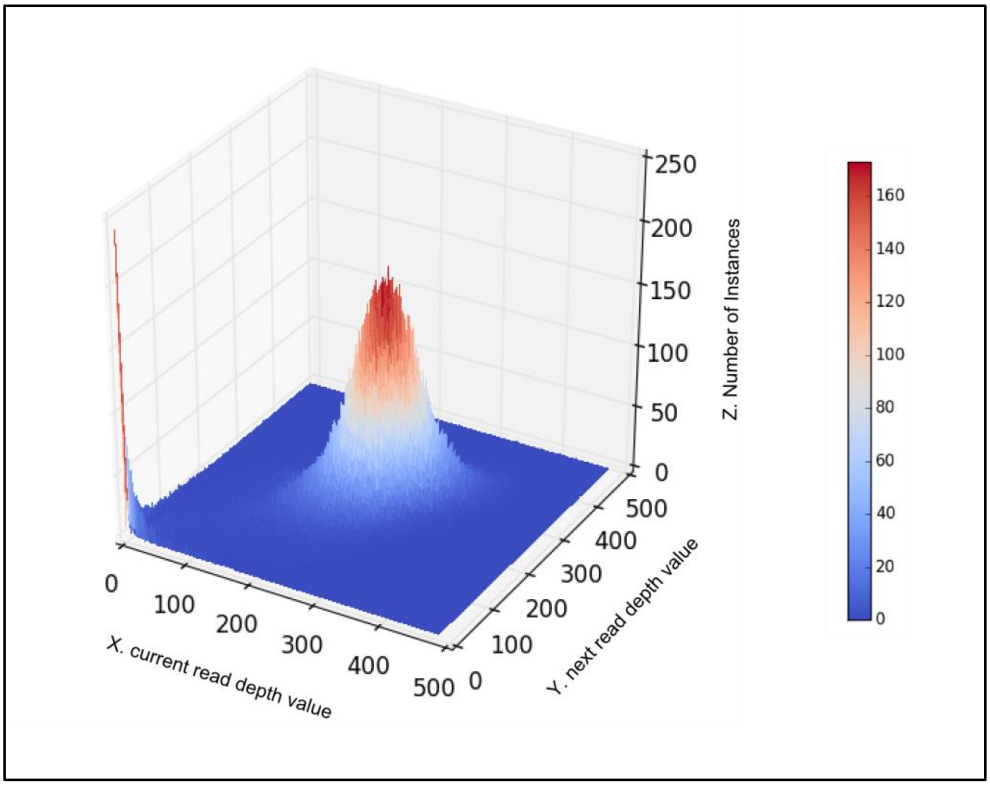

Figure 4.14 *Illustration of the spatial correlation of two adjacent data points in read-depth data. A 3-tuple (current_read_depth, next_read_depth, number_of_instances) represents how many times a specific combination of (current_read_depth, next_read_depth) is observed in the read-depth data.*

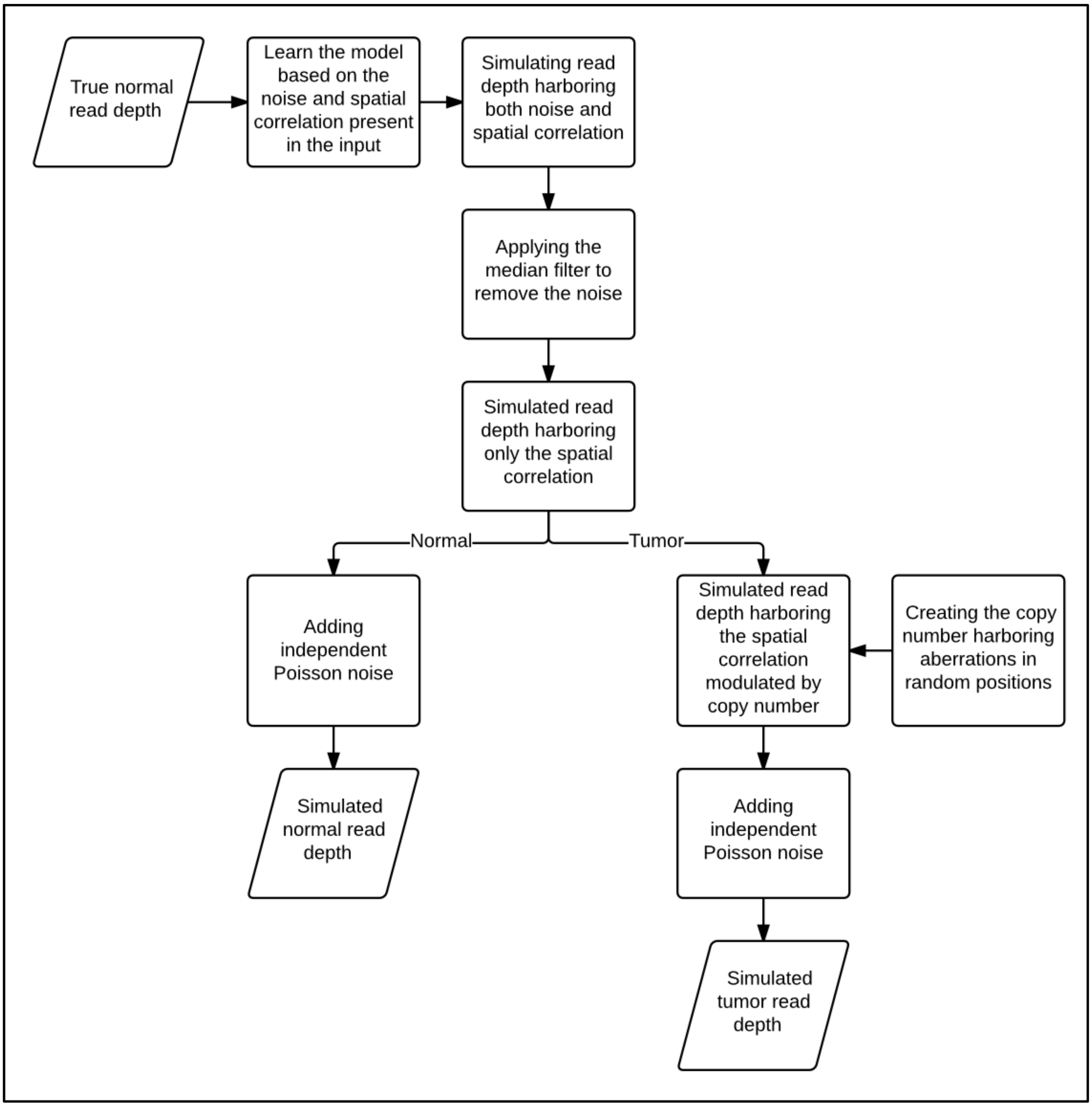
True normal read depth
Learn the model based on the noise and spatial correlation present in the input
Simulating read depth harboring both noise and spatial correlation
Applying the median filter to remove the noise
Simulated read depth harboring only the spatial correlation
Normal
Tumor
Adding independent Poisson noise
Simulated read depth harboring the spatial correlation modulated by copy number
Creating the copy number harboring aberrations in random positions
Simulated normal read depth
Adding independent Poisson noise
Simulated tumor read depth

Figure 4.15 *Pipeline of simulator.*

To construct the BAF data, first, heterozygous SNPs are randomly distributed across the genome. It was assumed that there is one heterozygous SNP per 1500 base pair in the human genome and the BAF for each of these heterozygous SNPs was calculated. The BAF for a heterozygous SNP is the fraction of the number of alternate allele to the total number of reads at the position of the SNP. The number of alternate allele at the heterozygous SNP was calculated using a *binomial distribution*. The binomial distribution has two parameters: n the number of trials (total number of reads at the position of the SNP) and p the probability of success (the probability that a read is coming from the alternate allele). The parameter n for each SNP is the read-depth at the position of the SNP in the genome. This value was extracted from the simulated normal read-depth at the position of the SNP. The probability of seeing an alternate allele at the position of a heterozygous SNP is 0.5. However, this probability changes because of changes in the copy number caused by aberrations. To calculate the parameter p, the two copy number tracks – explained earlier – was used. Once the number of alternate allele was calculated, it was used to calculate the BAF value and then it was assigned to the current heterozygous SNP.

Finally, ground truth is constructed using the two copy number tracks. This will be used in the evaluation of the developed tool. The simulated normal and tumor results were stored in a file with WIGGLE format. The BAF data was stored in a tab-separated file. The ground truth was also stored in file with SEG format for visualization purposes. Figure 4.15 illustrates the outputs of the simulator visualized in IGV.

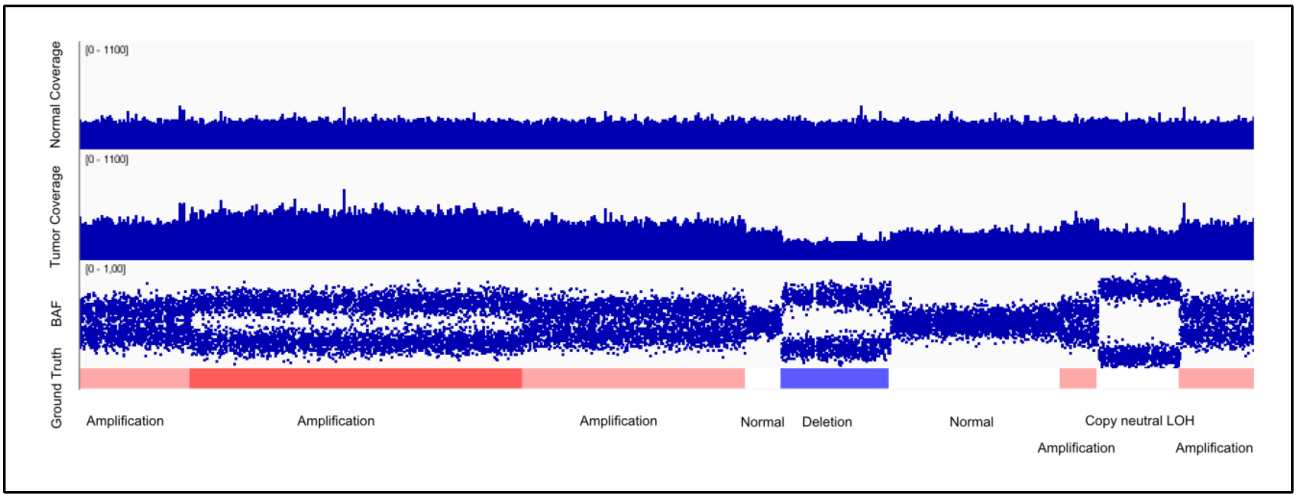

Figure 4.16 *Simulated data visualized in IGV.*

5. Results

This chapter represents the results of running the developed tool over real and simulated data. First, some examples of the tool's results visualized in IGV are presented. Furthermore, the results of the evaluation of the developed tool by running it over the simulated data are presented. Finally, the result of the benchmark against three other somatic copy number analysis tools and an explanation for each of the competing tools is presented.

5.1. Examples

The developed tool outputs a tab-delimited SEG file where each line in the file represents a detected segment, its position on the genome and the mean read-depth logratio of the segment. This file can be opened in IGV for visual inspection of the results. Figures 5.1 and 5.2 illustrate the results of the running of the tool over nine metastatic tumor samples obtained from different organs of a patient with metastatic prostate cancer. Figure 5.1 illustrates tumor protein p53 (*TP53*) copy number loss. Figure 5.2 illustrates the androgen receptor (*AR*) copy number amplification.

Figure 5.3 represents the visualized results of running the tool over samples obtained from two cohorts of ERG-positive and ERG-negative patients. The results illustrate that many of the patients from the ERG-positive cohort have obtained a copy number loss of size around 3 Mbp between ERG and *TMPRSS2* genes resulting in the formation of a *TMPRSS2-ERG* gene fusion. TMPRSS2-ERG gene fusion is a highly prevalent oncogenic alteration in prostate tumor cells and it leads to the androgenic induction of *ERG* expression (Tomlins, Rhodes et al. 2005, Furusato, Gao et al. 2008).

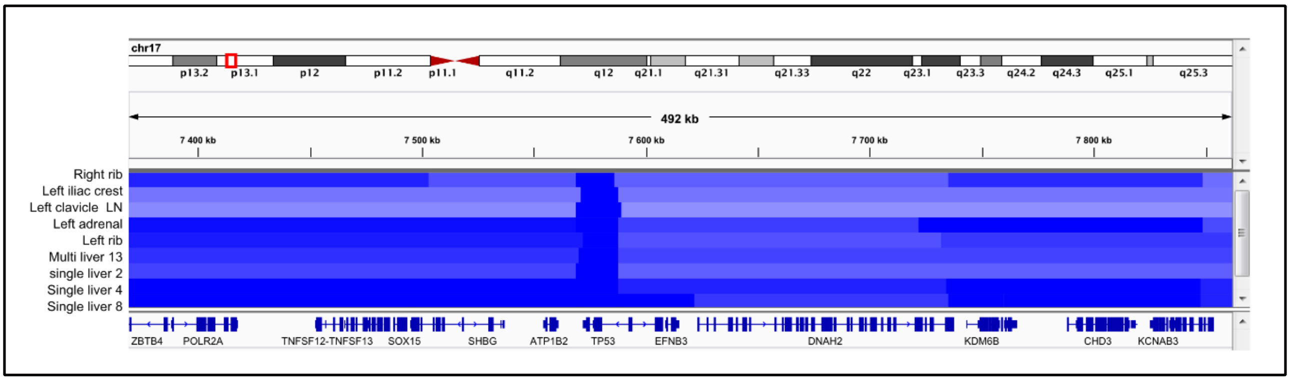

Figure 5.1 *Visualization of the results in IGV. TP53 copy number loss is observable in all the samples.*

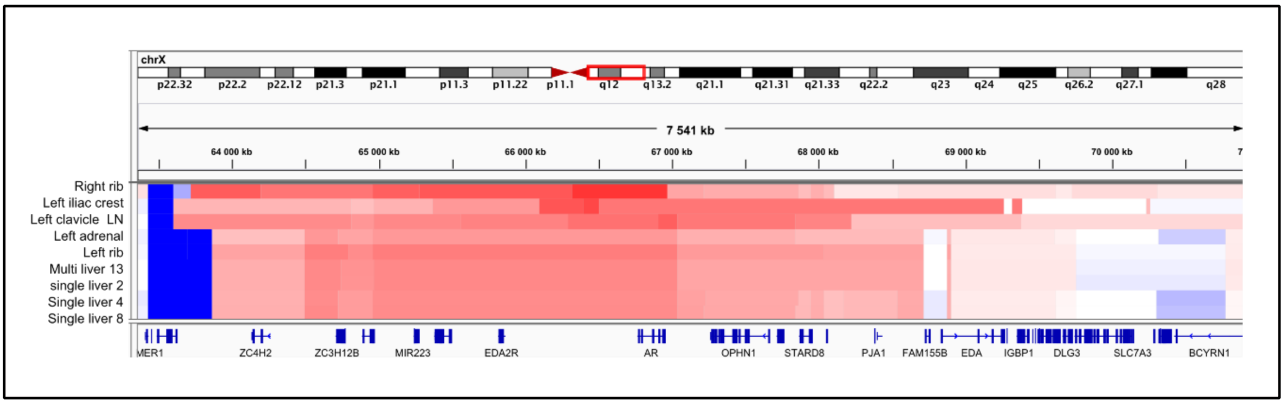

Figure 5.2 *Visualization of the results in IGV. AR copy number amplification is observable in all the samples.*

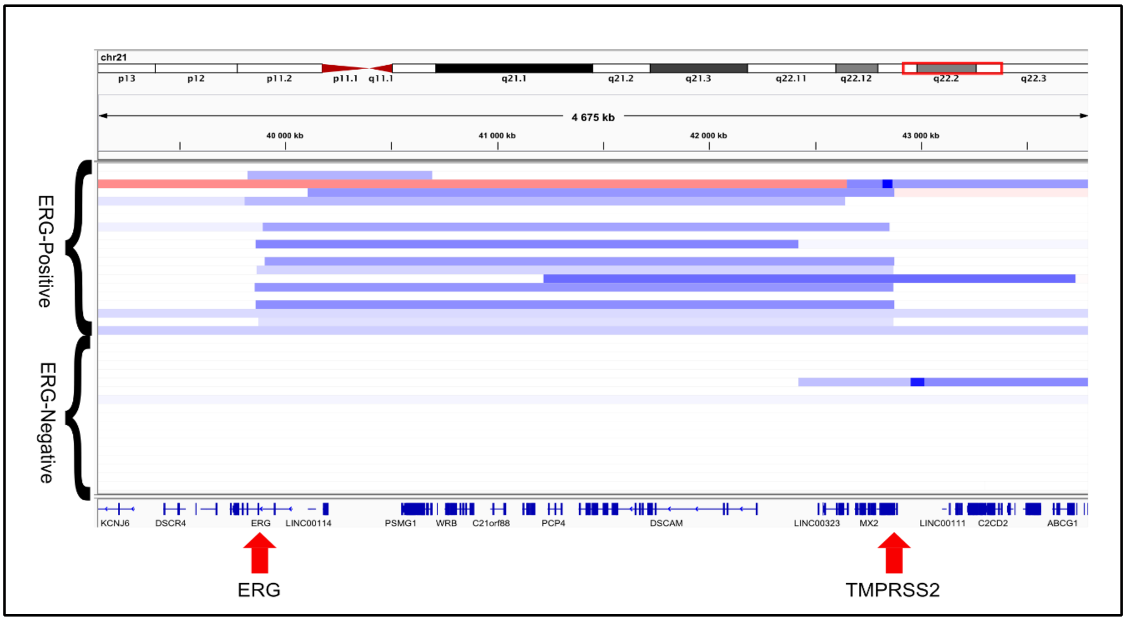

Figure 5.3 *Visualization of the results in IGV. A 3 megabase deletion between genes ERG and TMPRSS2 is observable in many of the ERG-positive samples.*

5.2. Results of Simulation Study

As explained earlier a simulator capable of simulating the read-depth for both normal and tumor samples and BAF based on events such as deletion, amplification and copy-neutral LOH was developed. The simulated data was analyzed by the tool and the results of the segmentation were checked against the 'ground truth'.

A piece of code was written to compute the number of *false positives*, *false negatives*, and *true positives* based on the comparison of the results from the tool against the ground truth in order to calculate the *false discovery ratio (FDR)*. False positive, in this context, refers to a situation where the tool places a breakpoint at a specific region of the genome where in reality there does not exist such a breakpoint. To calculate the number of false positives, the code checks whether for each detected breakpoints by the tools there is a breakpoint in the ground truth in a window of size 10 kbps centered on the detected breakpoint.

False negative, in this context, refers to a situation where in reality there exists a breakpoint at a specific region of the genome and the tool fails to place a breakpoint at that region. To calculate the number of false negatives, the code checks whether for each breakpoint in the ground truth the tool has detected a breakpoint within a window of size 10 kbps centered on the ground truth breakpoint.

True positive, in this context, refers to the actual breakpoint positions in the simulated genome. The number of true positives, in this context, can be directly extracted from the ground truth.

Having calculated the number of false positives and true positives, one can calculate the false discovery rate, defined as:

$$FDR = \frac{Number\ of\ False\ Positives}{(Number\ of\ False\ Positives + Number\ of\ True\ Positives)}$$

The lower the FDR, the better the tool is performing. The comparison of the results obtained from running the tool over simulated data against the ground truth revealed that, in general, the algorithm has a low false discovery rate of below 5%.

5.2.1. Characterization of Sources of False Positives and False Negatives

In order to characterize the sources of false positives and false negatives the output of the tool, the simulated read-depth data for both normal and tumor samples and the ground truth were loaded in IGV and the regions of the occurrence of the false positives and false negatives were inspected by eye. The visual inspection resulted in the identification of three sources of false negatives and false positives.

The first source of false negatives was small segments (typically a few kbps) in the genome. If a segment was smaller than half of the size of the smallest double sliding window, it was not detected. As a result, tool is only capable of detecting copy number segments with the resolution of half of the smallest double sliding window.

The second source of getting false negatives is caused by the random nature of the simulation. The simulator randomly places a breakpoint in the simulated tumor genome; as a result, there is a possibility for a scenario where some of the breakpoints in the ground truth fall in regions where there is not enough read-depth data. Consequently, the tool is not capable of detecting such breakpoints resulting in getting a false negative. This issue can be addressed in the future by redistributing the breakpoints, which are placed in regions of not enough data to other places with enough data.

The third source of getting both false positives and false negatives is that on some occasions the tool places the breakpoint farther away (more than 5 kbps away but not much farther) from the ground truth since the tool does not have access to enough data to decide upon the exact position of the breakpoint. As a result, both false positive and false negative counts are increased. Figure 5.4 illustrates such situation. The issue of determining the exact position of the breakpoint can be addressed, in the future, by an additional post-processing step where the sequencing reads will be used to pinpoint the exact breakpoint location up to a single-nucleotide resolution.

5.3. Benchmark

In order to benchmark the performance of the developed tool against competing tools a comprehensive search of the literature was performed and a list of copy number analysis tools was compiled. This list provides several pieces of information about each tool such as the name of each tool, the year in which each tool was published, the journal in which each tool was published, the inputs each tool requires, etc. Table 5.1 represents the list of 24 copy number analysis tools.

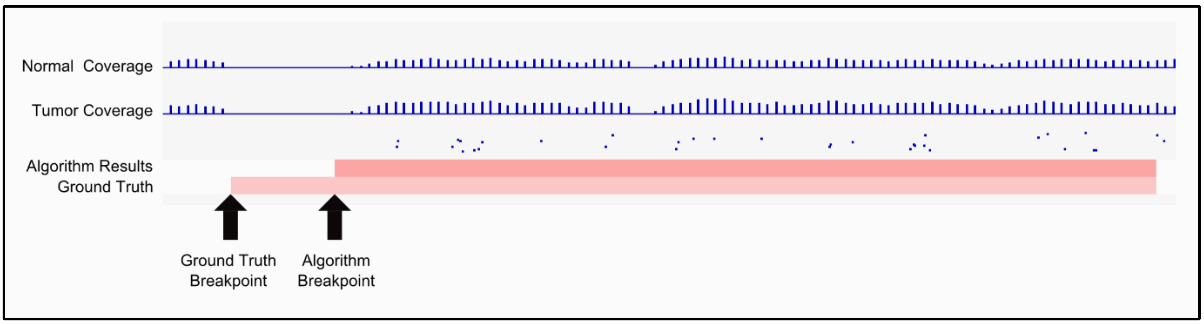

Figure 5.4 *Situation where a breakpoint is placed farther away from a ground truth breakpoint due to lack of data.*

Table 5.1 *List of copy number analysis tools.*

Name	Year	Journal	Input	Methods used	Need of control	Programming language	Requires BAF	Supported sequencing technology
SegSeq	2009	Nature Methods	BED	Statistical testing, CBS	Yes	Matlab	No	Massively parallel sequencing
readDepth	2011	PLOS ONE	BED	LOESS regression, CBS	No	R	No	Massively parallel sequencing
CNVnator	2011	Genome Res.	BAM	Mean shift Algorithm	No	C	No	WGS
RDXplorer (EWT)	2009	Genome Res.	BAM	Statistical testing	No	R, Python	No	WGS
CNAnorm	2012	Bioinformatics	SAM, BAM, GC content data	Linear regression, CBS	Yes	R	No	WGS
rSW-seq	2010	BMC Bioinformatics	Read-depth	Smith-Waterman Algorithm	Yes	C	No	WGS, Single-end sequencing
CNV-seq	2009	BMC Bioinformatics	Hits	Statistical testing	Yes	R, Perl	No	Shotgun sequencing
cn.MOPS	2012	Nucleic Acids Res.	BAM	Mixture of Poisson, EM, CBS	Multiple samples	R, C++	No	WES
JointSLM	2011	Nucleic Acids Res.	Data Matrix	HMM, ML estimator, Viterbi algorithm	Multiple samples	R, Fortran	No	WGS
GENSENG	2013	Nucleic Acids Res.	Triplet of RD signal, GC content data, Mapability data	HMM, Negative binomial regression	No	C++	No	WGS
PennCNV	2007	Genome Res.	SNP array data	HMM	No	Perl	Yes	SNP array
DNAcopy	2004	Biostatistics	aCGH data	CBS	-	R	No	Microarray
CLImAT	2014	Bioinformatics	BAM, GC content data, Mapability data	HMM	No	Matlab, C	Yes	WGS
ASCAT	2010	PNAS	logratio, BAF data	statistical modelling	optional	R	Yes	SNP array
GAP	2009	Genome Bio.	SNP array data	Pattern recognition	No	R	Yes	SNP array
GPHMM	2011	Nucleic Acid Res.	SNP array data, PFB data, GC data	HMM	No	Matlab, C	Yes	SNP array
MixHMM	2010	PLOS ONE	SNP array data	HMM	No	Python	Yes	SNP array
OncoSNP	2010	Genome Bio.	SNP array data	Single Unified Bayesian Framework	No	Matlab	Yes	SNP array
APOLLOH	2012	Genome Research	Allelic counts, Copy number segment [HMMCopy]	Non-stationary HMM	Yes	Matlab	Yes	WGS
ControlFREEC	2012	Bioinformatics	BAM, SAM, Pileup, SNP data, Mapability data	Lasso based	Optional	C++	Yes	WGS, WES
GIANT	2014	PLOS ONE	SNP array data	HMM	Yes	Matlab, C	Yes	SNP array
Patchwork	2013	Genome Biology	BAM, Pileup	CBS	Optional	R	Yes	WGS
SeqCNA	2014	BMC genomics	SAM	LOESS	No	R	No	WGS
WaveCNV	2013	Bioinformatics	Pileup	Wavelet transform	No	Matlab, Perl	Yes	WGS

5.3.1. Overview of Competing Tools

The criteria to choose the competing tools from the compiled list were that first, the tools were designed to work with WGS data. Second, the tools use BAF data in order to perform the copy number analysis; and finally, the tools were published in well-known journals or were recently published or they are popular among bioinformaticians. As can be seen from table 5.1, many of the available copy number analysis tools in this list are designed to work with SNP array data. Therefore, these tools were not considered as competing tools. Some of the tools were not publicly available at the time of compilation; these also were excluded from the list. Finally, some of the tools were not using the BAF data in order to perform the copy number analysis. In the end, three tools passed all the criteria. These tools were ControlFREEC (Boeva, Popova et al. 2012), Patchwork (Mayrhofer, DiLorenzo et al. 2013), and CLImAT (Yu, Liu et al. 2014). In what follows, these tools and their methodology are briefly described.

5.3.1.1. ControlFREEC

ControlFREEC (Control-FREE Copy number and allelic content caller) is a read-depth-based method for the detection of somatic copy number variations and LOH written in C++ programming language (Boeva, Popova et al. 2012a). ControlFREEC constructs the copy number and BAF profiles using the aligned reads and the genomic position data of known SNPs (retrieved from dbSNP database). Subsequently, ControlFREEC normalizes, segments, and analyzes the constructed profiles in order to determine the copy number and allelic content. ControlFREEC uses equally sized, non-overlapping windows in order to compute the read-depth ratios per each window. The use of control sample is optional. If a control sample is not available, ControlFREEC estimates a hypothetical read-depth for a given window using a polynomial function of the window's GC content. However, if the control sample is available, it can distinguish somatic from the germline variants. In order to perform the segmentation, ControlFREEC determines the breakpoints using Least Absolute Shrinkage eStimatOr (LASSO) regression. One of the features of the ControlFREEC is its ability to evaluate and correct for the normal cell contamination, GC-content and mapability biases while constructing the copy number profile of a tumor genome (Boeva, Popova et al. 2012). However, the users should determine the sample ploidy. If ploidy is not known, it is suggested to run the program several times with possible ploidy values and compare the results (Liu, Morrison et al. 2013).

5.3.1.2. Patchwork

Patchwork is a read-depth-based method for performing allele-specific somatic copy number analysis (Mayrhofer, DiLorenzo et al. 2013). Patchwork is written in the R programming language. Patchwork automatically calculates the average ploidy and purity of the tumor cells, and therefore does not require a prior knowledge of average ploidy (as opposed to ControlFREEC) or tumor cell content. Patchwork starts by taking aligned reads in BAM format. Then, it performs GC-normalization followed by a positional normalization. After the normalization, Patchwork uses equally sized windows of size 10 kbps in order to compute the normalized read-depth for each of the windows. These are then used in order to segment the genome using the circular binary segmentation (CBS) algorithm where each segment is assigned an average normalized coverage. Furthermore, using SAMtools the single nucleotide variant data is extracted and informative heterozygous variants using a list of known SNPs (from dbSNP database) are identified. Using these, the allelic imbalance ratios are calculated and assigned to each segment. Subsequently, Patchwork visualizes the allelic imbalance ratio and the normalized coverage for genomic segments for each chromosome. Using these visualizations, a user will be able to determine the parameters needed by Patchwork in order to assign allele-specific copy number to genomic segments or in other words, in order to determine different copy number states.

5.3.1.3. CLImAT

CLImAT (CNA and LOH Assessment in Impure and Aneuploid Tumors) suggested by Yu et al. (Yu, Liu et al. 2014) and written in C and Matlab programming language is a read-depth-based method for assessing the somatic copy number variation and LOH. CLImAT is also capable of estimating tumor impurity and ploidy. Furthermore, CLImAT does not require the use of a control sample. CLImAT starts by taking aligned reads in BAM format and a file containing the list of all known SNPs (retrieved from dbSNP database). SAMtools is used to extract read-depth data from the BAM file by counting the reads with starting position within 1000 base pair window centered on each of the known SNPs. Once the read-depth data is extracted, it is corrected for CG-content and mapability biases. Furthermore, BAF data for each SNP is calculated and normalized (using quantile normalization) in order to eliminate allelic bias. Allelic bias refers to the issue where most aligners prefer to align reads to reference allele than the alternative allele. Once the read-depth and BAF data are ready, CLImAT models them with an integrated HMM in order to infer somatic copy number variation, LOH, tumor ploidy, tumor cell content and tumor genotype.

5.4. Sample Datasets

In order to perform the benchmark of the developed tool against the competing tools, the BAM files for paired normal and tumor samples of 6 individuals diagnosed with prostate adenocarcinoma (PRAD) were selected from TCGA and downloaded from the CGHub (Appendix A contains the full name of the sample datasets used in this study). Each BAM file stores the aligned reads of a WGS experiment.

5.5. Segmentation Results

After downloading the sample datasets, parameters needed to run each of the tools were set based on the recommendations of the tool's manual. Mostly, default values for parameters were used. The configuration files and parameters values used for running the tools can be found under appendices section (Appendix B).

Different tools provide the results in different formats. In order to be able to inspect visually the results in IGV, the output of each tool were transformed into a SEG file. Following figures (figures 5.5 – 5.16) represent the visualized results obtained from running the tools over six TCGA PRAD datasets. The difference in the heatmap coloring is because the developed tool (called Segmentum in the figures) and Patchwork assign a logarithmic average copy number to a segment but ControlFREEC and CLImAT assign an integral copy number to a segment. In addition, the results illustrate only the autosomes since CLImAT was unable to perform the copy number analysis of sex chromosomes at the time of performing the analysis.

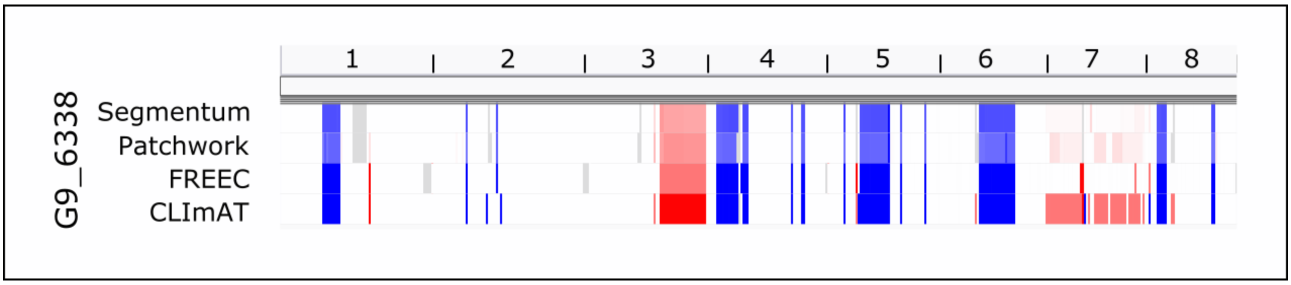

Figure 5.5 *Segmentation results obtained for sample G9_6338 (chromosomes 1-8).*

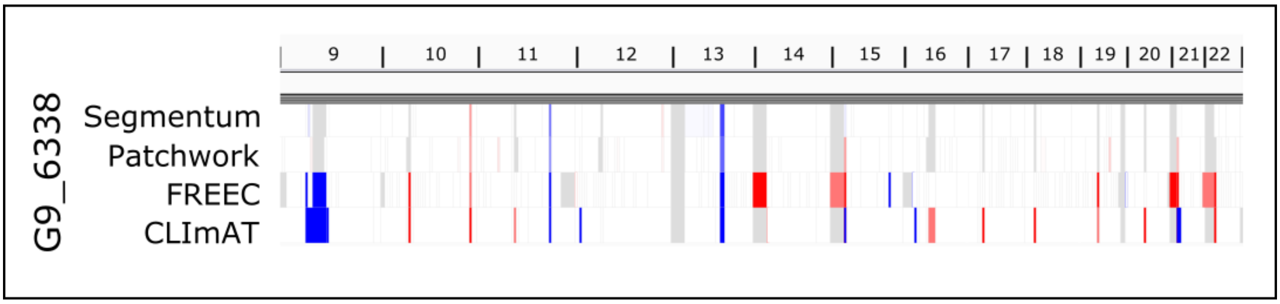

Figure 5.6 *Segmentation results obtained for sample G9_6338 (chromosomes 9-22).*

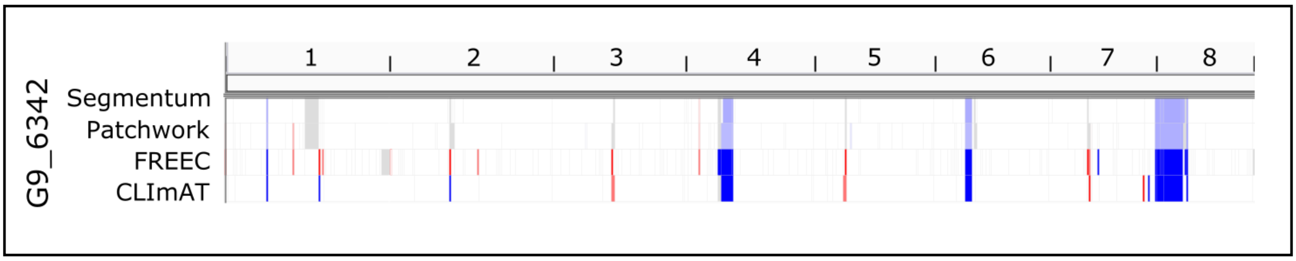

Figure 5.7 *Segmentation results obtained for sample G9_6342 (chromosomes 1-8).*

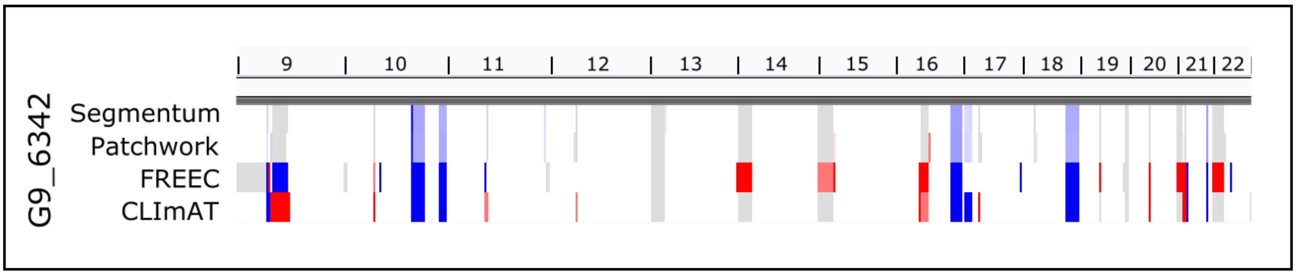

Figure 5.8 *Segmentation results obtained for sample G9_6342 (chromosomes 9-22).*

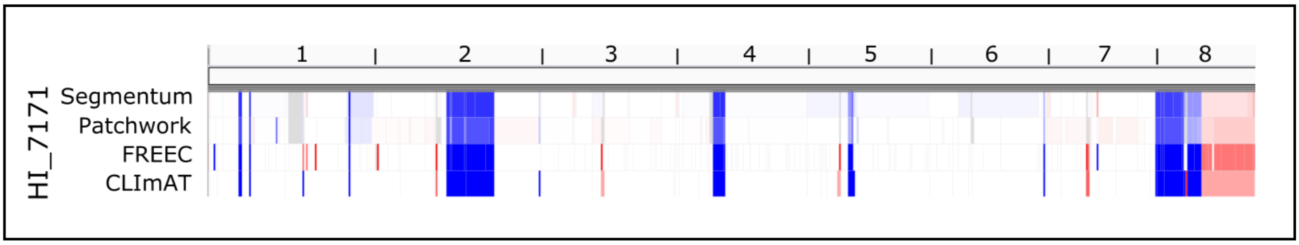

Figure 5.9 *Segmentation results obtained for sample HI_7171 (chromosomes 1-8).*

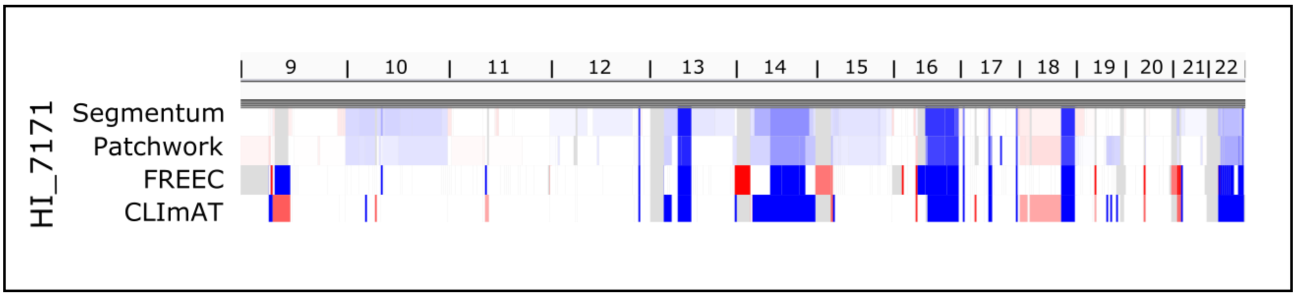

Figure 5.10 *Segmentation results obtained for sample HI_7171 (chromosomes 9-22).*

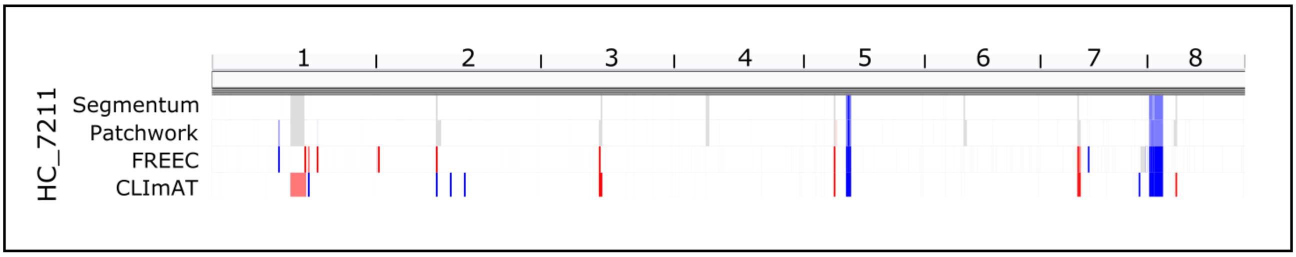

Figure 5.11 *Segmentation results obtained for sample HC_7211 (chromosomes 1-8).*

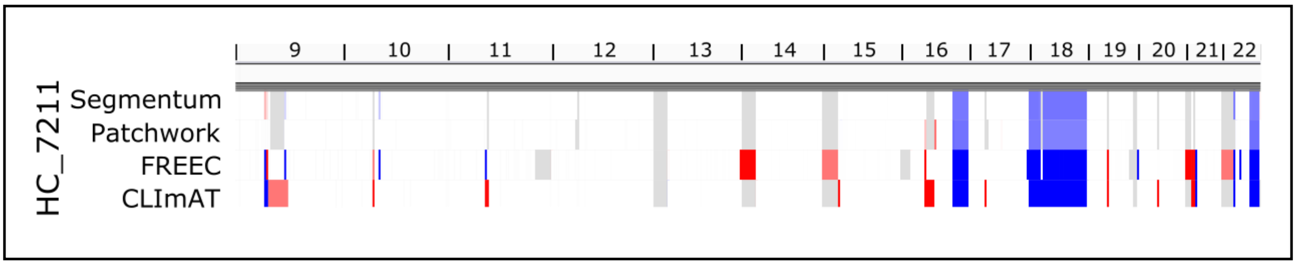

Figure 5.12 *Segmentation results obtained for sample HC_7211 (chromosomes 9-22).*

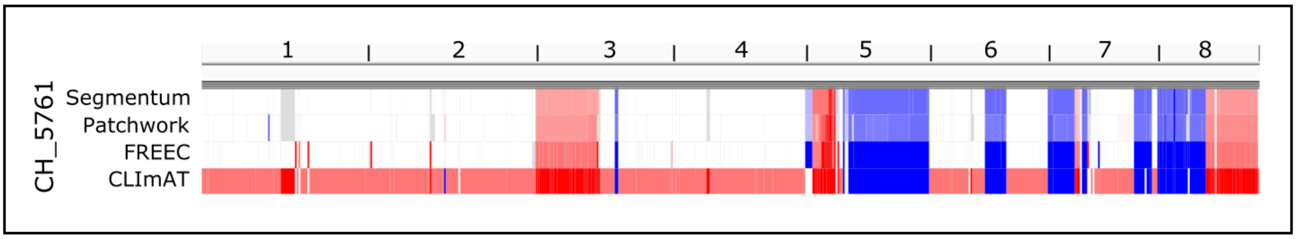

Figure 5.13 *Segmentation results obtained for sample CH_5761 (chromosomes 1-8).*

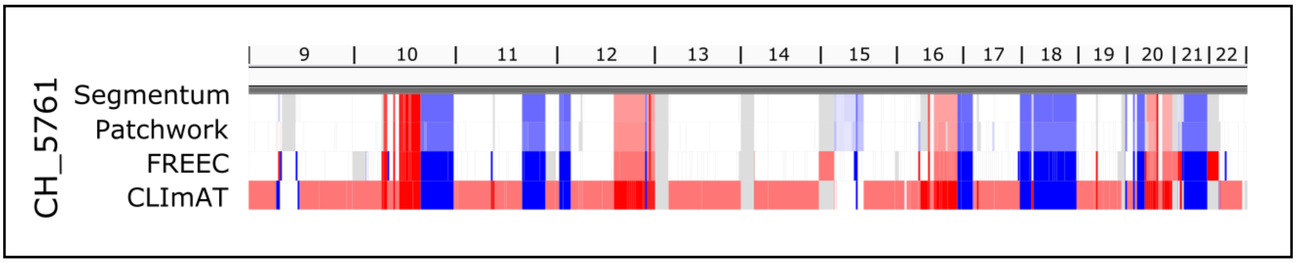

Figure 5.14 *Segmentation results obtained for sample CH_5761 (chromosomes 9-22).*

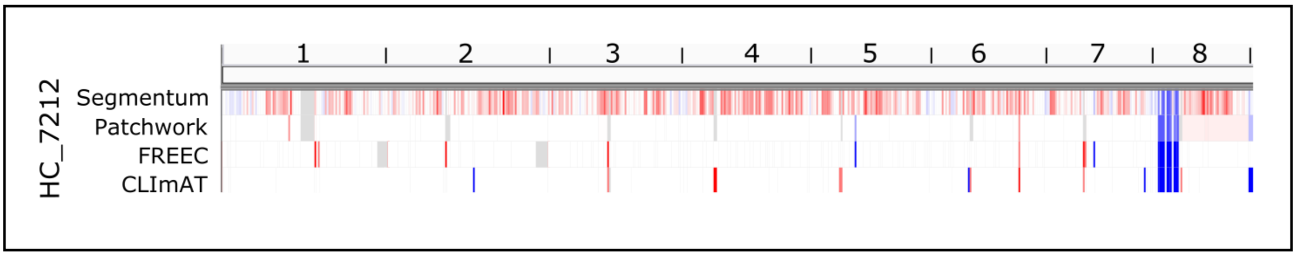

Figure 5.15 *Segmentation results obtained for sample HC_7212 (chromosomes 1-8).*

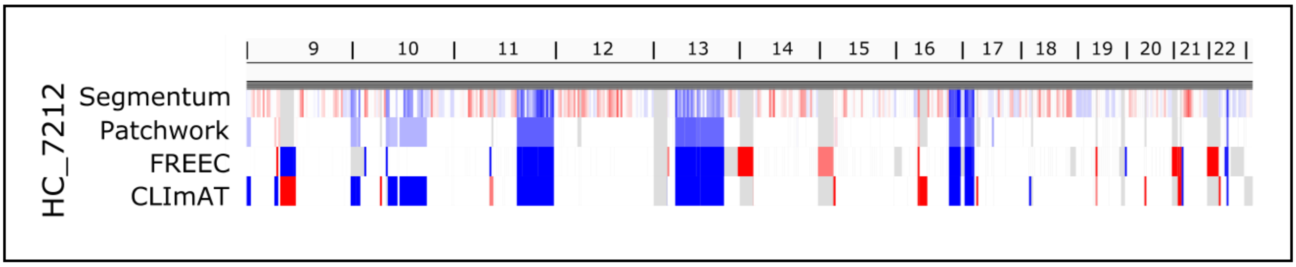

Figure 5.16 *Segmentation results obtained for sample HC_7212 (chromosomes 9-22).*

Inspecting the results visually, it can be seen that there are genomic regions annotated as aberrant by all of the four tools, or there are cases where a genomic region is only picked by three, two or only one of the tools. In order to quantitatively measure the percentage of regions throughout the whole genome (except the sex chromosomes) annotated by either all four, three, two or only one of the tools, a small program was developed. First, the program breaks the genome into blocks of same size (100 base pairs). Next, for each block, by inspecting the results from each of the tools the program checks whether the block is annotated as aberrant (either partially or completely) by any of the tools. If this is the case, it keeps track of the name of the tools that have annotated the block as aberrant. Subsequently, the program aggregates the results and calculates the percentages. Furthermore, the program calculates the pairwise concordances between tool-pairs. Pairwise concordance represents how much the results of two tools are concordant with each other. Following figures (figures 5.17 - 5.22) illustrate the results from the program in the form of Venn diagrams. For instance, from the figure 5.17, it can be seen that 47.69% of the blocks were annotated by all the tools to be aberrant. From the same figure, it can be seen that around 28% of the blocks are annotated as aberrant solely by CLImAT.

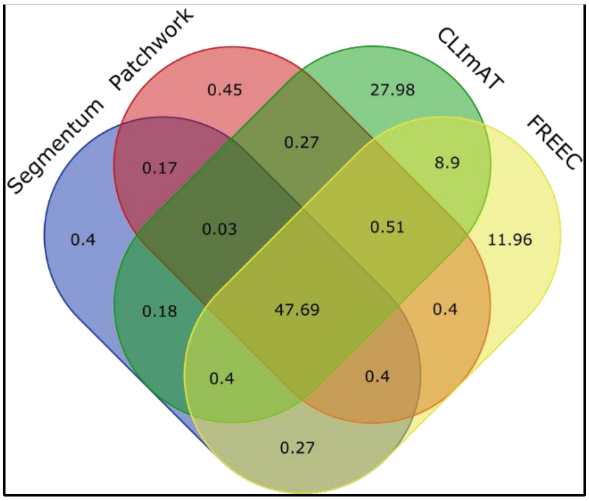

Figure 5.17 Venn diagram of the results for sample G9_6338.

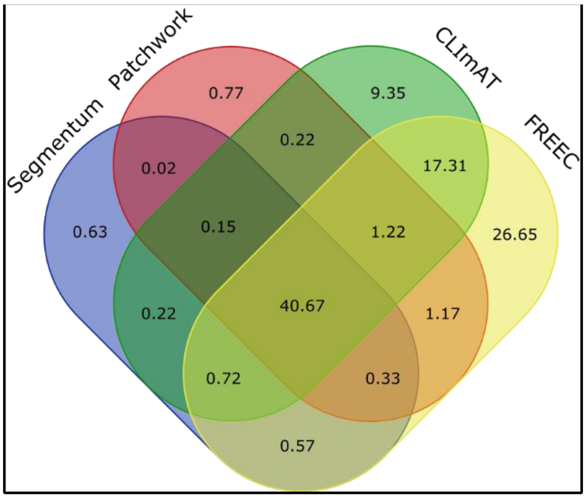

Figure 5.18 Venn diagram of the results for sample G9_6342.

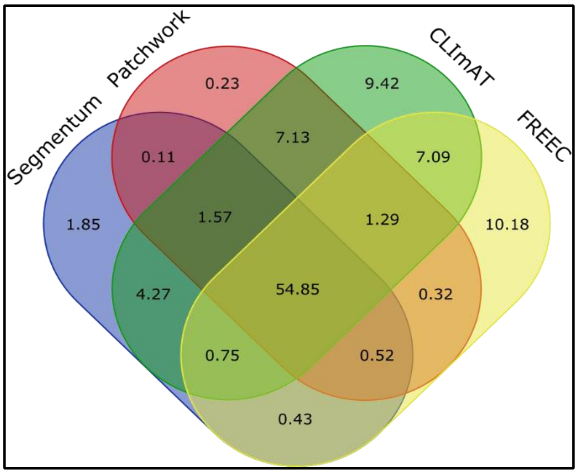

Figure 5.19 Venn diagram of the results for sample HI_7171.

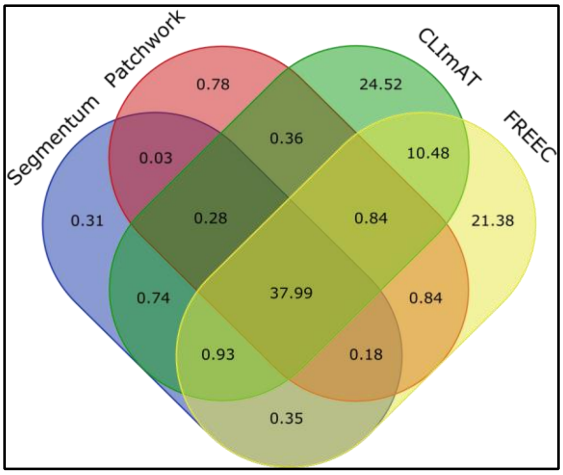

Figure 5.20 Venn diagram of the results for sample HC_7211.

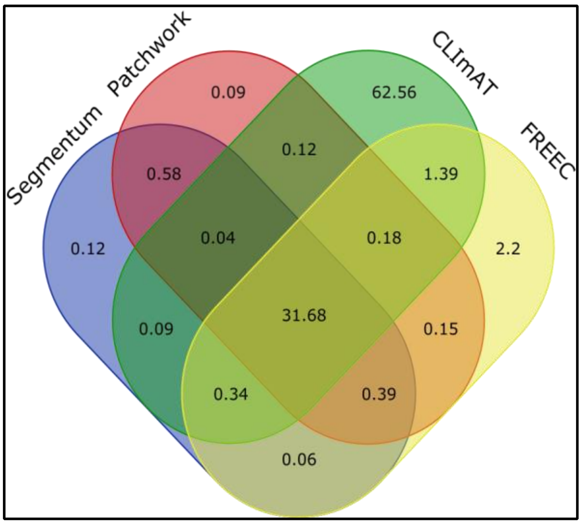

Figure 5.21 Venn diagram of the results for sample CH_5761.

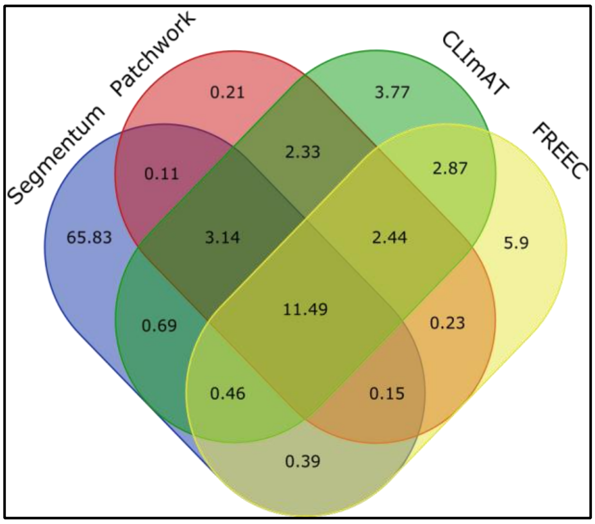

Figure 5.22 Venn diagram of the results for sample HC_7212.

From the Venn diagrams results it can be observed that, on average, only 37.4% of the regions are annotated as aberrant by all four tools across six samples. In addition, from the Venn diagrams (excluding the Venn diagrams for samples *CH_5761* and *HC_7212*), it can be seen that ControlFREEC and CLImAT tend to individually annotate as aberrant a large portion of the genome. For instance, in sample *HC_7211*, 24.52% of the annotated regions as aberrant by CLImAT are not detected by the other three tools. In the same sample, 21.38% of regions detected by ControlFREEC are not detected by the other three tools. At first, it might seem that these two tools are capable of detecting aberrations not detectable by the other two tools. However, one explanation for such behavior, as is evident from IGV visualizations, is that ControlFREEC and CLImAT tend to annotate the centromeres and telomeres as aberrant while the developed tool and Patchwork do not annotate these regions.

In sample *CH_5761*, 62.56% of the annotated regions as aberrant by CLImAT are not detected by the other three tools. Looking at the visualizations for this sample shows that these regions are annotated as *amplified* by CLImAT. CLImAT reports an estimated average cancer DNA index for each sample. DNA index is a measure of the ploidy of cells. It is defined as the ratio between the DNA content of tumor cell and the DNA content of a normal cell. As a result, DNA index equal 1.0 denotes a diploid cell and DNA indices below 0.96 or above 1.04 denote aneuploidy (Ormerod, 2008). The DNA index for sample CH_5761 estimated by CLImAT is 1.4228 suggesting that the whole genome has become aneuploid. However, it is difficult to determine whether this observation is real or not since the other three tools do not annotate this sample the same way as CLImAT does.

The same observation can be seen for sample *HC_7212*. 65.83% of the annotated regions as aberrant by the developed tool are not detected by the other three tools. The understanding of the underlying reasons for such a behavior requires more in-depth deliberation. This issue will be addressed in future studies.

As explained earlier, to see which pair of tools produces more concordant results, concordance measures were calculated for each tool-pair and for each sample. To calculate the concordance from the Venn diagram, the regions in the Venn diagram that are shared between both tools in a tool-pair is summed and then divided by the sum of the regions covered by both tools in a tool-pair. The following formula represents this mathematically:

$$\text{Pairwise concordance} = \frac{\bigcap pair}{\bigcup pair}$$

Table 5.2 represents the list of concordance values for each tool-pair for each sample. As it can be observed from the list (except sample HC_7212), the results from Patchwork and Segmentum are producing the most concordant results.

Table 5.2 *List of pairwise concordances.*

	G9_6338	G9_6342	HI_7171	HC_7211	CH_5761	HC_7212
Patchwork-Segmentum	94%	88%	78%	88%	97%	17%
FREEC-Patchwork	69%	48%	68%	54%	87%	48%
FREEC-Segmentum	68%	47%	68%	53%	87%	13%
CLImAT-FREEC	58%	61%	65%	51%	34%	51%
CLImAT-Patchwork	56%	59%	74%	51%	33%	70%
CLImAT-Segmentum	55%	59%	69%	52%	33%	17%

6. Discussion

The main objective of this study was to develop a tool capable of detecting genomic aberrations such as deletion, amplification and copy-neutral LOH events in the cancer genome using the WGS data. As explained in the literature review chapter, there are 4 main methods available in order to detect such aberrations. These four methods are (1) read-pair methods, (2) split-read methods, (3) sequence assembly methods, and (4) read-depth methods (Alkan, Coe et al. 2011). In this study, read-depth method was used. The reason behind this choice was that this method is intuitive and it is the mostly adapted method by the research community in order to perform copy number analysis.

One of the advantages of the developed tool over some other available CNA analysis tools is that it utilizes the BAF data in order to find the copy number segments. The use of BAF data not only improves the accuracy of copy number detection but also enables the identification of the copy-neutral LOH events. Tools that do not utilize such data are unable to detect copy-neutral LOH events. Even though there are many SNP array-based CNA analysis tools utilizing BAF data, there are not many WGS-based tools available. Another advantage of the developed tool is that it is fast. The developed tool is capable of performing the copy number analysis in less than two minutes.

The developed tool has originally been designed to analyze WGS data and currently, it cannot be applied to WES data. This is because the sparse and non-uniform distribution of exons across the genome requires different types of data processing (Krumm, Sudmant et al. 2012). However, some tools such as ControlFREEC are capable of analyzing both WGS and WES data (Boeva, Popova et al. 2012).

Currently, the developed tool requires a paired normal sample in order to correct for GC-content and mapability biases and to discriminate germline CNVs from the somatic CNAs. The assumption behind the requirement of a paired normal sample in order to correct for such biases is that GC-content and mapability biases affect the normal and the tumor samples in a similar way. As a result, during the calculation of the read-depth ratios, these biases are implicitly corrected. However, the availability of many tumor samples without accompanying matched normal sample calls for the development of a general-purpose copy number analysis tool capable of performing bias correction without requiring the matched normal sample. There are already some strategies available in order to address this issue. For instance, Scheinin et al. (2014) are employing a binning strategy. In this approach, the genome is partitioned into fixed-sized bins and then the median read counts for all bins with the same combinations of GC and mapability are found. Next, a LOESS surface is fit through the medians and finally, the GC and mapability biases is corrected by dividing the raw counts by its corresponding LOESS value (Scheinin, Sie et al. 2014).

The developed tool is capable of identifying copy number segments with the resolution of half of the size of the smallest double sliding window (typically a few kbps). As a result, in order to locate the exact position of segment boundaries up to a single-nucleotide resolution, the mapped sequencing reads can be inspected searching for the breakpoints in the reads. There are some tools available to perform such post processing of the results.

Furthermore, the developed tool currently assigns a mean read-depth logratio to each segment, which is the average read-depth logratio of all the data points that fall into a segment. As a result, the developed tool is incapable of assigning an integral copy number value to a segment. However, some

HMM-based segmentation tools such as CLImAT are capable of assigning an integral copy number value to a segment by means of their hidden states where each hidden state is corresponding to an integral copy number (Klambauer, Schwarzbauer et al. 2012).

Currently, the developed tool does not address the issues of normal cell contamination (tumor purity), tumor ploidy and tumor heterogeneity as opposed to tools such as Patchwork and CLImAT that are capable of inferring the tumor purity and tumor ploidy. Tumor heterogeneity refers to the observation that a tumor may be comprised of multiple clones of tumor cells. In principle, it is possible to infer the tumor purity, tumor ploidy and tumor heterogeneity from the observed read-depth logratios and BAF values. This is possible since there are only certain patterns of read-depth logratios and BAF values possible for a given copy number state and deviations from these patterns may lead to the inference of tumor ploidy, tumor purity and tumor heterogeneity (Liu, Morrison et al. 2013, Oesper, Mahmoody et al. 2013). For instance, in a triploid genomic region (i.e. copy number equal to 3) of a tumor sample with 100% of tumor purity, BAF values can be either 0, 0.33, 0.66 or 1 while for the same region with 50% of tumor purity, BAF values are shifted to 0, 0.415, 0.58 and 1.

In the result section, it was observed that different tools produce different results and on average only 37.4% of the regions was annotated as aberrant by all four tools across six samples. One can speculate few explanations for such observation. The difference in the results may be the byproduct of the underlying assumptions of different tools. Alternatively, it might be because of the parameter values that were set for the running of each tool. In this study, mostly (unless otherwise stated) the default parameter values were used assuming that the developers of the methods had better knowledge about the optimal parameter values (Appendix B contains the configuration files and the parameter values used in this study to run the tools). For instance, ControlFREEC requires the user to specify the ploidy of the samples and, in this study, to run this tool it was assumed that all the samples were diploid. The tool might have produced different results if prior knowledge about the ploidy of the samples was available.

Even though in this study a simulator capable of producing the read-depth and BAF data for normal and tumor sample was developed, the format of the simulated data was not compatible with the input requirements of the competing tools. As a result, the simulated data was only used in order to evaluate the performance of the developed tool. One approach that some studies adopt in order to evaluate the result of their tools is to evaluate the obtained result against the result of a SNP array experiment on the same datasets. In this approach, the tumor samples are also assayed by SNP array technology. Then, the SNP array data is analyzed by a SNP array-based copy number analysis tool such as ASCAT (Van Loo, Nordgard et al. 2010). In this approach, the results of the copy number analysis from the SNP array analysis is considered as the ground truth. This approach for instance was used by CLImAT (Yu, Liu et al. 2014). However, this approach can introduce more costs to the study. Unfortunately, the SNP array data was not available for the TCGA PRAD data used in this study. Recently, there have been attempts to simulate HTS data that harbors all the existing biases in the sequencing data; however, these attempts have not yet been successful in capturing all the features of tumor genome. As a result, the unavailability of a comprehensive ground truth or gold standard to evaluate the results obtained from different CNA detection tools remains a challenge and until such option is not available it would be difficult to evaluate objectively how the tools are performing (Liu, Morrison et al. 2013). Luckily, the research on developing simulated data is ongoing holding the promise that in the near future we will have access to such data.

Finally, in this study, only Venn diagram and the concordance measure were used in order to assess the results from different tools. However, one can employ different evaluation criteria in order to ensure the robustness of the evaluation study.

6.1. Further Research

The future work can address some of the limitations in this work. For instance, this study can be furthered by developing a data analysis scheme capable of correcting for biases such as GC-content and mapability without the requirement of a paired normal sample. Furthermore, the tool can be expanded in order to address the issues of normal cell contamination (tumor purity), tumor ploidy and tumor heterogeneity. In addition, the developed simulator can be expanded in a way that it produces the data usable by other tools. Once these issues are addressed, one can tackle the problem of tumor evolution.

7. Conclusion

In this study, the main aim was to develop a tool capable of detecting somatic copy number alterations such as deletion, amplification, and copy-neutral LOH in the tumor genome using WGS data. To do this, read-depth detection method was adapted. Read-depth methods detect the amplifications and deletions in the genome by examining respectively the increase and decrease in read-depth data extracted from BAM files. However, the read-depth data needs to be corrected for biases such as mapability and GC-content biases. In this study, it was assumed that the utilization of a matched normal sample in the calculation of the read-depth ratio of the tumor and normal sample would largely and implicitly correct for mapability and GC-content biases assuming that these biases affect both the normal and the tumor samples in a similar way. Furthermore, in order to detect genomic regions with copy-neutral LOH, the BAF data extracted from the BAM files were used. Finally, a double sliding window method was employed in order to segment the genome into either normal or aberrated regions harboring events such as deletion, duplication, or copy-neutral LOH based on both read-depth and BAF data.

As a result, a fast and easy-to-use tool capable of performing copy number analysis was developed. The developed tool is capable of detecting amplifications, deletion and copy-neutral LOH. Furthermore, a simulator capable of simulating the read-depth and BAF data was developed. The simulated data was used to perform a simulation study. The results from the simulation study were first used to pinpoint the issues in the implementation and to improve the performance of the tool in terms of sensitivity and specificity. Furthermore, the results from the simulation study showed that the developed tool was performing well.

In addition to the simulation study, a benchmark was performed where the developed tool along with three other competing tools analyzed six prostate adenocarcinoma samples downloaded from the cancer genome atlas (TCGA). The results from the benchmark showed that the developed tool was capable of producing comparable results in comparison to other competing tools. Finally, in the discussion chapter the limitations of the developed tool such as its inability to infer the tumor purity and ploidy was addressed and further possible research to improve the tool in the future was outlined.

As explained earlier, CNAs and copy-neutral LOH are two important drivers of genomic instability associated with cancer and accurate detection of these aberrations holds the promise to increase our knowledge of the underlying processes of cancer by identifying novel cancer related oncogenes and tumor suppressor genes. This consequently may help in finding new ways in treating cancer.

References

1. *Encyclopedic dictionary of polymers,* 2011, 2nd ed, Springer, New York; London.
2. *Encyclopedic reference of cancer,* 2001, Springer, Berlin; London.
3. Alkan, C., Coe, B.P. & Eichler, E.E. 2011, "Genome structural variation discovery and genotyping", *Nature reviews. Genetics,* vol. 12, no. 5, pp. 363-376.
4. Alkodsi, A., Louhimo, R. & Hautaniemi, S. 2015, "Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data", *Briefings in bioinformatics,* vol. 16, no. 2, pp. 242-254.
5. Barresi, V., Romano, A., Musso, N., Capizzi, C., Consoli, C., Martelli, M.P., Palumbo, G., Di Raimondo, F. & Condorelli, D.F. 2010, "Broad copy neutral-loss of heterozygosity regions and rare recurring copy number abnormalities in normal karyotype-acute myeloid leukemia genomes", *Genes, chromosomes & cancer,* vol. 49, no. 11, pp. 1014-1023.
6. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., Mc Henry, K.T., Pinchback, R.M., Ligon, A.H., Cho, Y.J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M.S., Weir, B.A., Tanaka, K.E., Chiang, D.Y., Bass, A.J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F.J., Sasaki, H., Tepper, J.E., Fletcher, J.A., Tabernero, J., Baselga, J., Tsao, M.S., Demichelis, F., Rubin, M.A., Janne, P.A., Daly, M.J., Nucera, C., Levine, R.L., Ebert, B.L., Gabriel, S., Rustgi, A.K., Antonescu, C.R., Ladanyi, M., Letai, A., Garraway, L.A., Loda, M., Beer, D.G., True, L.D., Okamoto, A., Pomeroy, S.L., Singer, S., Golub, T.R., Lander, E.S., Getz, G., Sellers, W.R. & Meyerson, M. 2010, "The landscape of somatic copy-number alteration across human cancers", *Nature,* vol. 463, no. 7283, pp. 899-905.
7. Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O. & Barillot, E. 2012, "Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data", *Bioinformatics (Oxford, England),* vol. 28, no. 3, pp. 423-425.
8. Broadinstitute.org 2015, *SEG: Segmented Data File Format.* Available: http://www.broadinstitute.org/igv/SEG [2015, January/22].
9. Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., Beroukhim, R., Pellman, D., Levine, D.A., Lander, E.S., Meyerson, M. & Getz, G. 2012, "Absolute quantification of somatic DNA alterations in human cancer", *Nature biotechnology,* vol. 30, no. 5, pp. 413-421.
10. Cghub.ucsc.edu 2015, *Cancer Genomics Hub.* Available: https://cghub.ucsc.edu/ [2015, January 23].
11. Chen, G.K., Chang, X., Curtis, C. & Wang, K. 2013, "Precise inference of copy number alterations in tumor samples from SNP arrays", *Bioinformatics (Oxford, England),* vol. 29, no. 23, pp. 2964-2970.
12. Chen, J.M., Cooper, D.N., Chuzhanova, N., Ferec, C. & Patrinos, G.P. 2007, "Gene conversion: mechanisms, evolution and human disease", *Nature reviews.Genetics,* vol. 8, no. 10, pp. 762-775.
13. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., Shi, X., Fulton, R.S., Ley, T.J., Wilson, R.K.,

Ding, L. & Mardis, E.R. 2009, "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation", *Nature methods,* vol. 6, no. 9, pp. 677-681.

14. Chin, L., Hahn, W.C., Getz, G. & Meyerson, M. 2011, "Making sense of cancer genomic data", *Genes & development,* vol. 25, no. 6, pp. 534-555.
15. Crisan, A. & Xiang, J. 2009, "Comparison of Hidden Markov Models and Sparse Bayesian Learning for Detection of Copy Number Alterations", *Canadian Student Conference on Biomedical Computing*.
16. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Graf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., METABRIC Group, Langerod, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Borresen-Dale, A.L., Brenton, J.D., Tavare, S., Caldas, C. & Aparicio, S. 2012, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups", *Nature,* vol. 486, no. 7403, pp. 346-352.
17. Dancey, J.E., Bedard, P.L., Onetto, N. & Hudson, T.J. 2012, "The genetic basis for cancer treatment decisions", *Cell,* vol. 148, no. 3, pp. 409-420.
18. Di Cristofano, A., Pesce, B., Cordon-Cardo, C. & Pandolfi, P.P. 1998, "Pten is essential for embryonic development and tumour suppression", *Nature genetics,* vol. 19, no. 4, pp. 348-355.
19. Dong, J.T. 2001, "Chromosomal deletions and tumor suppressor genes in prostate cancer", *Cancer metastasis reviews,* vol. 20, no. 3-4, pp. 173-193.
20. Duan, J., Zhang, J., Deng, H. & Wang, Y. 2013, "Comparative Studies of Copy Number Variation Detection Methods for Next-Generation Sequencing Technologies", *PLoS ONE,* vol. 8, no. 3, pp. 1-12.
21. Eddy, S.R. 2004, "What is a hidden Markov model?", *Nature biotechnology,* vol. 22, no. 10, pp. 1315-1316.
22. Ensembl.org 2015, *WIG File Format - Definition and supported options*. Available: http://www.ensembl.org/info/website/upload/wig.html [2015, January/22].
23. Flury, B. 1997, *A first course in multivariate statistics,* Springer, New York.
24. Furusato, B., Gao, C.L., Ravindranath, L., Chen, Y., Cullen, J., McLeod, D.G., Dobi, A., Srivastava, S., Petrovics, G. & Sesterhenn, I.A. 2008, "Mapping of TMPRSS2-ERG fusions in the context of multi-focal prostate cancer", *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc,* vol. 21, no. 2, pp. 67-75.
25. genome.ucsc.edu 2015, *UCSC Genome Browser: Wiggle Track Format (WIG)*. Available: http://genome.ucsc.edu/goldenpath/help/wiggle.html [2015, January/22].
26. genomewiki.ucsc.edu 2015, *Selecting a graphing track data format - Genomewiki*. Available: http://genomewiki.ucsc.edu/index.php/Selecting_a_graphing_track_data_format[2015, January/22].
27. Globocan.iarc.fr 2014, *GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012*. Available: Globocan.iarc.fr [2014, November 12].
28. Griffiths, A.J.F., Gelbart, W.M., Miller, J.H. & Lewontin, R.C. 1998, *Modern Genetic Analysis,* W H Freeman & Co.
29. Guo, Y., Sheng, Q., Samuels, D.C., Lehmann, B., Bauer, J.A., Pietenpol, J. & Shyr, Y. 2013, "Comparative study of exome copy number variation estimation tools using array

comparative genomic hybridization as control", *BioMed research international,* vol. 2013, pp. 915636.

30. Gusnanto, A., Wood, H.M., Pawitan, Y., Rabbitts, P. & Berri, S. 2012, "Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data", *Bioinformatics (Oxford, England),* vol. 28, no. 1, pp. 40-47.
31. Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., Shumansky, K., Chin, S.F., Turashvili, G., Hirst, M., Caldas, C., Marra, M.A., Aparicio, S. & Shah, S.P. 2012, "Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer", *Genome research,* vol. 22, no. 10, pp. 1995-2007.
32. Hanahan, D. & Weinberg, R.A. 2011, "Hallmarks of cancer: the next generation", *Cell,* vol. 144, no. 5, pp. 646-674.
33. Hanahan, D. & Weinberg, R.A. 2000, "The hallmarks of cancer", *Cell,* vol. 100, no. 1, pp. 57-70.
34. Hemmer, S., Wasenius, V.M., Haglund, C., Zhu, Y., Knuutila, S., Franssila, K. & Joensuu, H. 2001, "Deletion of 11q23 and cyclin D1 overexpression are frequent aberrations in parathyroid adenomas", *The American journal of pathology,* vol. 158, no. 4, pp. 1355-1362.
35. Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E.E. & Sahinalp, S.C. 2010, "Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery", *Bioinformatics (Oxford, England),* vol. 26, no. 12, pp. i350-7.
36. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. 2012, "De novo assembly and genotyping of variants using colored de Bruijn graphs", *Nature genetics,* vol. 44, no. 2, pp. 226-232.
37. Isola, J., Chu, L., DeVries, S., Matsumura, K., Chew, K., Ljung, B.M. & Waldman, F.M. 1999, "Genetic alterations in ERBB2-amplified breast carcinomas", *Clinical cancer research : an official journal of the American Association for Cancer Research,* vol. 5, no. 12, pp. 4140-4145.
38. Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F. & Pinkel, D. 1992, "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors", *Science,* vol. 258, no. 5083, pp. 818.
39. Kallioniemi, A., Visakorpi, T., Karhu, R., Pinkel, D. & Kallioniemi, O.P. 1996, "Gene Copy Number Analysis by Fluorescence in Situ Hybridization and Comparative Genomic Hybridization", *Methods,* vol. 9, no. 1, pp. 113.
40. Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.A., Mitterecker, A., Bodenhofer, U. & Hochreiter, S. 2012, "cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate", *Nucleic acids research,* vol. 40, no. 9, pp. e69.
41. Korbel, J.O., Abyzov, A., Mu, X.J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M. & Gerstein, M.B. 2009, "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data", *Genome biology,* vol. 10, no. 2, pp. R23-2009-10-2-r23.
42. Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., NHLBI Exome Sequencing Project, Quinlan, A.R., Nickerson, D.A. & Eichler, E.E. 2012, "Copy number

variation detection and genotyping from exome sequence data", *Genome research,* vol. 22, no. 8, pp. 1525-1532.

43. LaFramboise, T. 2009, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances", *Nucleic acids research,* vol. 37, no. 13, pp. 4181-4193.
44. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. 2009, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome", *Genome biology,* vol. 10, no. 3, pp. R25-2009-10-3-r25. Epub 2009 Mar 4.
45. Li, H. & Durbin, R. 2009, "Fast and accurate short read alignment with Burrows-Wheeler transform", *Bioinformatics (Oxford, England),* vol. 25, no. 14, pp. 1754-1760.
46. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup 2009, "The Sequence Alignment/Map format and SAMtools", *Bioinformatics (Oxford, England),* vol. 25, no. 16, pp. 2078-2079.
47. Li, H., Ruan, J. & Durbin, R. 2008, "Mapping short DNA sequencing reads and calling variants using mapping quality scores", *Genome research,* vol. 18, no. 11, pp. 1851-1858.
48. Liu, B., Morrison, C.D., Johnson, C.S., Trump, D.L., Qin, M., Conroy, J.C., Wang, J. & Liu, S. 2013, "Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges", *Oncotarget,* vol. 4, no. 11, pp. 1868-1881.
49. Marenne, G., Rodriguez-Santiago, B., Closas, M.G., Perez-Jurado, L., Rothman, N., Rico, D., Pita, G., Pisano, D.G., Kogevinas, M., Silverman, D.T., Valencia, A., Real, F.X., Chanock, S.J., Genin, E. & Malats, N. 2011, "Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study", *Human mutation,* vol. 32, no. 2, pp. 240-248.
50. Martinez, J., Taylor Parker, M. & Fultz, K. 2003, "Molecular Biology of Cancer" in *Burger's Medicinal Chemistry and Drug Discovery; Volume 5: Chemotherapeutic Agents*, ed. D. Abraham, 6th ed, John Wiley & Sons, Inc., , pp. 1-50.
51. Marusyk, A. & Polyak, K. 2010, "Tumor heterogeneity: causes and consequences", *Biochimica et biophysica acta,* vol. 1805, no. 1, pp. 105-117.
52. Mayrhofer, M., DiLorenzo, S. & Isaksson, A. 2013, "Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue", *Genome biology,* vol. 14, no. 3, pp. R24-2013-14-3-r24.
53. Metzker, M.L. 2010, "Sequencing technologies - the next generation", *Nature reviews.Genetics,* vol. 11, no. 1, pp. 31-46.
54. Moore, C.M. & Best, R.G. 2001, "Chromosomal Genetic Disease: Structural Aberrations", *eLS.*
55. Moore, S.R., Persons, D.L., Sosman, J.A., Bobadilla, D., Bedell, V., Smith, D.D., Wolman, S.R., Tuthill, R.J., Moon, J., Sondak, V.K. & Slovak, M.L. 2008, "Detection of copy number alterations in metastatic melanoma by a DNA fluorescence in situ hybridization probe panel and array comparative genomic hybridization: a southwest oncology group study (S9431)", *Clinical cancer research : an official journal of the American Association for Cancer Research,* vol. 14, no. 10, pp. 2927-2935.

56. Mosén-Ansorena, D., Aransay, A. & Rodríguez-Ezpeleta, N. 2012, "Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data", *BMC Bioinformatics,* vol. 13, no. 1, pp. 192.

57. Nowell, P.C. 1976, "The clonal evolution of tumor cell populations", *Science (New York, N.Y.),* vol. 194, no. 4260, pp. 23-28.

58. Oesper, L., Mahmoody, A. & Raphael, B.J. 2013, "THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data", *Genome biology,* vol. 14, no. 7, pp. R80-2013-14-7-r80.

59. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. 2004, "Circular binary segmentation for the analysis of array-based DNA copy number data", *Biostatistics (Oxford, England),* vol. 5, no. 4, pp. 557-572.

60. Oostlander, A.E., Meijer, G.A. & Ylstra, B. 2004, "Microarray-based comparative genomic hybridization and its applications in human genetics", *Clinical genetics,* vol. 66, no. 6, pp. 488-495.

61. Piazza, R., Magistroni, V., Pirola, A., Redaelli, S., Spinelli, R., Redaelli, S., Galbiati, M., Valletta, S., Giudici, G., Cazzaniga, G. & Gambacorti-Passerini, C. 2013, "CEQer: a graphical tool for copy number and allelic imbalance detection from whole-exome sequencing data", *PloS one,* vol. 8, no. 10, pp. e74825.

62. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W. & Albertson, D.G. 1998, "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays", *Nature genetics,* vol. 20, no. 2, pp. 207-211.

63. Popova, T., Manie, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E. & Stern, M.H. 2009, "Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays", *Genome biology,* vol. 10, no. 11, pp. R128-2009-10-11-r128. Epub 2009 Nov 11.

64. Robinson, J.T., Thorvaldsdóttir, H. & Mesirov, J.P. 2011, *Integrative Genomics Viewer*. Available: http://www.broadinstitute.org/igv/ [2015, January/16].

65. Rowley, J.D. 2001, "Chromosome translocations: dangerous liaisons revisited", *Nature reviews. Cancer,* vol. 1, no. 3, pp. 245-250.

66. Sathirapongsasuti, J.F. 2015, "Pushing the boundaries of somatic copy-number variation detection: advances and challenges", *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO,* vol. 26, no. 1, pp. 11-12.

67. Sathirapongsasuti, J.F., Lee, H., Horst, B.A., Brunner, G., Cochran, A.J., Binder, S., Quackenbush, J. & Nelson, S.F. 2011, "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV", *Bioinformatics (Oxford, England),* vol. 27, no. 19, pp. 2648-2654.

68. Schaaf, C.P., Wiszniewska, J. & Beaudet, A.L. 2011, "Copy number and SNP arrays in clinical diagnostics", *Annual review of genomics and human genetics,* vol. 12, pp. 25-51.

69. Scheinin, I., Sie, D., Bengtsson, H., van de Wiel, M.A., Olshen, A.B., van Thuijl, H.F., van Essen, H.F., Eijk, P.P., Rustenburg, F., Meijer, G.A., Reijneveld, J.C., Wesseling, P., Pinkel, D., Albertson, D.G. & Ylstra, B. 2014, "DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly", *Genome research,* vol. 24, no. 12, pp. 2022-2032.

70. Seiser, E.L. & Innocenti, F. 2015, "Hidden Markov Model-Based CNV Detection Algorithms for Illumina Genotyping Microarrays", *Cancer informatics,* vol. 13, no. Suppl 7, pp. 77-83.

71. Speicher, M.R. & Carter, N.P. 2005, "The new cytogenetics: blurring the boundaries with molecular biology", *Nature reviews.Genetics,* vol. 6, no. 10, pp. 782-792.

72. Staaf, J., Lindgren, D., Vallon-Christersson, J., Isaksson, A., Goransson, H., Juliusson, G., Rosenquist, R., Hoglund, M., Borg, A. & Ringner, M. 2008, "Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays", *Genome biology,* vol. 9, no. 9, pp. R136-2008-9-9-r136. Epub 2008 Sep 16.

73. Tan, R., Wang, Y., Kleinstein, S.E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A.S. & Zhu, M. 2014, "An evaluation of copy number variation detection tools from whole-exome sequencing data", *Human mutation,* vol. 35, no. 7, pp. 899-907.

74. Tcga-data.nci.nih.gov 2015, *The Cancer Genome Atlas - Data Portal*. Available: https://tcga-data.nci.nih.gov/tcga/ [2015, January 23].

75. The Cancer Genome Atlas 2015, *Missions and Goals*. Available: http://cancergenome.nih.gov/abouttcga/overview/missiongoal [2015, January/16].

76. Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J.E., Shah, R.B., Pienta, K.J., Rubin, M.A. & Chinnaiyan, A.M. 2005, "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer", *Science (New York, N.Y.),* vol. 310, no. 5748, pp. 644-648.

77. Trapnell, C. & Salzberg, S.L. 2009, "How to map billions of short reads onto genomes", *Nature biotechnology,* vol. 27, no. 5, pp. 455-457.

78. Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., Perou, C.M., Borresen-Dale, A.L. & Kristensen, V.N. 2010, "Allele-specific copy number analysis of tumors", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, no. 39, pp. 16910-16915.

79. Wang, H. & Elbein, S.C. 2007, "Detection of allelic imbalance in gene expression using pyrosequencing", *Methods in molecular biology (Clifton, N.J.),* vol. 373, pp. 157-176.

80. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H. & Bucan, M. 2007, "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data", *Genome research,* vol. 17, no. 11, pp. 1665-1674.

81. Weinberg, R.A. 2013, *The Biology of Cancer,* 2nd ed, Garland Science, New York.

82. Yakhini, Z. & Jurisica, I. 2011, "Cancer computational biology", *BMC bioinformatics,* vol. 12, pp. 120-2105-12-120.

83. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. 2009, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads", *Bioinformatics (Oxford, England),* vol. 25, no. 21, pp. 2865-2871.

84. Yu, Z., Liu, Y., Shen, Y., Wang, M. & Li, A. 2014, "CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data", *Bioinformatics (Oxford, England),* pp. 1-8.

85. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. 2013, "Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives", *BMC bioinformatics,* vol. 14 Suppl 11, pp. S1-2105-14-S11-S1. Epub 2013 Sep 13.

Appendices

In this section the appendices for this study is presented. Appendix A contains the complete name of the TCGA PRAD samples used for the benchmark. Appendix B presents the parameter values and the configuration files used for the running of the competing tools. Finally, appendix C represents the sets of commands in the Pypette package used to extract the read-depth and BAF data from BAM files required by the developed tool.

Appendix A. Sample Dataset Names

This appendix provides the complete name of the downloaded TCGA PRAD datasets from CGHub.

TCGA-G9-6338-01A-12D-1957_130222_SN1222_0174_BC1RNEACXX_s_5_rg.sorted.bam ➔ G9_6338_t

TCGA-G9-6338-10A-01D-1957_130222_SN1222_0174_BC1RNEACXX_s_6_rg.sorted.bam ➔G9_6338_n

TCGA-HC-7211-01A-11D-2111_130517_SN590_0230_AD24B4ACXX_s_2_rg.sorted.bam ➔ HC_7211_t

TCGA-HC-7211-11A-01D-2111_130517_SN590_0230_AD24B4ACXX_s_3_rg.sorted.bam ➔ HC_7211_n

TCGA-G9-6342-01A-11D-1957_130222_SN1222_0174_BC1RNEACXX_s_7_rg.sorted.bam ➔ G9_6342_t

TCGA-G9-6342-10A-01D-1957_130222_SN1222_0174_BC1RNEACXX_s_8_rg.sorted.bam ➔ G9_6342_n

TCGA-HI-7171-01A-12D-2111_130607_SN590_0232_BD26R5ACXX_s_7_rg.sorted.bam ➔ HI_7171_t

TCGA-HI-7171-10A-01D-2111_130607_SN590_0232_BD26R5ACXX_s_8_rg.sorted.bam ➔ HI_7171_n

TCGA-HC-7212-01A-11D-2111_130517_SN590_0230_AD24B4ACXX_s_4_rg.sorted.bam ➔ HC_7212_t

TCGA-HC-7212-10A-01D-2111_130517_SN590_0230_AD24B4ACXX_s_5_rg.sorted.bam ➔ HC_7212_n

TCGA-CH-5761-01A-11D-1572_130103_SN1222_0166_BD1K7EACXX_s_7_rg.sorted.bam ➔ CH_5761_t

TCGA-CH-5761-11A-01D-1572_130103_SN1222_0166_BD1K7EACXX_s_8_rg.sorted.bam ➔ CH_5761_n

Appendix B. Parameters and Configuration Files

Under this appendix the parameter settings and content of the configuration files used to run the competing tools is presented.

ControlFREEC

ControlFREEC requires a SNP file and a configuration file in order to run. The SNP file was downloaded from https://xfer.curie.fr/get/QKFgcU5caZd/hg19_snp138.SingleDiNucl.1based.txt.gz

Here, the content of the configuration file for each of the samples is presented.

G9_6338

```
[general]
chrFiles = /data/csb/datasets/tcga_prad/wgs/hg19_broad_variant
chrLenFile = /home/afyounia/controlFREEC_run/hg19.chrom.sizes
coefficientOfVariation =  0.05
#window = 5000
contaminationAdjustment = TRUE
intercept = 1
outputDir = /home/afyounia/controlFREEC_run/results/g9_6338_cv
ploidy = 2
samtools = samtools
sex = XY
numberOfProcesses = 5

[sample]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/G9_6338_t.pileup.gz
inputFormat=pileup
mateOrientation=FR

[control]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/G9_6338_n.pileup.gz
inputFormat=pileup
mateOrientation=FR

[BAF]
SNPfile = /data/csb/datasets/tcga_prad/wgs/snp_file/hg19_snp138_1based_nochr_FREEC.txt
```

HC_7211

```
[general]
chrFiles = /data/csb/datasets/tcga_prad/wgs/hg19_broad_variant
chrLenFile = /home/afyounia/controlFREEC_run/hg19.chrom.sizes
coefficientOfVariation =  0.05
#window = 5000
contaminationAdjustment = TRUE
intercept = 1
outputDir = /home/afyounia/controlFREEC_run/results/hc_7211_cv
ploidy = 2
samtools = samtools
sex = XY
numberOfProcesses = 5

[sample]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/HC_7211_t.pileup.gz
inputFormat=pileup
mateOrientation=FR

[control]
mateFile = /data/csb/datasets/tcga_prad/wgs/pileups_freec/HC_7211_n.pileup.gz
inputFormat=pileup
mateOrientation=FR
```

```
[BAF]
SNPfile = /data/csb/datasets/tcga_prad/wgs/snp_file/hg19_snp138_1based_nochr_FREEC.txt
```

G9-6342

```
[general]
chrFiles = /data/csb/datasets/tcga_prad/wgs/hg19_broad_variant
chrLenFile = /home/afyounia/controlFREEC_run/hg19.chrom.sizes
coefficientOfVariation =  0.05
#window = 5000
contaminationAdjustment = TRUE
intercept = 1
outputDir = /home/afyounia/controlFREEC_run/results/g9_6342_cv
ploidy = 2
samtools = samtools
sex = XY
numberOfProcesses = 5

[sample]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/G9_6342_t.pileup.gz
inputFormat=pileup
mateOrientation=FR

[control]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/G9_6342_n.pileup.gz
inputFormat=pileup
mateOrientation=FR

[BAF]
SNPfile = /data/csb/datasets/tcga_prad/wgs/snp_file/hg19_snp138_1based_nochr_FREEC.txt
```

HI-7171

```
[general]
chrFiles = /data/csb/datasets/tcga_prad/wgs/hg19_broad_variant
chrLenFile = /home/afyounia/controlFREEC_run/hg19.chrom.sizes
coefficientOfVariation =  0.05
#window = 5000
contaminationAdjustment = TRUE
intercept = 1
outputDir = /home/afyounia/controlFREEC_run/results/hi_7171_cv
ploidy = 2
samtools = samtools
sex = XY
numberOfProcesses = 5

[sample]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/HI_7171_t.pileup.gz
inputFormat=pileup
mateOrientation=FR

[control]
mateFile = /data/csb/datasets/tcga_prad/wgs/pileups_freec/HI_7171_n.pileup.gz
inputFormat=pileup
mateOrientation=FR

[BAF]
SNPfile = /data/csb/datasets/tcga_prad/wgs/snp_file/hg19_snp138_1based_nochr_FREEC.tx
```

HC-7212

```
[general]
chrFiles = /data/csb/datasets/tcga_prad/wgs/hg19_broad_variant
chrLenFile = /home/afyounia/controlFREEC_run/hg19.chrom.sizes
coefficientOfVariation =  0.05
#window = 5000
contaminationAdjustment = TRUE
```

```
intercept = 1
outputDir = /home/afyounia/controlFREEC_run/results/hc_7212_cv_2
ploidy = 2
samtools = samtools
sex = XY
numberOfProcesses = 5

[sample]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/HC_7212_t.pileup.gz
inputFormat=pileup
mateOrientation=FR

[control]
mateFile = /data/csb/datasets/tcga_prad/wgs/pileups_freec/HC_7212_n.pileup.gz
inputFormat=pileup
mateOrientation=FR

[BAF]
SNPfile = /data/csb/datasets/tcga_prad/wgs/snp_file/hg19_snp138_1based_nochr_FREEC.txt
```

CH-5761

```
[general]
chrFiles = /data/csb/datasets/tcga_prad/wgs/hg19_broad_variant
chrLenFile = /home/afyounia/controlFREEC_run/hg19.chrom.sizes
coefficientOfVariation =  0.05
#window = 5000
contaminationAdjustment = TRUE
intercept = 1
outputDir = /home/afyounia/controlFREEC_run/results/ch_5761_cv
ploidy = 2
samtools = samtools
sex = XY
numberOfProcesses = 5

[sample]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/CH_5761_t.pileup.gz
inputFormat=pileup
mateOrientation=FR

[control]
mateFile =  /data/csb/datasets/tcga_prad/wgs/pileups_freec/CH_5761_t.pileup.gz
inputFormat=pileup
mateOrientation=FR

[BAF]
SNPfile = /data/csb/datasets/tcga_prad/wgs/snp_file/hg19_snp138_1based_nochr_FREEC.txt
```

Patchwork

Patchwork, in order to run, requires a reference genome. The documentation recommends creating a reference genome using $patchwork.createreference$() function with three normal BAM files. The following normal sample BAM files were used in order to create the reference genome: *G9_6338_n*, *HI_7171_n* and *CH_5761_n* (refer to appendix A for the complete sample names).

Furthermore, Patchwork uses $patchwork.copynumbers$() function in order to assign copy number and allele ratio for each of the segments. This function receives some parameters. These parameters are ***cn2***, ***delta***, ***het***, and ***hom***. In order to set the parameters, the user is required to interpret one of the chromosomal plots generated by the $patchwork.plot$() function. **cn2** denotes the position of copy number 2 in the plot. **delta** denotes the difference between two copy numbers on the coverage axis. **het** denotes the position of heterozygous copy number 2 on the allelic imbalance axis. Finally, **hom** denotes the position of homozygous, LOH, copy number 2 on the allelic imbalance axis. The Patchwork documentation (available at: http://patchwork.r-forge.r-project.org) provides an example illustrating how these parameter values are extracted from the plot. Figure B.1 represents this example plot. From this example figure, the following parameter values can be extracted from the plot: $\boldsymbol{cn2} = 0.8$, $\boldsymbol{delta} = 0.28$, $\boldsymbol{het} = 0.21$ and $\boldsymbol{hom} = 0.79$. In this study, the same procedure was used in order to extract the parameter values for the six TCGA PRAD samples.

CLImAT

There are two steps in performing copy number analysis using CLImAT. First, the read-depth data need to be extracted from the BAM files using *DFExtract* tool. After having extracted the read-depth data, in order to perform the copy number analysis CLImAT is used.

DFExtract requires a SNP file and a mapability file in order to extract the read-depth data. The SNP file was downloaded from

http://bioinformatics.ustc.edu.cn/CLImAT/download.html.

The recommended mapability (wgEncodeCrgMapabilityAlign36mer.bw) file was downloaded from

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/.

Furthermore, some parameter values need to be passed to the *DFExtract*. Table B.1 lists the default parameter values for running *DFExtract*. In addition, CLImAT requires a configuration file in order to run. Table B.2 lists the default parameter values used to create the configuration file:

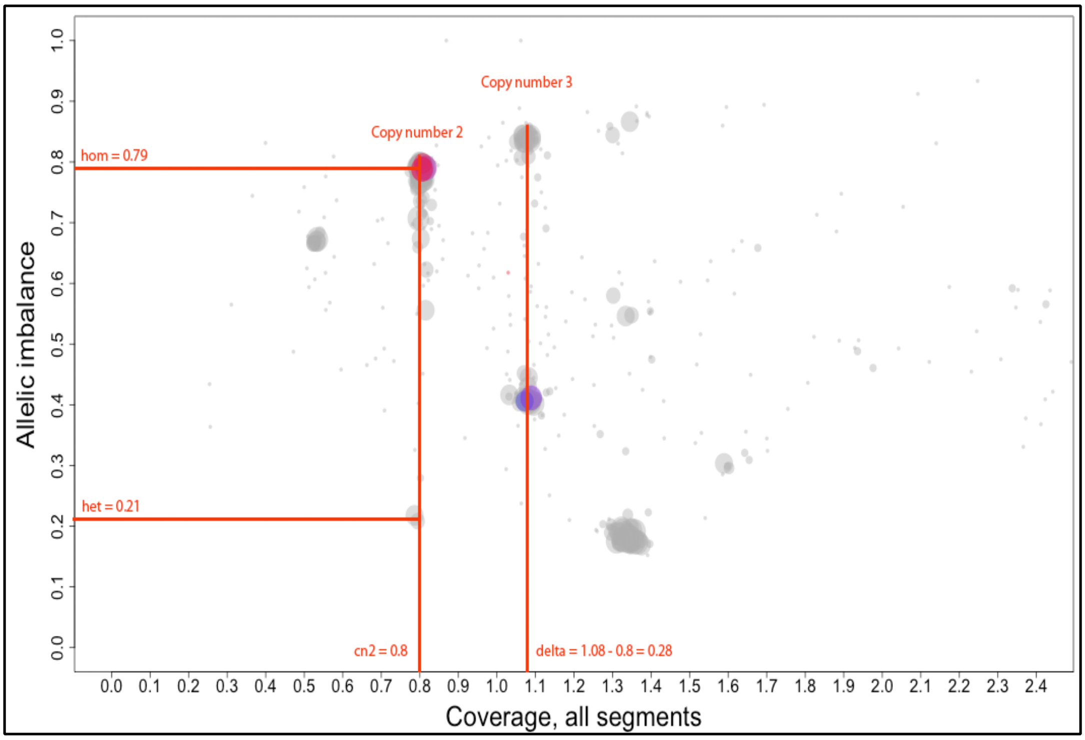

Figure B.1 *Extraction of parameter values from the plot created by Patchwork.*

Table B.1 Default parameter values used for running DFExtract.

Options	Description	Default value
-w, --window	set the size of windows	1000
-Q, --baseQ	threshold value for base quality	10
-q, --mapQ	threshold value for mapping quality	20
-d, --minDepth	minimum read-depth for a position to be considered	10

Table B.2 Default parameter values used for running CLImAT.

Parameters	Description	Default value
minDepth	minimum read-depth for a position to be considered	10
maxDepth	maximum read-depth for a position to be considered	300
minGC	minimum GC-content for a position to be considered	0
minMapScore	minimum mapability score for a position to be considered	0
maxMapScore	maximum mapability score for a position to be considered	0.98

Appendix C. Extraction of Read-depth and BAF Data from BAM Files using Pypette Package

In order to extract the read-depth data from BAM files following command is used:

```
coverage tiled <bam_file> <window_size> [-s N] [-q N] [-S|-1|-2] [-P|-M]

Options:
  -q --quality=N    Minimum alignment quality [default: 10].
  -s --step=N       Step size for window placement [default: window size / 2].
  -S --single       Use all reads for coverage calculation, not just paired.
  -P --plus         Calculate coverage only for the plus strand.
  -M --minus        Calculate coverage only for the minus strand.
```

In order to extract the BAF values from BAM files, the following commands are used sequentially:

```
variant call <genome_fasta> <bam_files>... [-r REGION] [--ref=N:R] [--hetz=N:R]
[--homz=N:R] [-q N] [-Q SAMPLES] [--keep-all]

variant keep samples <vcf_file> <regex>

variant heterozygous bases <vcf_file> <pos_file>

variant discard samples <vcf_file> <regex>

variant allele fractions <vcf_file> <pos_file>

Options:
  -r <region>       Restrict analysis to chromosomal region
  -q N              Minimum mapping quality score [default: 10]
  -Q SAMPLES        Samples for which mapping quality is ignored [default: ]
  --ref=N:R         Minimum evidence for homozygous reference [default: 8:0.9]
  --hetz=N:R        Minimum evidence for heterozygous [default: 4:0.25]
  --homz=N:R        Minimum evidence for homozygous alt [default: 4:0.8]
  --keep-all        Show sites even if they are all homozygous reference
```