Inaugural dissertation
for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics,
Engineering and Natural Sciences
of the
Ruprecht - Karls - University
Heidelberg

Presented by

M.Sc. Zhao Yuan

Born in: Tianjin, China

Oral examination: 22.11.2024

Analysis of somatic copy number alterations in liquid biopsies from cancer patients

Referees:

Prof. Dr. Benedikt Brors

Prof. Dr. Holger Sülltmann

Abstract

Compared to tissue biopsies, ctDNA provides a more comprehensive tumor landscape, allowing for repeated non-invasive sampling. Therefore, ctDNA detection has a broad application prospect in tumor diagnosis, treatment and monitoring. However, due to the low content of ctDNA in cfDNA, ctDNA detection requires a high sequencing depth to achieve high sensitivity. Currently, the commonly used ctDNA testing scheme is panel sequencing with high coverage, so that genes of interest can be tested at a lower cost. However, panel sequencing has limited ability to detect other variants such as SV and CNV. Another option is lcWGS. lcWGS is able to identify CNVs, providing valuable insights into genomic alterations. Based on the data of HIPO-K34 and INFORM, this study explored the ability of ctDNA detection by the two schemes. The HIPO-K34 project focused on patients with ALK gene-fused non-small cell lung cancer (NSCLC) and contains multi-time point sequencing data from lcWGS and panel sequencing. The INFORM project consists of liquid biopsy samples and tissue samples taken from the same patient at the same time point. In addition, various detection tools were benchmarked using simulated data with known tumor DNA fractions and CNVs. In order to improve the detection performance, the tools were optimized and the tool with the best performance was selected. Finally, a pipeline combining the panel analysis process with the optimized lcWGS analysis process was established for the accurate analysis of liquid biopsy samples.

Keywords: liquid biopsy; ctDNA; lcWGS; tumor DNA fraction; CNV

Zusammenfassung

Im Vergleich zu Gewebebiopsien bietet ctDNA eine umfassendere Tumorlandschaft und ermöglicht wiederholte, nicht-invasive Probenahmen. Daher hat die ctDNA-Detektion ein breites Anwendungspotential in der Tumordiagnose, -behandlung und -überwachung. Aufgrund des niedrigen ctDNA-Gehalts in cfDNA erfordert die ctDNA-Detektion jedoch eine hohe Sequenziertiefe, um eine hohe Sensitivität zu erreichen. Derzeit ist das häufig verwendete ctDNA-Testverfahren die Panel-Sequenzierung mit hoher Abdeckung, sodass Gene von Interesse zu geringeren Kosten getestet werden können. Die Panel-Sequenzierung hat jedoch eine begrenzte Fähigkeit, andere Varianten wie SV und CNV zu erkennen. Eine andere Option ist lcWGS. lcWGS kann CNVs identifizieren und wertvolle Einblicke in genomische Veränderungen bieten. Basierend auf den Daten von HIPO-K34 und INFORM wurde in dieser Studie die Fähigkeit der ctDNA-Detektion durch die beiden Verfahren untersucht. Das HIPO-K34-Projekt konzentrierte sich auf Patienten mit ALK-genfusioniertem nicht-kleinzelligem Lungenkrebs (NSCLC) und enthält Sequenzierungsdaten zu mehreren Zeitpunkten von lcWGS und Panel-Sequenzierung. Das INFORM-Projekt besteht aus Flüssigbiopsieproben und Gewebeproben, die vom selben Patienten zum selben Zeitpunkt entnommen wurden. Darüber hinaus wurden verschiedene Detektionstools unter Verwendung von simulierten Daten mit bekannten Tumor-DNA-Fraktionen und CNVs bewertet. Um die Detektionsleistung zu verbessern, wurden die Tools optimiert und das leistungsstärkste Tool ausgewählt. Schließlich wurde eine Pipeline entwickelt, die den Panel-Analyseprozess mit dem optimierten lcWGS-Analyseprozess kombiniert, um eine genaue Analyse von Flüssigbiopsieproben zu ermöglichen.

Schlüsselwörter: Flüssigbiopsie; ctDNA; lcWGS; Tumor-DNA-Fraktion; CNV

Content

Figure of Content

Table of Content

1. Introduction

The text was written by Zhao Yuan. It has been proofread and edited by ChatGPT.

1.1 Cancer pathogenesis

Cancer is caused by the accumulation of genomic changes in somatic cells. There are many reasons for these mutations, including mismatches during DNA replication, DNA repair defects, and exposure to exogenous or endogenous mutagens[1]. In 1982, the first oncogene RAS was discovered in bladder cancer cells. Mutations in the RAS gene can inhibit the senescence and death of cells, leading to cell canceration[2]. In addition, studies have shown that RAS gene mutations can maintain the stable synthesis of PD-L1 protein which can respond to PD-1 on the surface of T cells so that cancer cells have the ability to promote immunosuppression[3]. Subsequently, the first tumor suppressor gene RB1 was discovered, which plays an indispensable role in inhibiting the occurrence of a variety of tumors, such as retinoblastoma, small cell lung cancer, osteosarcoma, pancreatic cancer, and breast cancer[4]. The tumor suppressor effect of RB1 is closely related to its regulation of cell cycle, cell differentiation, cell senescence, cell apoptosis, and growth inhibition[5]. The activation of oncogenes and the loss of tumor suppressor genes can all lead to the occurrence of cancer. Microbial gene integration is also one of the important causes of cancer. Zapatka M et al. analyzed the whole genome and part of the transcriptome data of 38 tumor types from 2,658 cases and detected virus genes in about 13% of the samples. These include Epstein-Barr Virus (EBV), Hepatitis B Virus (HBV), and Human Papilloma Virus (HPV)[6]. In addition, epigenetic changes may also cause cancer by changing chromosome structure and gene expression[7].

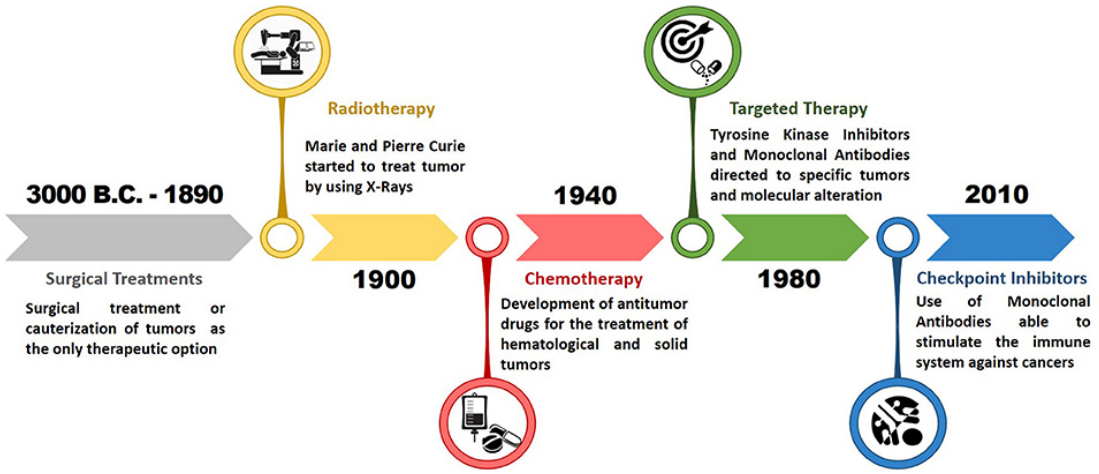


Figure 1-1 Timeline of tumor therapeutic option. In the early 1900, radiotherapy is the main means of cancer treatment. After that, with the discovery of the first chemotherapeutic drugs, chemotherapy gradually became one of the means of cancer treatment. In 1980, Targeted therapy went into people's vision because of the progress of Medical Oncology. In the past few years, the emergence of immune checkpoint inhibitors provided new ideas for advanced and metastatic tumors (from Falzone L, et al. 2018)[9].

Over the past century, there have been tremendous advancements in cancer treatment research. (Figure 1-1) In the past few decades, researchers have conducted extensive studies on the mechanism of cancer and made considerable progress. However, mutations in cancer genomes vary widely between different tumor types and different cases. For example, some cancer genomes contain more than 100,000 point mutations, while others have less than 1,000.[9] Some childhood cancers carry very few mutations in the genome, while cancers that have been exposed to mutagens for a long time, such as lung cancer caused by smoking, contain numerous mutations.[10, 11] In addition to point mutations, the rearrangement of the cancer genome is also very complicated[12]. These characteristics of cancer have brought great challenges to cancer research. Next-generation sequencing (NGS) methods can help researchers understand these changes in the cancer genome more comprehensively.

1.2 Cancer diagnosis: tissue biopsy and liquid biopsy

Tumor tissue is the current gold standard source for diagnosing and characterizing cancer. Morphological and immunohistochemical analysis of tissue samples can provide critical information to determine the type and characteristics of cancer and provide information on its grade and extent of spread[13]. By detecting mutations in specific genes or abnormal expression of proteins in tumor tissue, it is possible to predict a patient's sensitivity to certain targeted therapeutic drugs. This provides an important basis for the development of individualized treatment plans, which are critical to providing optimal patient care[14]. Clinically, tissue cells are usually chemically or frozen processed, and then further studied by microscopic observation and sequencing. Because of directly sampling the tumor tissue, tissue biopsy can obtain a higher concentration of tumor cells, which is very helpful for obtaining sufficient tumor biological information. In addition, when paired normal samples exist, comparing the patient's tumor cell genome with the normal cell genome through high-throughput sequencing can easily eliminate individual patient bias[15].

However, the shortcomings of tissue biopsy cannot be ignored. First, sampling is difficult. For some patients, such as advanced cancer patients and lung cancer patients with pneumothorax[16], the sampling of tissue biopsy is relatively difficult. Second, clinical complications. The tissue biopsy process can cause trauma that is difficult to heal. Especially for those patients which need repeated sampling and have a poor biological function, it is easy to cause complications. The difficulties brought by these complications in the treatment are undoubtedly worsening the situation for the patients[17]. Third, it is difficult to preserve samples. One method of tissue preservation in clinical is fresh frozen. However, due to the relative higher cost of fresh frozen fixation and the need for the resected tissue to be quickly frozen in liquid nitrogen in a short period of time, formalin fixation and paraffin embedding (FFPE) is an alternative to fresh frozen[18]. But FFPE can easily cause DNA cross-linking in the sample[19]. Fourth, is the heterogeneity of tumors. Tissue biopsy is conducted for a specific part,

and the sample can only represent the tumor information at the specific part at the time of sampling. The change of tumor is an evolving process, and there is heterogeneity among different tumor cells[20]. Tissue biopsy cannot reflect the heterogeneity of tumor.

Compared with tissue biopsy, liquid biopsy has the advantage of non-invasiveness, reflecting the whole picture of the tumor, and facilitating real-time monitoring of patients. There is cell-free circulating DNA (cfDNA) in human plasma and serum. These cfDNA fragments usually come from apoptotic cells in healthy humans. (Figure 1-2) DNA fragments are released into the blood after cell apoptosis, the size of these DNA fragments is usually between 150 and 200bp[21]. In cancer patients, besides the DNA fragments produced by normal apoptotic cells, there are also DNA fragments from necrotic tumor cells and apoptotic tumor cells, or DNA fragments carried in exosomes released by tumor cells[22]. These DNA fragments are called cell-free circulating tumor DNA (ctDNA). Therefore, ctDNA can be used as a marker for tumor detection. Studies have shown that there are higher cfDNA levels in cancer patients than in healthy people[23, 24, 25]. This is because phagocytes cannot effectively remove the residues of apoptosis and necrosis in tumors, causing DNA fragments to aggregate and release into the blood.

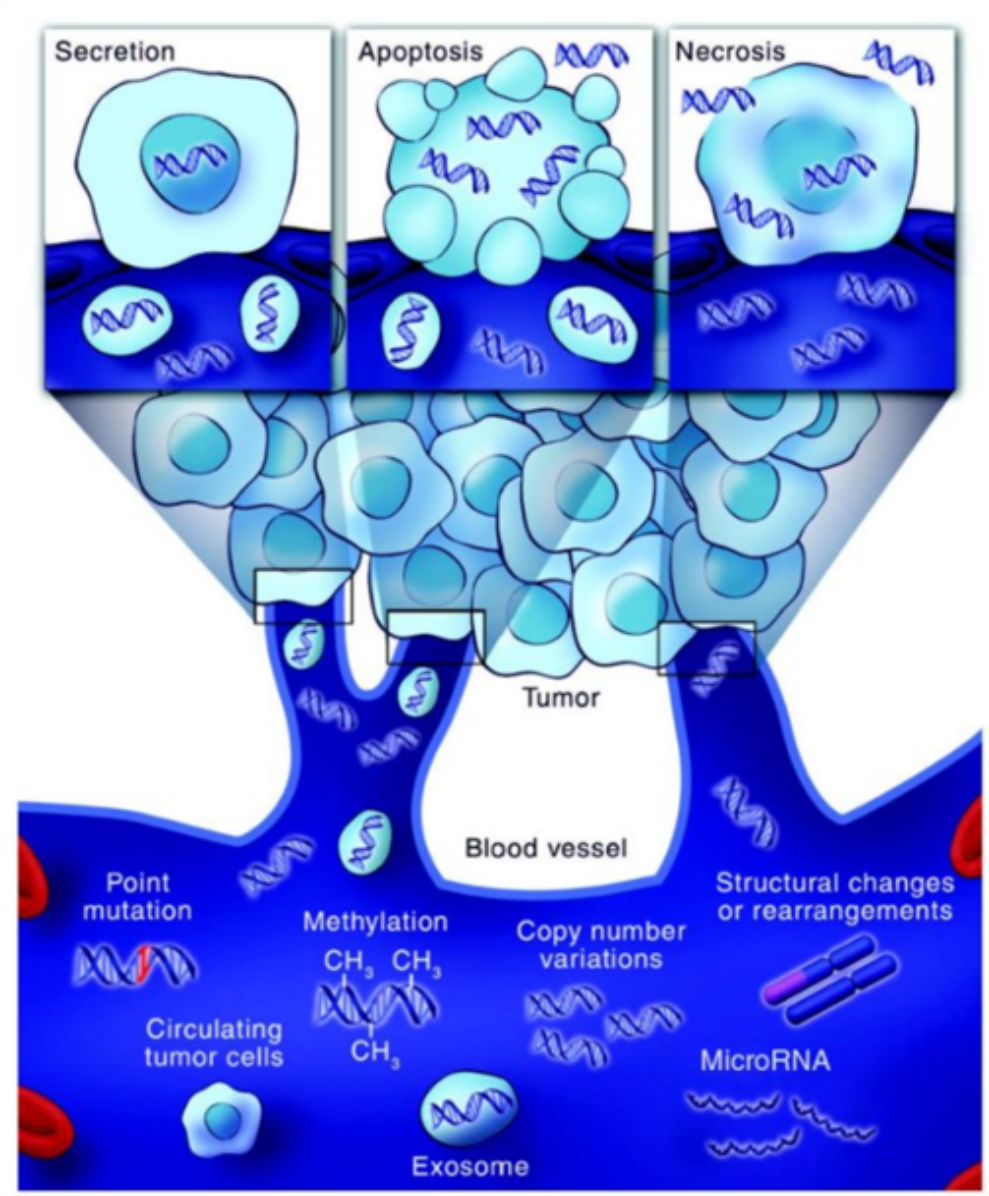


Figure 1-2 The release of DNA fragments from tumor cells into the blood circulation. *The DNA fragments of tumor cells are released through secretion, apoptosis, and necrosis, accumulate in the tissues, and finally enter the circulatory system. By sequencing cfDNA in the blood, point mutations in cfDNA, CNV, chromosomal rearrangements, changes in methylation levels, etc. can be detected (from Diaz Jr, et al. 2014)*[19].

Liquid biopsy also has its own limitations. At present, the accuracy of the ctDNA detection is insufficient, and the false negative of the detection cannot infer the absence of a tumor. For different types of cancer, the detection rate of liquid biopsy varies greatly[26]. Compared with other cancers, ctDNA has a lower detection rate in primary brain, kidney, prostate, or thyroid

cancers. This is because physical barriers like the blood-brain barrier and mucins can prevent ctDNA from entering the circulation[27].

1.3 Next Generation Sequencing and Cancer Diagnosis

In the past few years, next-generation sequencing technology has made rapid progress and development. Compared with the traditional first-generation sequencing technology, NGS has the advantages of high throughput and low cost. In 2001, the cost of sequencing the entire human genome was about $100,000,000. With the development of next-generation sequencing technology, the cost has dropped to less than $1,000 in 2021[28]. The reduction of sequencing cost makes it possible to use sequencing technology to assist in the diagnosis and treatment of cancer. Using next-generation sequencing technology to detect the mutation in patients to support the design and adjustment of targeted drugs or immunotherapy, is a common auxiliary means in cancer therapy[29].

The application of NGS is of great help to the diagnosis and treatment of cancer[30]. As mentioned earlier, cancer is caused by the accumulation of mutations in the genome of somatic cells. Even for the same type of cancer, the genetic mutations between different individuals are very different. Next-generation sequencing strategies include whole-genome sequencing (WGS), whole-exome sequencing (WES), transcriptome sequencing (RNA-seq), targeted sequencing, Bisulfite-seq, ChIP-seq, etc. Figure 1-3 shows the application of NGS in cancer research and clinical application.

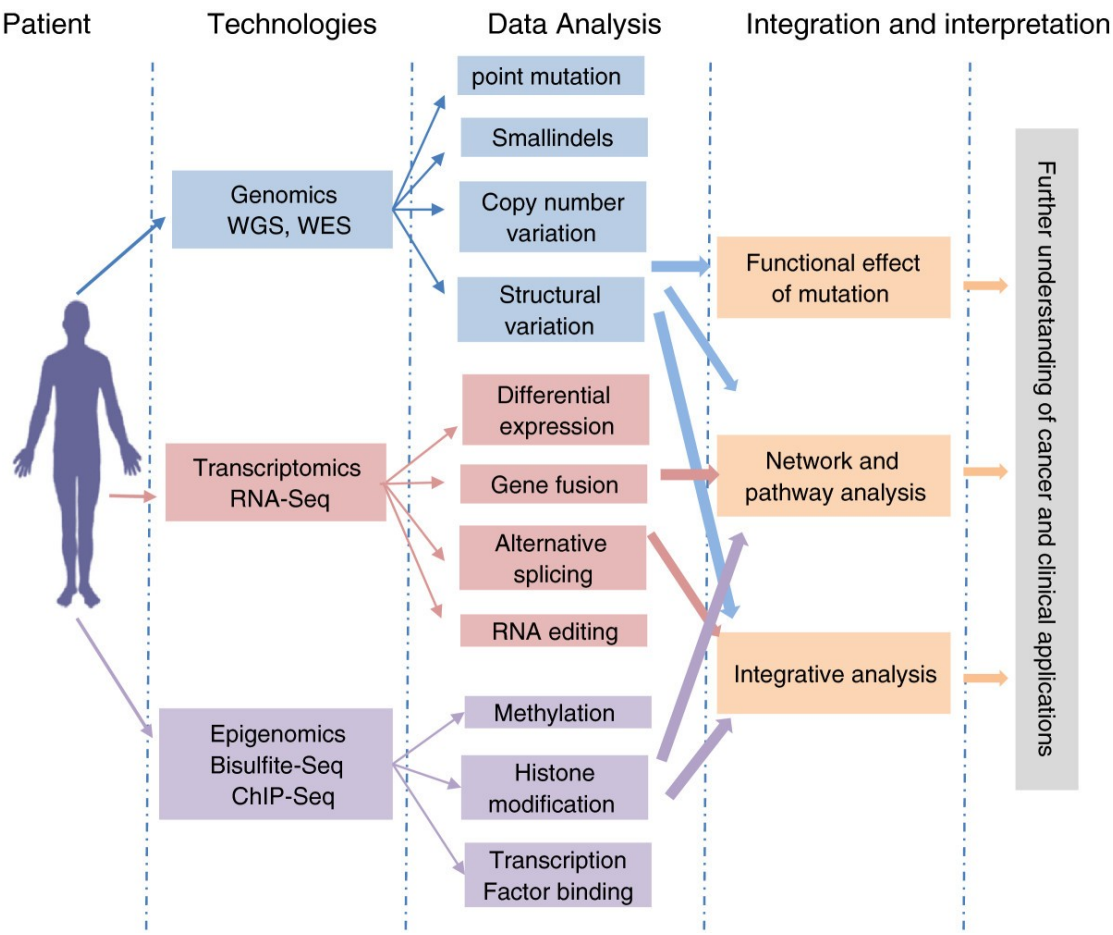


Figure 1-3 The application of second-generation sequencing in cancer research and clinical application. The application of second-generation sequencing includes WGS and WES at the genomic level, RNA-seq at the transcriptome level, and bisulfite-seq and ChIP-seq at the epigenetic level. A variety of bioinformatics tools are used to analyze data to help us better understand the mechanism of cancer occurrence and formulate cancer treatment plans (from Shyr D, et al. 2013)[31].

1.3.1 Development history of sequencing

Sequencing technologies for proteins and RNAs have been around long before DNA sequencing technologies emerged. In 1949, Frederick Sanger developed a technique for determining the amino-terminal sequences of the two peptide chains of insulin, and in 1953, the amino acid sequence of insulin was determined[32]. Edman also proposed the protein N-terminal sequencing technology in 1950 and later developed the protein automatic

sequencing technology on this basis[33]. Sanger et al. invented the small fragment sequencing method of RNA in 1965 and completed the determination of 120 nucleotides of E. coli 5S rRNA[34]. During the same period, Holley completed the sequencing of yeast alanine-transporting tRNA[35].

Compared to RNA sequencing, DNA sequencing technology appeared relatively late. In 1975 Sanger and Coulson invented the addition and subtraction method to determine DNA sequence[36]. In 1977, after the introduction of dideoxynucleoside triphosphate (ddNTP), the dideoxy chain termination method was formed, which greatly improved the efficiency and accuracy of DNA sequence determination[37]. Maxam and Gilbert also reported in 1977 the chemical degradation method to determine the sequence of DNA[38]. In the same year, Frederick Sanger invented the first sequencer and used it to sequence the first genome, phage X174 with a full length of 5375 bases[39]. The dideoxy chain termination method, also known as the first-generation sequencing technology, remains widely used today. This method enables the sequencing of a range of 700-1000 bases in a single run, demonstrating high accuracy and effective handling of repetitive sequences[40]. However, its limitation of detecting only one template at a time makes it a time-consuming process. Consequently, it is unable to meet the urgent needs of modern scientific development for the acquisition of modern scientific development for the acquisition of biological gene sequences.

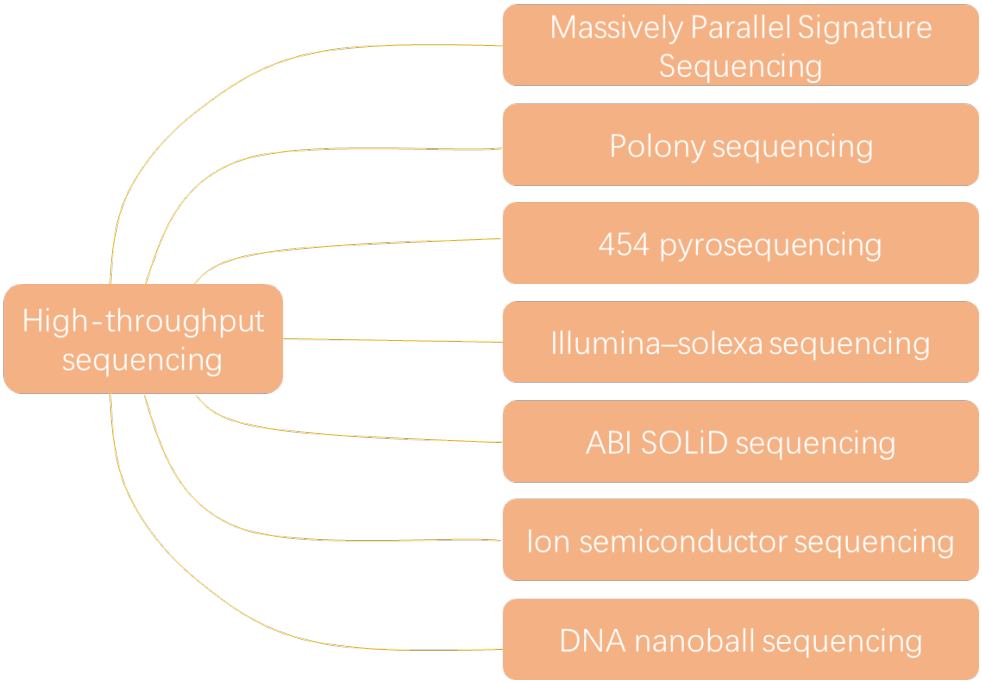


Figure 1-4 The next-generation sequencing technologies. *According to the development history, sequencing principles, and technologies, the next-generation sequencing can be mainly divided into massively parallel signature sequencing, polony sequencing, 454 pyrosequencing, Illumina–solexa sequencing, ABI SOLiD sequencing, Ion semiconductor sequencing, and DNA nanoball sequencing.*

Next-generation sequencing, also called high-throughput sequencing (HTS) is a revolutionary change to traditional Sanger sequencing technology, which can sequence up to millions of nucleic acid molecules at a time. The emergence of high-throughput sequencing technology has made it possible to conduct a detailed and comprehensive analysis of the genome and transcriptome of a species. There are currently several representative next-generation sequencing technologies (Figure 1-4), like 454 technology of Roche, SOLiD technology of ABI, and Solexa technology of Illumina[41]. Invented by Jonathan Rothberg in 2005, 454 was the first next-generation sequencing technology to be invented, which led life science research into the era of high-throughput sequencing[42]. The DNA fragment does not need to be fluorescently labeled or electrophoresed. Synthesis and sequencing are performed simultaneously. A pyrophosphate will be removed when the base is added to the sequence, and the base will be recognized by detecting the pyrophosphate. This technology is also called pyrosequencing. SOLiD technology was developed from the

ligase sequencing method. Leroy Hood designed the first automatic fluorescent sequencer using the ligase method in the middle 1980s[43]. Based on the sequential ligation synthesis of four-color fluorescently labeled oligonucleotides, SOLiD replaces the traditional polymerase ligation reaction and enables large-scale amplification and high-throughput parallel sequencing of DNA fragments. In the Solexa technology, synthesis and sequencing are also carried out at the same time[44] (Figure 1-5). In the process of sequencing, the modified DNA polymerase and dNTPs with four kinds of fluorescent labels are added. Because the 3' hydroxyl terminus of dNTPs bears a chemically cleavable moiety, it only allows the incorporation of a single base per cycle. The surface of the reaction plate is scanned with the laser so that the type of nucleotides polymerized in each round of the reaction of each template sequence can be determined according to the fluorescence of dNTPs. After the cycle of "synthesis-cleaning-photographing", the sequence of the target fragment is finally obtained.

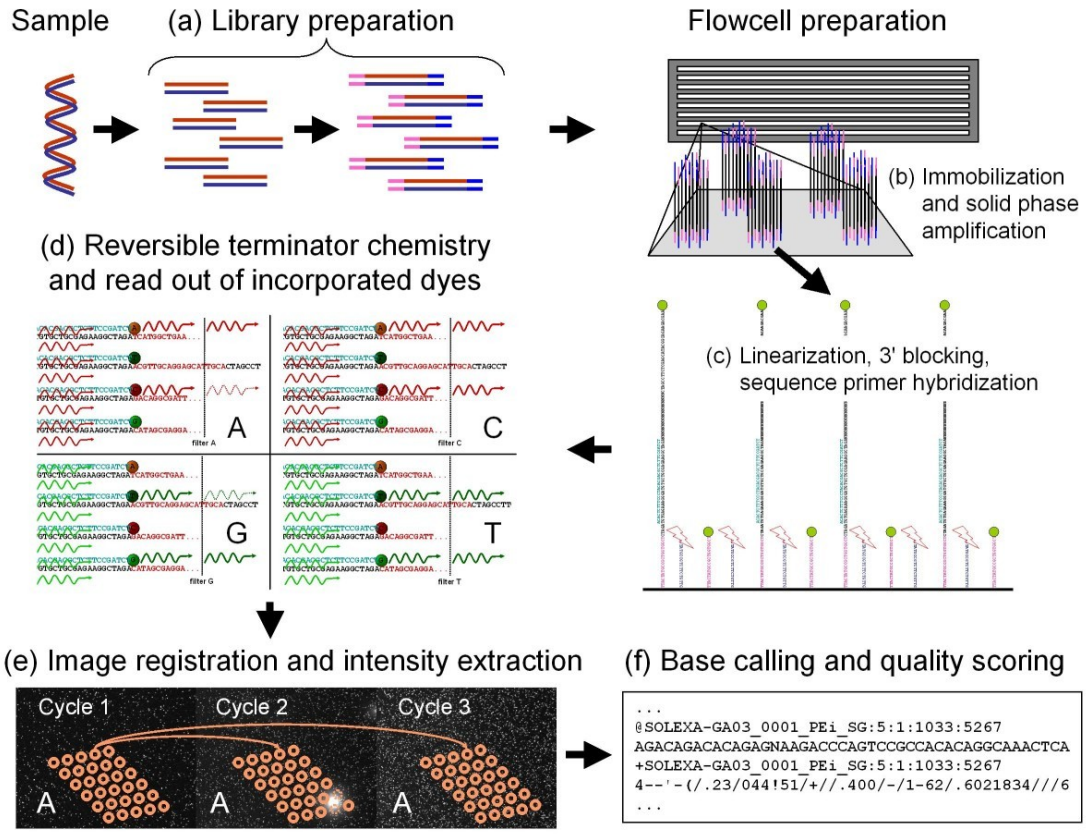


Figure 1-5 Principle and workflow of Illumina sequencing. *In the first step of sequencing, the*

sample DNA needs to be sheared into a specific size, and then the adapters are added to the end of the DNA fragments to prepare the sequencing library. Then the prepared library is loaded into the flowcell and amplified by bridge PCR. The dNTP with fluorescent-label is added to the system for sequencing. The dNTP also contains an azide group so it cannot extend normally during sequencing. Therefore, the extension of the sequence will stop after each nucleic acid is added. At this time, the nucleic acid being synthesized can be read under the observation equipment according to the fluorescence color emitted by the nanowells. After observation, the azide group and fluorescent group are hydrolyzed by specific enzymes, so that the next dNTP can enter the extension sequence normally (from Kircher M, et al. 2011)[45].

The novel sequencing technology represented by PacBio's SMRT technology and Oxford Nanopore Technologies' nanopore single-molecule technology is called the third-generation sequencing technology (TGS). Compared with the previous sequencing technology, it doesn't need PCR amplification during sequencing, so no GC preference is introduced. It can achieve an average read length of over 10kb[46]. Furthermore, methylation information can be directly detected in third-generation sequencing techniques, and epigenetic recognition can be performed simultaneously.

1.3.2 Next-generation sequencing strategies

1.3.2.1 DNA sequencing

The advent of DNA sequencing methods has greatly facilitated research and discovery in biology and medicine. DNA sequencing has become an indispensable tool in basic biological research and numerous applications, such as the diagnosis of cancer or other diseases, biotechnology, forensic biology, and biosystematics[47, 48].

WGS sequenced the entire genome to provide the most comprehensive genome features. It

can obtain all gene sequences and help to clarify the factors that affect the occurrence and progression of diseases. Among next-generation sequencing strategies, WGS is more expensive. However, it provides extensive information on point mutations, gene fusions, indels, and copy number variations (CNV), as well as information on complex rearrangements of chromosomes. In addition, WGS can detect genomic mutations outside the coding region of genes[49]. This includes non-coding somatic mutations such as promoters, enhancers, introns, non-coding RNAs, and unannotated regions.

WES can detect mutations of coding genes in the genome. Although it has limitations in detecting structural variation[50] and is not able to detect non-coding region variation, the cost and analysis time of WES are greatly reduced, and the coverage of the region of interest and the accuracy of mutation information have been improved[51]. Compared with WGS, WES is mainly used to characterize the defects of single-gene diseases (Mendelian genetic diseases), which can cause rare familial diseases[52]. In addition, WES also has great potential in non-hereditary or new mutation-related diseases. Therefore, it can be used to detect known mutations and new mutations in tumor samples[53].

Targeted sequencing is a research strategy in which genomic regions of interest are enriched and sequenced by techniques such as gene probe capture and PCR amplification. According to different applications, ultra-high sensitivity and accuracy can be obtained with a small amount of data[54]. Compared with whole genome sequencing and whole exome sequencing, targeted sequencing focuses on the region of interest and eliminates the interference of redundant data. With low sequencing cost and deep sequencing depth, it can maximize the use of sequencing reads, especially in clinical applications[55]. For example, the size of the whole human genome is about 3Gb while the exon region only accounts for 2% (about 60M). A single WGS sample of 30-50X has an output data of 90-150Gb, and a single WES sample of 100-200X has an output data of 6-12Gb. For a panel with a target region size of 2M, the data volume is only 4Gb when the sequencing depth reaches 2000X. In recent years, many commercial companies and scientific researchers have developed their own gene panels. For

example, TruSight Cancer of Illumina Company provides a gene panel of 94 genes and 287 SNPs related to breast cancer, and the Ion AmpliSeq Cancer Hotspot Panel of Ion Torrent Company provides tumor detection panel which contains 2800 hot spot mutations[56, 57]. In addition, Welch et al. used a gene panel of their own design in a study of drug response to decitabine in patients with acute myeloid leukemia and myelodysplastic syndrome[58].

1.3.2.2 RNA sequencing

RNAs are critical to gene expression, both in the form of mRNAs and in the form of non-coding RNAs that regulate transcription, such as lncRNAs[59] or snRNA[60]. There is evidence that RNA processing is systematically altered in cancer cells, demonstrating that RNA has an important impact on tumorigenesis, growth and progression[61]. RNA-Seq analysis is a useful way to obtain insights into cancer genome alterations. RNA-seq extracts transcriptome RNA from biological samples, obtains cDNA by reverse transcription, and then sequences the cDNA. Through RNA-seq, a complete transcriptome sequence can be obtained to reflect the gene expression in the sample. Because the content of the transcriptome is highly variable in the body, the analysis of the transcriptome can only represent the expression of genes in the body when the transcriptome is obtained. RNA-seq is highly sensitive and effective in detecting gene fusion, somatic mutation, and gene expression[62]. In cancer research, the use of RNA-seq to detect gene expression and transcriptome changes helps to understand the classification and progression of tumors[63].

1.3.2.3 Bisulfite sequencing

The term epigenetics was defined by Riggs et al. as "*any heritable changes in gene function that cannot be explained by changes in the DNA sequence*"[64]. Its important feature is that the

DNA sequence is unchanged, but gene expression and phenotype undergo changes that can be stably transmitted during development and cell proliferation. For decades since the discovery of the DNA double helix, it has been assumed that genes determine all biological phenotypes. But there are still unexplained phenomena, such as identical twins who share the same genome but have vastly different personalities and health when raised in the same environment. With the deepening of research, the concept of epigenetics is used to explain these phenomena that cannot be explained by classical genetic theory. Epigenetics includes DNA methylation, histone methylation, Non-coding RNA interference, etc. Among them, due to the close relationship between DNA methylation and tumors, such as the inactivation of tumor suppressor gene transcription caused by CpG island methylation, the global hypomethylation that induces genomic instability[65], and the unwanted activation of transposons leading to further genetic damage[66], DNA methylation has become an important focus in cancer research.

1.3.3 Application of Next Generation Sequencing in Cancer Diagnosis

NGS can provide detailed information about the tumor genome and provide data support for researchers to understand the generation and development of tumors. In the early stage of the development of NGS technology, tumor research mainly focused on exome sequencing with a relatively small amount of data[67]. With the development of NGS technology, the improvement of sequencing throughput and the continuous reduction of sequencing costs, the research field gradually expanded to multi-omics research involving a greater amount of data[68]. Recently, the research on liquid biopsy, immunotherapy, and the relationship between microorganisms and tumors has attracted more and more attention. In general, the analysis of tumor genome sequencing data mainly includes the following: point mutation, indel, copy number variation, structural variation, methylation, pathogen integration (such as HBV, HPV)[69], etc.

Oncogene mutations are generally functional or active mutations, showing hot spot mutations, while tumor suppressor genes are inactivated mutations, showing scattered mutations[70]. The research of point mutations and indels mainly focuses on oncogenes and tumor suppressor genes, as well as the related genes of some specific cancer species. Some cancer patients' cancer cell genomes show large-scale copy number variation. For example, in ovarian cancer, pancreatic cancer, prostate cancer, and other tumors[71], homologous recombination deficiency causes DNA double-strand break repair to rely excessively on low fidelity DNA damage repair pathways such as non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ)[72] and single-strand annealing (SSA), leads to insertion/deletion of nucleic acid sequence and abnormal copy number[73]. Structural variation is also one of the characteristics of tumor cells, including rearrangement deletion, amplification, translocation, and so on. A study has shown that SVs can be classified into 16 different patterns (Figure 1-6), and these patterns show uneven distribution in different tumor types[74]. Viruses are closely related to the occurrence of cancer. For example, about 99.7% of all cases of cervical cancer are caused by human papillomavirus (HPV)[75]. Besides, Epstein Barr virus has been shown to cause many different types of cancer, such as lymphoma, gastric cancer and nasopharyngeal carcinoma[76]. Abnormal DNA methylation can lead to the activation of oncogenes or the inactivation of tumor suppressor genes[77]. In addition, studies have shown that genome-wide methylation can be applied to ctDNA early tumor detection and detection of measurable residual disease (MRD)[78, 79, 80, 81].

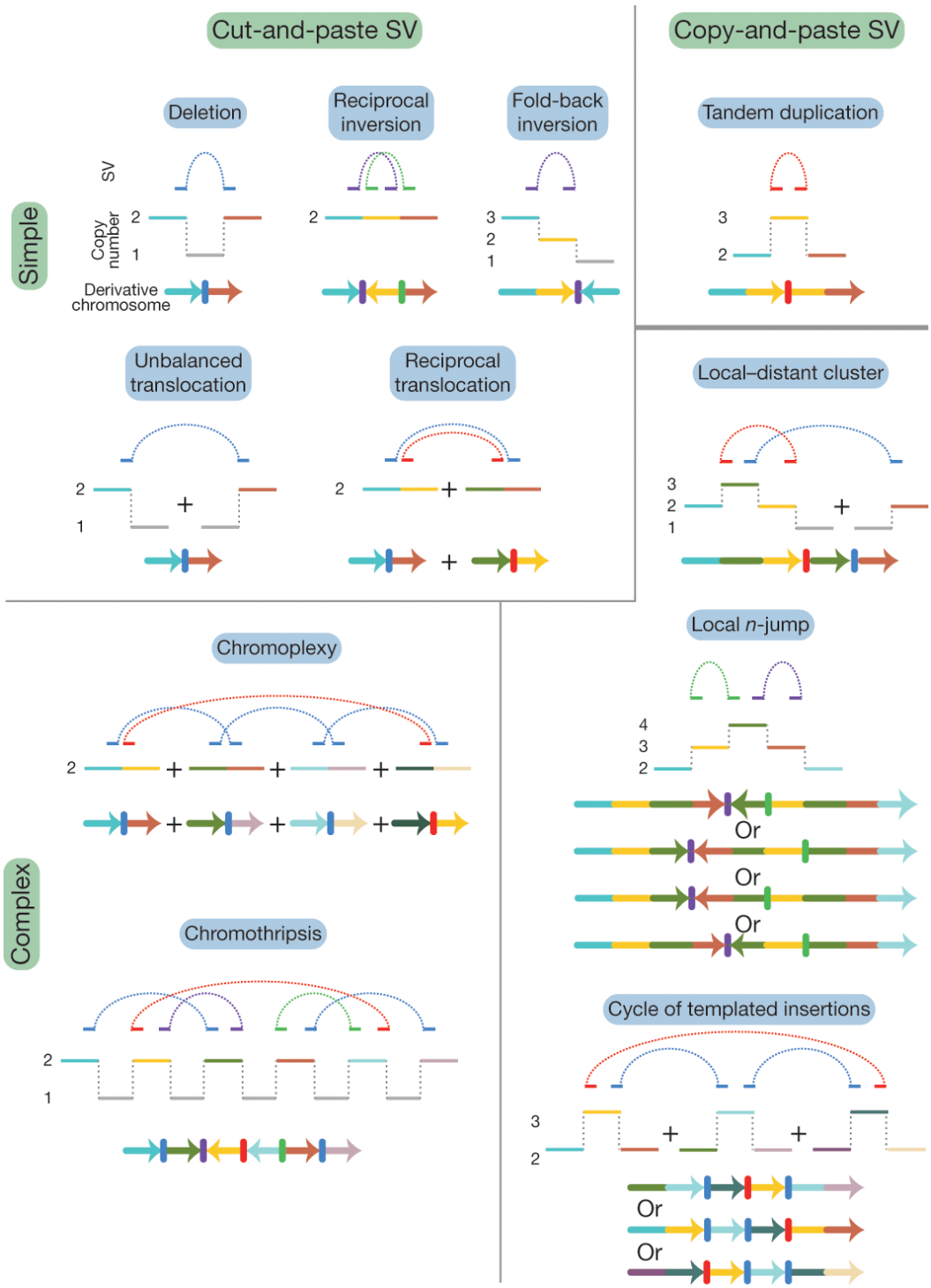


Figure 1-6 Schematic diagram of the major types of structural variants. Each type is divided into three parts, the top dotted arcs represent rearrangement junctions connecting the two chromosomal segments, and the middle part represents the copy number of the genomic segments. The bottom shows the final chromosome configuration (from Li Y, et al. 2020)[74].

1.4 ctDNA sequencing method

As mentioned above, tissue biopsy has inherent limitations, such as invasiveness and tumor heterogeneity. Compared with tissue biopsy, liquid biopsy is getting more and more attention in tumor diagnosis and treatment because it is non-invasive and makes it easy to monitor the tumor progress of patients.

1.4.1 Tumor DNA fraction detection and CNV detection in cfDNA

At present, imaging methods are commonly used in clinical practice to detect early tumors or monitor tumor progression. Commonly used imaging methods include computer tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI). However, the evaluation of the results of these methods is sometimes subjective and limited, and it is often difficult to distinguish when the tumor size is small[82]. The detection of tumor DNA fraction can provide important information about tumor progression, treatment response, and prognosis, helping to guide clinical decision making and individualized treatment[83]. When using cfDNA to detect tumor, tumor DNA fraction represents the proportion of ctDNA in the cfDNA. A higher tumor DNA fraction is generally associated with a poorer prognosis and shorter survival[84]. If the tumor DNA fraction continues to increase, current treatment options may need to be reevaluated and additional treatment options considered.

As mentioned in the previous section, CNV detection plays a key role in studying the mechanism of tumorigenesis, guiding treatment decisions, and evaluating prognosis. First, CNV detection can determine the copy number change of a gene or chromosome region in a tumor cell, helping to detect the deletion of tumor suppressor genes and the expansion of oncogenes. Deletion of tumor suppressor genes may lead to abnormal cell proliferation and

tumor formation, while expansion of oncogenes may promote tumor cell growth and survival[85]. Second, certain drug targets are often associated with copy number changes in specific genes or chromosomal regions. For example, HER2 amplification in breast cancer is associated with sensitivity to Trastuzumab treatment[86]. In terms of detecting tumor DNA fraction, CNV detection is also an important basis. This will be described in detail in the following section.

1.4.2 Tumor DNA fraction estimation methods

1.4.2.1 Tumor DNA fraction estimation based on CNV

One approach is to determine tumor DNA fraction in cfDNA based on CNV detection. In general, the CNV-based method to detect the tumor DNA fraction in total cfDNA starts by dividing the genome into bins of the same length. These bins usually range from kilobase to megabase[87]. The reads number or average depth in these bins is observed, and the bins with adjacent positions and similar reads number or average depth are combined to achieve genome segmentation. Of course, because of the effects of GC bias and mapping bias[88], paired normal samples need to be used to eliminate these biases. At present, the commonly used method to eliminate bias is to generate a fitting curve by loess regression[89] to correct the original sequencing data. After segmentation, the likelihood of different copy numbers corresponding to the obversed reads number or average depth are calculated, and the optimal tumor DNA fraction and CNV status can be calculated through maximum likelihood estimation. Common tools are ABSOLUTE[90], AbsCN-seq[91], and so on. In addition, some tools are used to detect allele specific CNV and predict the tumor DNA fraction by replacing reads number or average depth with allele specific frequency of SNP sites, such as FACETS[92], Sequenza[93].

The above methods are designed based on WGS or WES data of tumor tissue samples, which

have certain requirements on the depth of sequencing data and need paired normal samples to eliminate bias. In 2017, the researchers developed ichorCNA, which is suitable for low-coverage whole-genome sequencing (lcWGS) of ctDNA, to detect tumor DNA fraction[94]. It counts genomic reads using the tool in the HMMcopy Suite[95] and then normalizes the read counts to correct the GC content and mapping bias. Specifically, ichorCNA uses the cfDNA sequencing results of 27 healthy donors as the standard reference dataset, calculates the log2 copy ratio between the sample to be tested and the reference in each bin, and then uses the Hidden Markov model (HMM) to predict the segments with copy number changes. Finally, according to the above results, the corresponding clones, tumor DNA fraction, and subclone information can be estimated.

1.4.2.2 Tumor DNA fraction estimation based on specific variation frequency

In addition to the above CNV-based method, tumor DNA fraction can also be detected based on specific variation frequency[96], such as single nucleotide variations, structural variations, etc. Because the tumor content in cfDNA is generally low, detecting mutations usually requires very high coverage. High coverage sequencing of the entire genome or exome is expensive, so it is not a good option for many patients. The researchers developed a method, CAncer Personalized Profiling by deep Sequencing (CAPP-Seq), to address this challenge. Based on mutations frequently observed in the Catalogue of Somatic Mutations in Cancer (COSMIC) database, as well as mutations in WGS data from The Cancer Genome Atlas (TCGA) database, CAPP-Seq designed probes to cover exon and intron regions in genes containing common mutations. CAPP-seq efficiently concentrates the sequencing segment to only 0.004% of the total genome size, enabling subsequent ultra-deep sequencing.[97] This technique is capable of detecting tumor-derived ctDNA with high sensitivity and specificity, while also being cost-effective.

The Avenio ctDNA analysis kit used in subsequent articles is based on CAPP-Seq technology. This is a kit for ctDNA analysis launched by Roche. It detects variations from 17 important lung cancer and colorectal cancer-related genes and uses the molecular-barcoding method to reduce sequencing errors. In addition, AVENIO's ctDNA analysis software (Roche) leverages integrated digital error suppression (iDES) to remove PCR duplicates and stereotypical errors[98]. At present, this kit is only for scientific research. It offers a comprehensive genomic map across four mutation categories: single nucleotide variation (SNV), indel, copy number variation (CNV), and fusion, with the aim of helping researchers to explain the genomic complexity of tumors. This kit was used in project K34R of this study.

1.4.3 Limitations of the tumor DNA fraction estimation methods

It is crucial to recognize that both the CNV-based and the variation-based methods have their respective limitations. CNV-based approaches encounter difficulties when the tumor genome approximates diploidy. Furthermore, in certain cancer types like thyroid carcinoma (THCA) and kidney renal clear cell carcinoma (KIRC), CNV occurrence is infrequent[99]. Due to the lack of sufficient aneuploidy and chromosomal instability, tools such as ichorCNA and ACE[100] may not provide reliable estimates of the tumor DNA fraction. The shortcoming of the method of estimating tumor DNA fraction by variation is that the probes cover only a small part of the genome. Some patients do not detect enough variants in the CAPP covered area, but this does not rule out tumor positivity.

1.4.4 Future directions

In summary, while the aforementioned methods have demonstrated promising outcomes in numerous studies, it is important to acknowledge their inherent limitations. To enhance the accuracy of estimation of ctDNA tumor DNA fraction, it is crucial to consider the specific

characteristics of different cancer types. Selecting appropriate methods based on these characteristics or employing a comprehensive approach can improve the reliability of results.

Researchers have shown that the length of ctDNA fragments is typically shorter than that of cfDNA fragments derived from normal cells. In rat, the main fragments length of ctDNA derived from human glioblastoma multiforme and hepatocellular carcinoma was about 134-144 bp. The length of the main fragments in normal sample was about 167bp. The same thing happened in melanoma. In addition, the selection of cfDNA with shorter fragments lengths in lung cancer can increase the frequency of detection of EGFR mutations.[101] Based on these findings, efforts should be made to enrich fragments approximately 140bp. By focusing on this specific fragment size range, it may be possible to improve the precision and reliability of CNV detection.

The corresponding leukocyte sequencing data can be used to reduce the impact of clonal hematopoiesis. A study has indicated that most somatic mutations detected in cfDNA of lung cancer patients are attributed to clonal hematopoiesis, which are non-recurrent. Compared to tumor-derived mutations, clonal hematopoietic mutations tend to occur on longer cfDNA segments and lack the mutational signature associated with smoking.[102]

1.5 Aim of this study

With the continuous development of technology, ctDNA detection has a broad application prospect in tumor diagnosis, treatment, and monitoring. In recent years, the investigation of ctDNA has been increasingly discussed as an alternative to tumor tissue analysis.[14, 103, 104, 105] Compared with tissue biopsy, the use of ctDNA makes it easy to monitor the tumor progress by allowing repeated noninvasive sampling. However, due to the low content of ctDNA in cfDNA, ctDNA detection requires high sequencing depth to achieve high sensitivity. In this case, because of the high cost, it is unrealistic to sequence the whole genome at high

coverage. To solve this problem, two commonly used schemes at present include low-coverage WGS and high-coverage targeted sequencing. However, both methods have certain limitations. With high-coverage targeted sequencing, SNVs in the detection interval can be accurately detected, but the detection ability of other variations such as SVS and CNV is limited. With lcWGS, while the detection of SNVs may be limited, it still enables the identification of CNVs, providing valuable insights into genomic alterations. Combining the advantages of the two sequencing methods may provide a more holistic view of the tumor development and progression.

Based on the data of HIPO-K34 and INFORM, the main purpose of this study is to investigate and enhance the detection capability of ctDNA by the two methods. The HIPO-K34 study focused on patients with non-small cell lung cancer (NSCLC) with ALK gene fusion, incorporating multiple time-point sequencing data obtained from both lcWGS and panel sequencing. The INFORM project includes both liquid biopsy samples and tissue samples collected at the same time point from the same patient.

In this study, the consistency of the tumor DNA fraction estimation ability of CAPP-seq and lcWGS was tested based on the data of HIPO-K34. After that, various tools for tumor DNA fraction estimation and CNV detection were benchmarked using simulated data with predetermined CNVs. And then, the performance of those tools was further examined by enriching short fragments of cfDNA. Finally, the tool with the best performance was selected and optimized to improve its accuracy. Since the INFORM project includes both liquid biopsy samples and tissue samples, the accuracy of ctDNA detection by the optimized tool could be verified against the results of tissue samples, thereby evaluating the effectiveness of the optimized tool. At last, a pipeline was established to combine the panel analysis process with the optimized lcWGS analysis process to realize the accurate analysis of liquid biopsy samples.

2. Comparison of Tumor DNA Fraction Estimation Method for Liquid Biopsy Samples

In this chapter, the analysis was based on project HIPO-K34. In this project, my job was to perform bioinformatics analysis of the sequenced data. To use liquid biopsies for monitoring tumor diseases, it is a prerequisite to determine the tumor DNA fraction in cfDNA samples. In this chapter, I compared two methods to detect the tumor DNA fraction for liquid biopsy samples.

2.1 Introduction

The samples of the project HIPO-K34 are from non-small cell lung cancer (NSCLC) patients with ALK gene fusion. Anaplastic lymphoma kinase (ALK) fusion gene may lead to NSCLC. In most ALK-positive cases, the EML4 gene which is located at the 5' end of chromosome 2p inverses and fuses with ALK[106] (Figure 2-1). Due to the EML4 promoter, the fusion gene (EML4-ALK) is activated and expressed, thereby inducing cell proliferation and development of the tumor.

At present, a variety of drugs have a good therapeutic effect on this type of cancer. They specifically bind to the fusion gene through tyrosine kinase inhibitors (TKIs), thereby inhibiting the expression of the fusion gene and improving the survival of NSCLC patients. Unfortunately, almost all current targeted therapies against this mutation inevitably suffer from resistance problems[107]. One of the common reasons for drug resistance is the secondary mutation at the position where the original TKI is bound so that the original TKI cannot be well combined with

the fusion gene (Figure 2-2). The drug resistance caused by this reason can be solved by replacing a new generation of TKIs, which have a different binding site[108]. Therefore, we need to use ctDNA sequencing method that is less harmful to patients, convenient and low-cost, to clearly know the patient's cancer development status, tumor DNA fraction, and whether there are mutations at certain key sites.

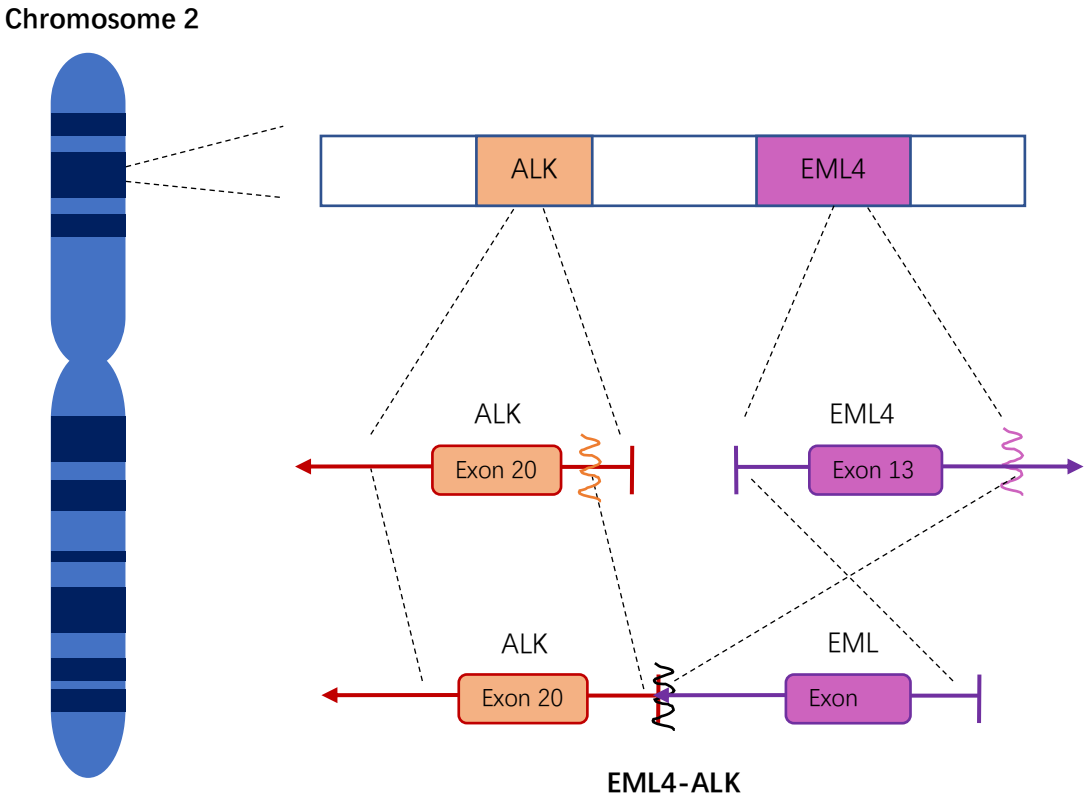


Figure 2-1 ***Schematic diagram of the fusion process of EML4 and ALK genes.*** *The ALK gene and the EML4 gene are both located on chromosome 2. The two genes are on opposite chains and with a distance of about 10MB in between. The fusion gene of ALK and EML4 generates the fusion protein EML4-ALK, which leads to abnormal activation of downstream signaling pathways such as AKT, STAT3, and ERK1/2, so that cells can proliferate, resist apoptosis, and eventually induce cancer. (Soda M et al., 2007)*[109].

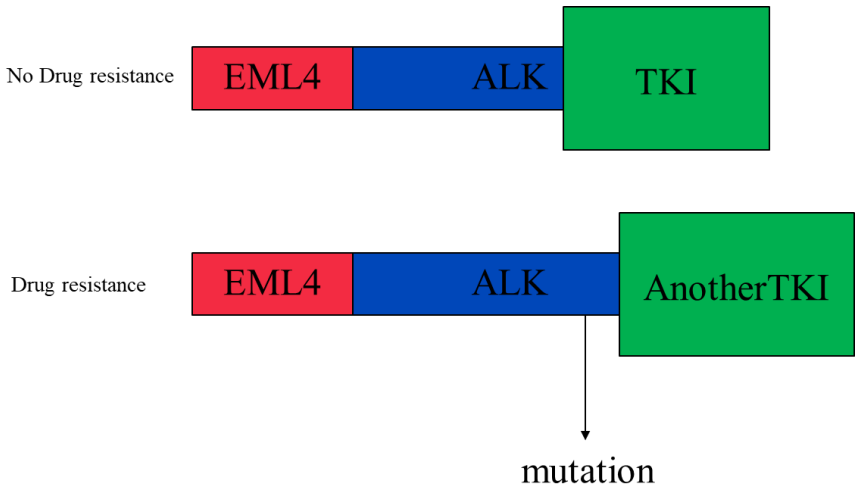


Figure 2-2 ***Secondary mutations lead to drug resistance.*** *TKIs specifically bind to the fusion gene, thereby inhibiting the expression of the fusion gene. And the secondary mutation at the TKI binding site leads to drug resistance. The drug resistance caused by this reason can be solved by changing the targeted drugs.*

Liquid biopsy is rapidly becoming an important minimally invasive aid for standard tumor biopsy. ctDNA sequencing can be used to monitor tumor progression and the development of drug resistance mechanisms[110]. Because the amount of ctDNA can reflect the patient's tumor load, for a better understanding of tumor evolution and drug resistance mechanisms, a method needs to be developed to accurately describe the tumor load for patients.

Here I used two methods to estimate the tumor DNA fraction. One of the methods was to estimate the tumor DNA fraction through the SNVs' allele frequency in the sample, and the other was to use the CNV detection tool. In this project, the CNV detection tool I used was ichorCNA[94].

2.2 Methods

2.2.1 Data

In this project, plasma samples were collected at several time points from 87 ALK-positive patients. The patients with metastatic NSCLC had received TKIs treatment at the Thoraxklinik Heidelberg, Germany, and the Lungenclinic Großhansdorf, Germany[111]. In total, 416 lcWGS data and 402 panel sequencing data were acquired. 395 samples contain both lcWGS and panel sequencing data (Figure 2-3). In this project, panel sequencing (average 4100x coverage) was used to detect mutations and fusions in the target genes, and lcWGS (average 0.5x coverage) was used for global copy number variant analysis from cfDNA. The data I used in this project included the BAM files of the lcWGS samples, and the results of the panel samples, which were analyzed by the bioinformatics analysis workflow of the Avenio platform. The Avenio analysis workflow reported the potential variants for the 402 panel sequencing samples. Most of the variants exhibited a MAF value below 0.2. However, it is crucial to note that the MAF values of some samples were concentrated around 0.5 and 1. Since all the samples in this project were from liquid biopsy, the proportion of ctDNA within cfDNA was relatively low. The mutation frequencies around 0.5 and 1 likely indicated germline mutations rather than specific mutations in tumor cells. Therefore, I removed variants with a MAF value larger than 0.4. The remaining variants were used for subsequent analysis. My job was to conduct downstream bioinformatics analysis based on these data.

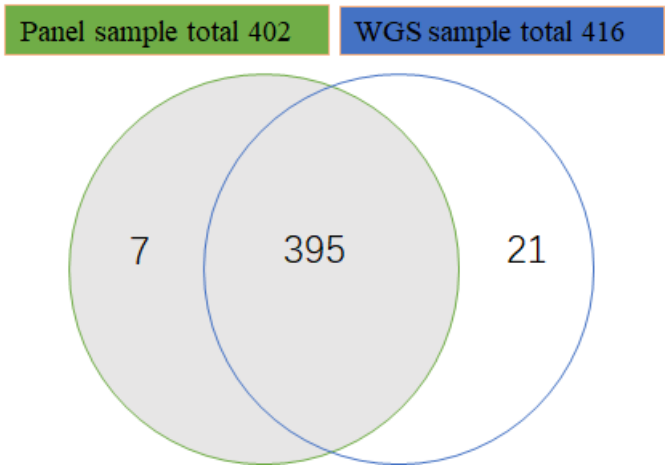


Figure 2-3 Sample composition of project HIPO-K34. Among them, 402 samples were panel sequenced, and 416 samples were whole genome sequenced with low coverage. A total of 395 samples were subjected to both panel sequencing and lcWGS.

For 87 patients, sampling was performed every two months, and the samples were subjected to lcWGS and panel sequencing. Of these, 21 patients were sampled only once, and 13 patients were sampled at two time points. The remaining 53 patients had greater than or equal to 3 sampling times at different time points. Taking multiple samples from the same patient at different time points can well track the development of cancer and detect drug resistance. One of the patients underwent 31 samplings spanning five years (Figure 2-4).

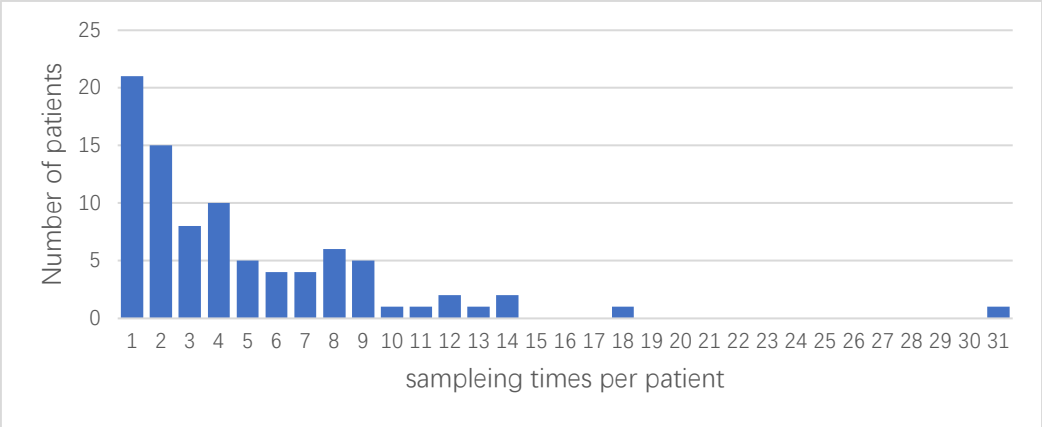


Figure 2-4 Distribution of sampling times per patient. The x-axis represents sampling times, the y-axis represents patients' numbers. More than half of the patients had sampling times greater than 2 at different time points.

2.2.2 Tool and parameter selection

In this project, I used ichorCNA (implemented in R 3.3.1) to do CNV detection for the lcWGS samples. Bin size is an important parameter for ichorCNA. In the process of genome segmentation, bins with similar reads number that are adjacent to each other on the genome need to be merged. Different bin sizes will affect the segmentation results[112]. IchorCNA provided three bin sizes, namely 10kb, 500kb, and 1MB. A comparison was carried out to select the best parameter. The command was as follows to test different bin sizes by changing the parameter '*--window*'.

```
readCounter \
        --window 1000000 --quality 20 \
        --
chromosome "1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X,Y" \

        bamfile > wigfile
```

2.2.3 Quality control

2.2.3.1 Sequencing quality control by fastqc

In this study, I conducted rigorous quality control procedures to ensure the reliability of sequencing data from 416 cfDNA lcWGS samples in BAM format. To accomplish this, I used the FastQC [113] software (version 0.11.9) for the purpose of assessing data quality, complemented by the application of the MultiQC [114] software (version 1.10) for the aggregation and visualization of results derived from FastQC. FastQC is designed to assess the quality of sequencing data. It offers a multitude of quality-related graphs and statistical

information, facilitating the swift recognition of issues and enabling the implementation of appropriate corrective measures. MultiQC is a tool specifically designed for the integration and visualization of reports from multiple analysis tools. It offers the invaluable capability to consolidate and visualize results from all samples in the project within a single report, providing a comprehensive overview of data quality across the entire dataset. Prior to FastQC analysis, I secured the lcWGS sequencing data for 416 cfDNA samples, available in their original BAM file format. The procedural framework for quality control is as follows:

1. Software Installation: Initially, the installation of FastQC and MultiQC was executed, accompanied by the validation of their configurations. In this study, FastQC version 0.11.9 and MultiQC version 1.10 were employed.
2. Running FastQC: For each sample's FASTQ file, the following command was executed to run FastQC:

```
fastqc sample.bam
```

   Here, sample.bam represents the BAM file under analysis.
3. Running MultiQC: To run MultiQC and display results for FastQC from the directory containing FastQC result files, the following command was executed:

```
multiqc /path/to/fastqc_results/
```

   Here, /path/to/fastqc_results/ represents the directory path where the FastQC result files were located.
4. Interpreting MultiQC Results: MultiQC generated a comprehensive HTML report that compiles information from FastQC results, allowing for the review of data quality for multiple samples within a single report.

2.2.3.2 GC bias removal

PCR amplification plays an important role in the generation of GC bias[115]. GC bias will affect the accuracy of CNV detection, so it is necessary to remove GC bias. ichorCNA provides its own function '*correctReadCounts*' to remove GC bias from samples. To check the sample quality, I modified this function and output the log2 ratio for each bin as a scatter plot. The steps are as follows.

1, Output reads number after GC bias removal. Below is the code of function '*correctReadCounts'*, the part marked with '#' was added to output the reads number after GC bias removal.

```
correctReadCounts <- function(x, chrNormalize = c(1:22), mappability = 0.9, samplesize
 = 50000, verbose = TRUE) {
    if (length(x$reads) == 0 | length(x$gc) == 0) {
        stop("Missing one of required columns: reads, gc")
    }
    chrInd <- as.character(seqnames(x)) %in% chrNormalize
    if(verbose) { message("Applying filter on data...") }
    x$valid <- TRUE
    x$valid[x$reads <= 0 | x$gc < 0] <- FALSE
    x$ideal <- TRUE
    routlier <- 0.01
    range <- quantile(x$reads[x$valid & chrInd], prob = c(0, 1 - routlier), na.rm = TRUE)
    doutlier <- 0.001
    domain <- quantile(x$gc[x$valid & chrInd], prob = c(doutlier, 1 - doutlier), na.rm = T
RUE)
    if (length(x$map) != 0) {
        x$ideal[!x$valid | x$map < mappability | x$reads <= range[1] |
        x$reads > range[2] | x$gc < domain[1] | x$gc > domain[2]] <- FALSE
    } else {
        x$ideal[!x$valid | x$reads <= range[1] |
        x$reads > range[2] | x$gc < domain[1] | x$gc > domain[2]] <- FALSE
    }
    if (verbose) { message("Correcting for GC bias...") }
    set <- which(x$ideal & chrInd)
    select <- sample(set, min(length(set), samplesize))
    rough = loess(x$reads[select] ~ x$gc[select], span = 0.03)
    i <- seq(0, 1, by = 0.001)
    final = loess(predict(rough, i) ~ i, span = 0.3)
    x$cor.gc <- x$reads / predict(final, x$gc)
    if (length(x$map) != 0) {
        if (verbose) { message("Correcting for mappability bias...") }
        coutlier <- 0.01
```

```
        range <- quantile(x$cor.gc[which(x$valid & chrInd)], prob = c(0, 1 - coutlier), na.r
m = TRUE)
        set <- which(x$cor.gc < range[2] & chrInd)
        select <- sample(set, min(length(set), samplesize))
        final = approxfun(lowess(x$map[select], x$cor.gc[select]))
        x$cor.map <- x$cor.gc / final(x$map)
    } else {
    x$cor.map <- x$cor.gc
    }
    x$copy <- x$cor.map
    x$copy[x$copy <= 0] = NA
    x$copy <- log(x$copy, 2)
###################################
# Output log2 ratio of reads number of GC content distribution
###################################
    write.table(x$copy,"outfilepath.csv",row.names=FALSE,col.names=TRUE,sep=",")
    return(x)
}
```

2, Plot the log2 ratio. The reads number after GC bias removal in each bin of a tumor sample was divided by the reads number in the corresponding bin of the healthy donor (provided by ichorCNA). And then logarithm was taken to obtain the log2 ratio. For all bins in the sample, a scatter plot was drawn with GC content as the x-axis and log2 ratio as the y-axis.

2.2.3.3 Genomic Fingerprints

The purpose of genomic fingerprinting is to uniquely identify genomes from the same person. It can be used to check samples swaps. In this project, genomic fingerprinting played a crucial role due to there were not only samples from multiple timepoints but also samples from different sequencing types (Panel sequencing and lcWGS) per patient. It played a significant role in ensuring that samples associated with the same Patient Identifier (PID) indeed originated from the same patient. This method greatly enhanced the credibility and reproducibility of the study, particularly when dealing with samples collected at different time points. The details are as follows:

1. Format Conversion: Because Panel Sequencing samples were aligned using GRCh38, while lcWGS employed GRCh37, it was necessary to initially utilize the LiftOver[116] （default parameter），an assembly converter software to convert the mutation data from Panel Sequencing samples from GRCh38 format to GRCh37 format. This ensures that all samples' variant information was compared on the same genome version.
2. Collection of Variant information from all Panel Sequencing samples. First, the variants that appeared in all panel sequencing samples were collected, totaling 432 variants. The chromosome positions, reference bases, and variant bases of these 432 variants were recorded. Next, for each sample, a fingerprint file (with the extension .fp) was generated. This fingerprint file contains four columns: the chromosome position, the reference base, the variant base, and marker. If the genotype of the sample at one chromosome position matched the reference base, the marker was labeled as 0. If a mutation occurred and the variant matched the variant base, the marker was labeled as 1. If there was no coverage at the position, the marker was labeled as 0. The final fingerprint file consisted of 432 rows, with each row representing the variant information for one specific genomic position.
3. Sample Comparison and Spearman Correlation Calculation: By comparing the .fp files of any two samples and examining the base differences at the 432 variants, the values of the marker column were utilized to calculate the spearman correlation coefficient, which represents the similarity between the pairs of samples.
4. Visualization of Correlation Coefficients: The correlation coefficients between all pairs of samples were depicted as a heatmap, with colors indicating the strength of the relationships. Through clustering analysis, samples with higher correlation coefficients were grouped together, aiding in the identification of potential sample swaps. The heatmap was generated using the 'pheatmap' package in R (version 3.3.1). Clustering was achieved by setting the parameters '*cluster_rows*' and '*cluster_cols*' to TRUE.

2.2.4 Estimate tumor DNA fraction

For monitoring the tumor development of the patients, a method was needed to evaluate the tumor DNA fraction. The two commonly used methods are CNV based method and mutation-based method.

In this chapter I used ichorCNA to estimate the tumor DNA fraction of lcWGS data. The default parameters of ichorCNA are described in Table 2-1. IchorCNA presets different ploidy and normal contamination in the initial stage. For each preset, a CNV state suitable for the whole genome is calculated based on the observed reads number in each bin. Then, the optimal solution is determined by judging the log-likelihood score of each preset. The optimal preset will be taken as the ploidy and tumor DNA fraction of the sample.

Table 2-1 The main default parameter of IchorCNA. IchorCNA sets the initial ploidy as 2 and 3, and the initial normal contamination as 0.5-0.9. To reduce errors caused by excessive copy number, total clonal CN states is set as 5. For subclonal copy number states, only 1 and 3 are considered.

Initial ploidy	2 and 3
Initial normal contamination	0.5, 0.6, 0.7, 0.8 and 0.9
Total clonal copy number states	5
Subclone copy number states	1 and 3

In this project, I also used the mutation-based approach to estimate tumor DNA fraction of the panel sequencing data. Mutation allele frequency (MAF) is an important index of tumor DNA content. First, single nucleotide polymorphism sites were excluded. Second, considering the subclonal diversity of tumor cells, the presence of mutations with different frequencies

should be observed. The mutations occurred in most tumor cells can reflect the tumor DNA fraction. In theory, the histogram of the MAFs should draw to find the identifiable peaks and the MAF value corresponding to the highest peak should be selected for tumor DNA fraction estimation. In this project, the number of mutations was small in most samples, so the highest MAF value was multiplied by 2 to represent the tumor DNA fraction of the sample. The MAF was multiplied by 2 because biallelic mutations at the site are uncommon.

To show more clearly how to determine tumor DNA fraction based on the MAF value, I used a sample (K34R-2VL6V1_tumor1-b13) to illustrate. As shown in Table 2-2, a total of 4 mutations were detected in sample K34R-2VL6V1_tumor1-b13, and their MAF value were 0.00964957, 0.00457038, 0.0353101 and 0.00375235, respectively. Among them, the highest MAF value was 0.0353101, and the tumor DNA fraction of this sample was estimated by 0.0706202, which was twice that of 0.0353101.

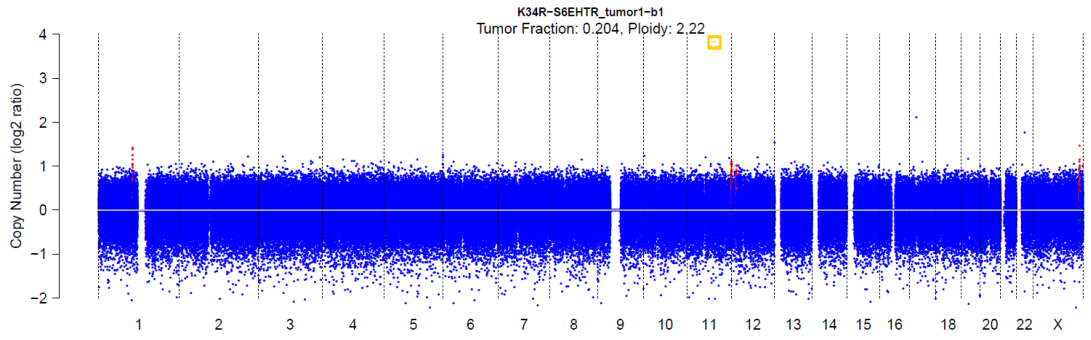
Table 2-2 MAF value of detected mutations of sample K34R-2VL6V1_tumor1-b13 after removal of SNP sites. There are 4 SNVs in this sample, among which the largest MAF value is 0.0353101 occurring at the chr2:29220765 position (green background). The tumor DNA fraction of this sample is twice the mutation frequency, which is 0.0706202.

Chromosome	Position	Ref	Alt	MAF value
chr2	29220747	C	T	0.00964957
chr2	29220759	G	A	0.00457038
chr2	29220765	G	T	0.0353101
chr2	29222347	A	G	0.00375235

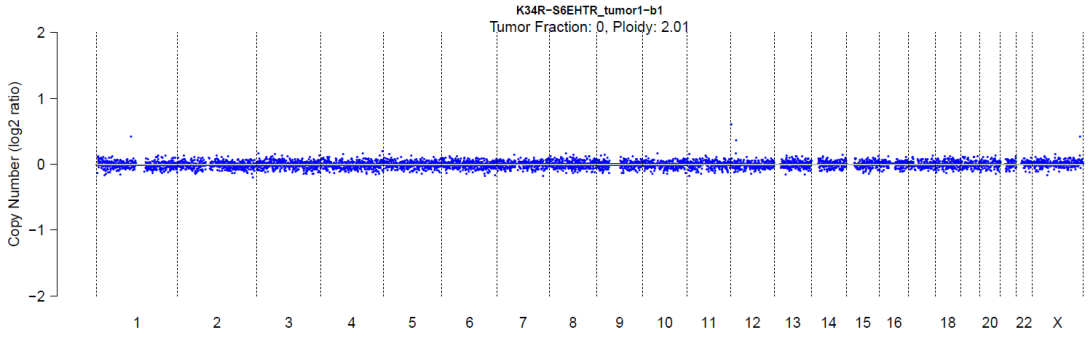
2.3 Results

2.3.1 Bin size selection

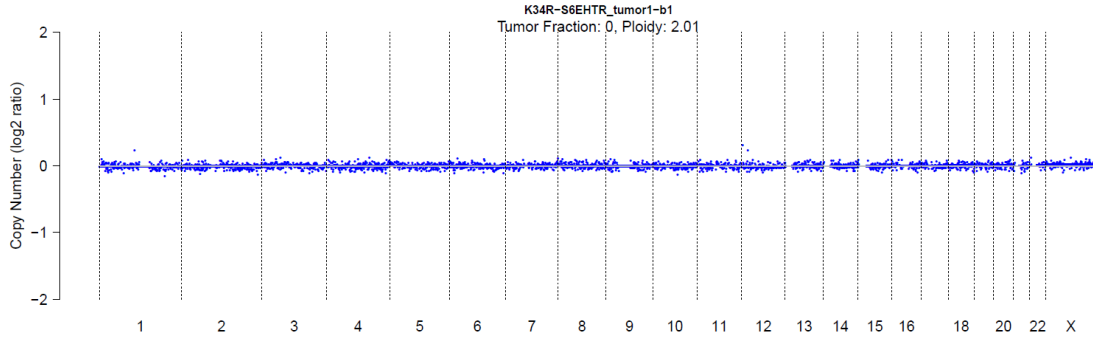
To determine the bin size of ichorCNA, I selected a sample in which no mutations were detected, including SNVs, CNVs, and fusions. The CNV results (by ichorCNA) of this sample (K34R-S6EHTR_tumor1-b1) according to different bin sizes are shown in Figure 2-5.



(a)



(b)



(c)

Figure 2-5 The influence of different bin sizes on the analysis results by ichorCNA (sample K34R-S6EHTR_tumor1-b1). *The x-axis represents different chromosomes, the y-axis represents the log2 ratio of copy number. In addition, ichorCNA also reported the corresponding tumor fraction and ploidy. The bin size of (a) is 10kb, the bin size of (b) is 500kb, and the bin size of (c) is 1Mb.*

It can be found that when 10kb was selected, the tool considered part of the noise to be CNV and detected that the sample contains tumor with a tumor fraction of 0.204. When a larger bin size was selected, (b) and (c) yield different results from (a), both indicating that the sample did not contain a tumor. Because there were no mutations in sample K34R-S6EHTR_tumor1-b1 through panel sequencing, (b) and (c) appear to be more credible. In the CNV detection of lcWGS, the largest challenge is the extremely low tumor DNA fraction in the sample. This causes noise to have a huge impact on CNV detection, so it is very important to improve the ability to resist noise[117]. Therefore, after comparing the analysis results of three different bin sizes, the bin size of 1MB was used for subsequent analysis.

2.3.2 QC results

2.3.2.1 MultiQC results

Before doing other data analysis, it is necessary to conduct basic quality control of the 416 lcWGS samples. Only when the quality of the sample is qualified it can be used for subsequent analysis.

By MultiQC, QC results for all samples were compared together. Among all the QC results, the more important indicators include the duplication rate and the mean quality score. Among all the samples, 409 samples had a duplication rate below 15%, and 7 samples had a

duplication rate between 15% and 20% (Figure 2-6). As can be seen from Figure 2-6, the duplication rates of all samples are concentrated in two ranges, one is around 6%, the other is around 11%. Samples in the two ranges are from different batches. As shown in Figure 2-7, all lcWGS samples came from 5 batches, and the distribution of duplication rates of each batch was different. The distribution of duplication rates of the samples on August 15, 2018, April 5, 2019 and June 26, 2020 was relatively wide. The duplication rates vary from 6-15%, 6-11%, 1.5% to 18% respectively. The distribution of duplication rates of the samples on November 29, 2018 and July 23, 2019 is relatively small, ranging from 5-7.5% and 3-5% respectively. Therefore, the batch effect may be the main reason for the bimodal distribution of all samples. Overall, the duplication rate of these samples was within the expect.

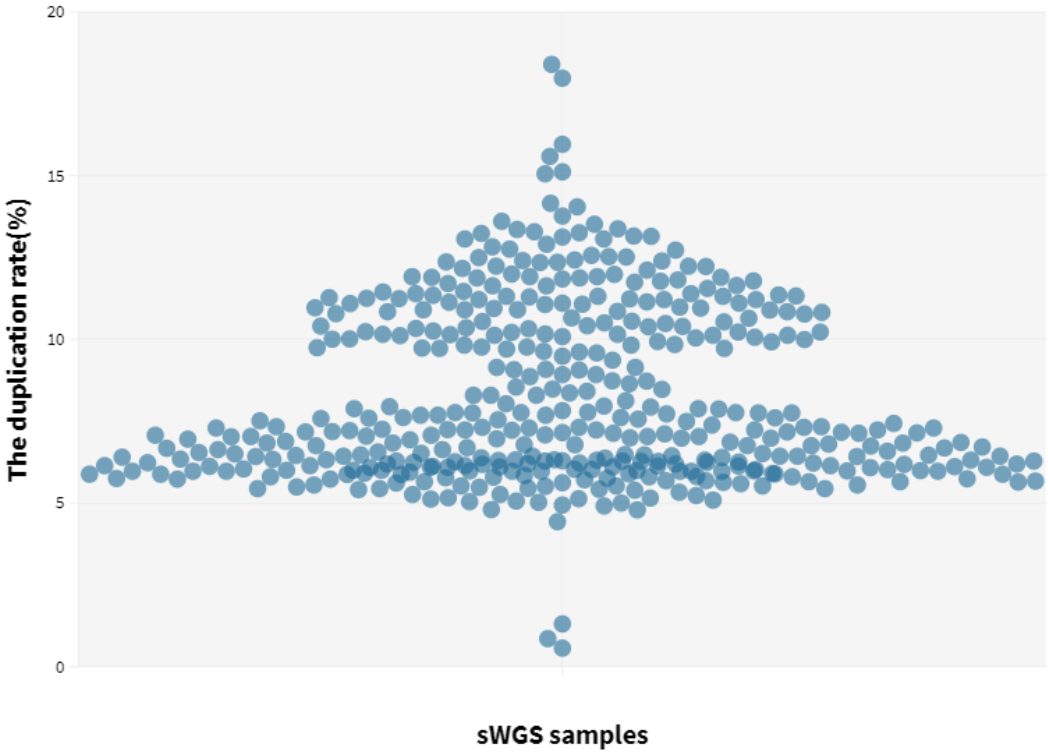


Figure 2-6 The duplication rate distribution of all 416 samples. *Each dot represents one sample, and the duplication rate of the samples is concentrated between 5%-7% and 10%-12%. Seven samples had a duplication rate of more than 15 percent.*

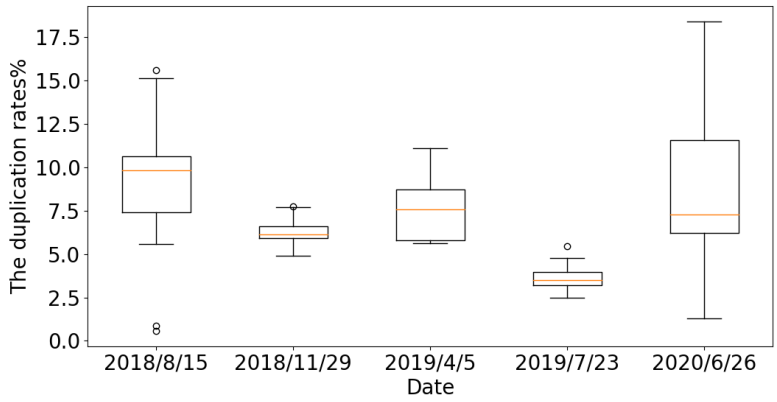


Figure 2-7 The duplication rate distribution of different batch. *The X-axis represents the batch date, and the Y-axis represents the distribution of duplication rate.*

The mean quality score is shown in Figure2-8. On most second-generation sequencing platforms, the quality of sequencing will gradually decrease with the length of sequencing. It can be seen from the figure that the mean quality score of most lcWGS samples was in line with expectations. However, there were still three samples (the red and yellow line) with low mean quality score. The red line was K34R-4UWF2Y_tumor2-b4 with a quality of 33.7. The two yellow lines were K34R-S77QY2_tumor1-b2 with a quality of 35.3, and K34R-QBKEXL_tumor1-b5 with a quality of 35.7. These three samples were still included in subsequent analyses, but they were recorded as low-quality samples.

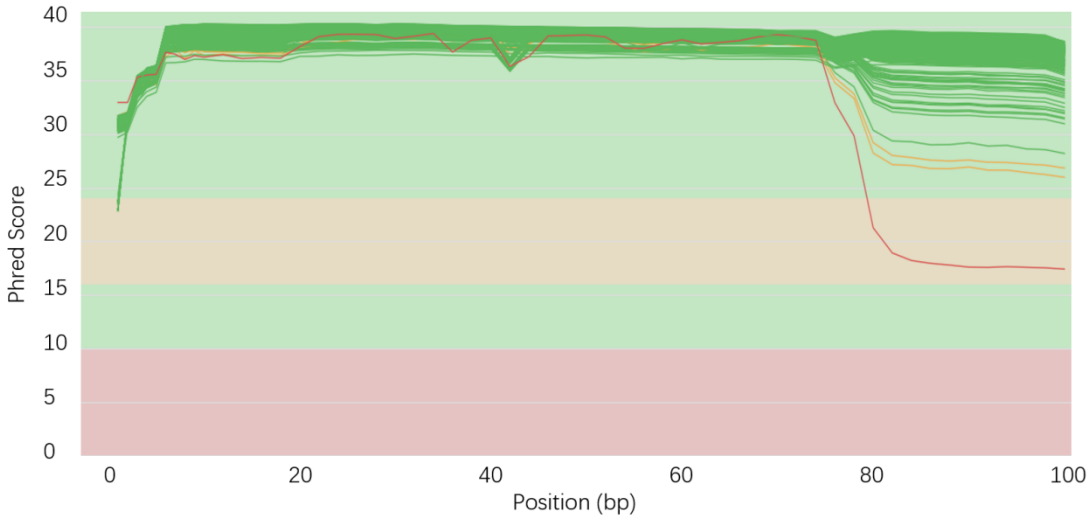
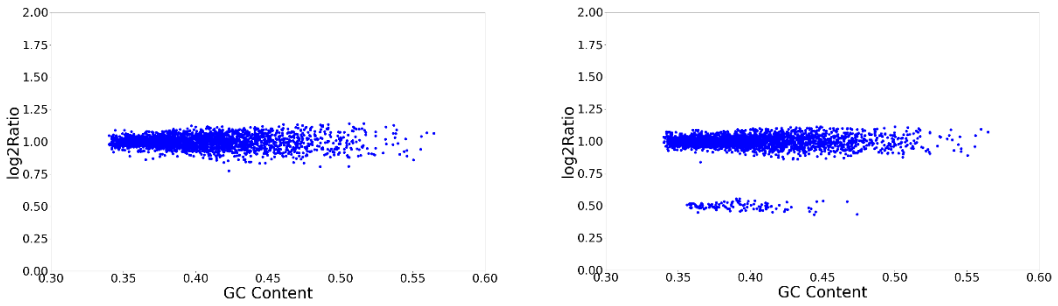


Figure 2-8 The mean quality score of 416 samples. *Among them, the x-axis represents the*

position of the reads. The y-axis represents the quality score. The higher the score, the higher the quality of base detection. The green background indicates very good quality, the orange background indicates reasonable quality, and red indicates poor quality.

2.3.2.2 GC bias removal

As described in section 2.2.3.2, ichorCNA provides the function of GC bias removal. I modified this function to output the read number after GC bias removal in each bin of the sample. After that, I took GC content as the x-axis and log2 ratio as the y-axis to draw a scatter plot. The scatter plot of the normal sample is shown in Figure 2-9, where the female sample has one horizontal line (Figure 2-9a) and the male sample has two horizontal lines (Figure 2-9b). The reason for the two straight lines in the male sample is that the male sex chromosome contains one X and one Y, and the number of reads is half of the other chromosomes.

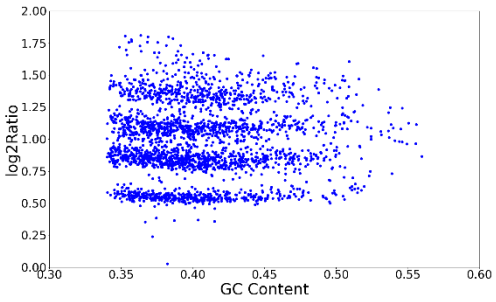


(a) (b)

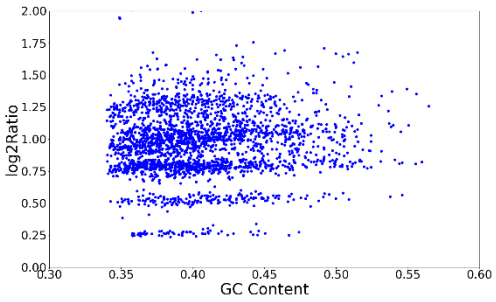
Figure 2-9 The relationship between GC content and reads number after GC correction. *X-axis represents GC content, y-axis represents the normalized reads number per bin. (a) is a female sample (K34R-1DQJQJ_tumor1-b2) while (b) is a male sample (K34R-6GMRNB_tumor1-b5).*

Among the total of 416 samples, there were six samples exhibiting distinct images (Figure 2-10). Multiple lines appear in (a) and (b), indicating that there were different copy numbers in the samples. In the figure (c) to (f), indicating that the reads number of these sample bins with

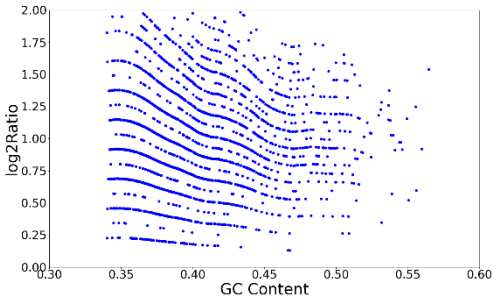
similar GC content were very different, which probably because of quality problems. This phenomenon may be attributed to inadequate coverage or DNA degradation in the samples. When a sample has insufficient coverage, the read count becomes more volatile. By checking the coverage of (c) to (f), it was found that they all had extremely low sequencing depth, respectively 0.0011, 0.000893, 0.000299, and 0.001736. The four samples were labeled as low quality. However, due to the intention of evaluating the performance of the detection tools on liquid biopsy samples with exceptionally low coverage, these samples were still included in subsequent analyses.



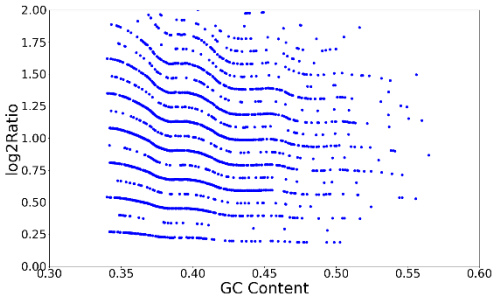
(a) K34R-DHKDDL_tumor1-b1



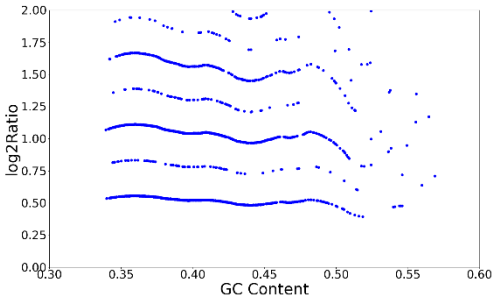
(b) K34R-QJRC8U_tumor1-b1



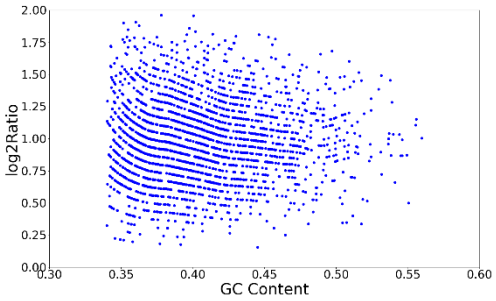
(c) K34R-4UWF2Y_tumor1-b2



(d) K34R-PKN8UU_tumor1-b2



(e) K34R-PKN8UU_tumor1-b8



(f) K34R-4UWF2Y_tumor2-b4

Figure 2-10 ***The six samples with low sequencing quality.*** *After GC bias correction, their shape is not linear, which means that their sequencing quality is poor, and their CNV detection results are untrustworthy. Samples (a) and (b) indicate different copy numbers, samples (c) to (f) exhibit significantly low coverage, lower than 0.002. These factors contribute to the variations observed in the scatter plot.*

2.3.2.3 Genomic Fingerprints

The genomic fingerprinting was used to test whether two samples are from the same patient. After construction of the heatmap of Genomic Fingerprints, all samples from the same patient can be clustered to check whether the samples correspond to the patient ID correctly. After this process, 4 sample swaps were discovered. One sample swap was shown in Figure 2-11. The figure shows the pairwise correlation between samples of two patients (K34R-1DQJQJ and K34R-DBD7LR), including both lcWGS data and panel data at all sampling times. There were suspected swaps of two samples (K34R-1DQJQJ_tumor1-b3 and K34R-DBD7LR_tumor1-b6) in the figure. By verifying the experimental records, the 4 incorrectly labeled samples were corrected.

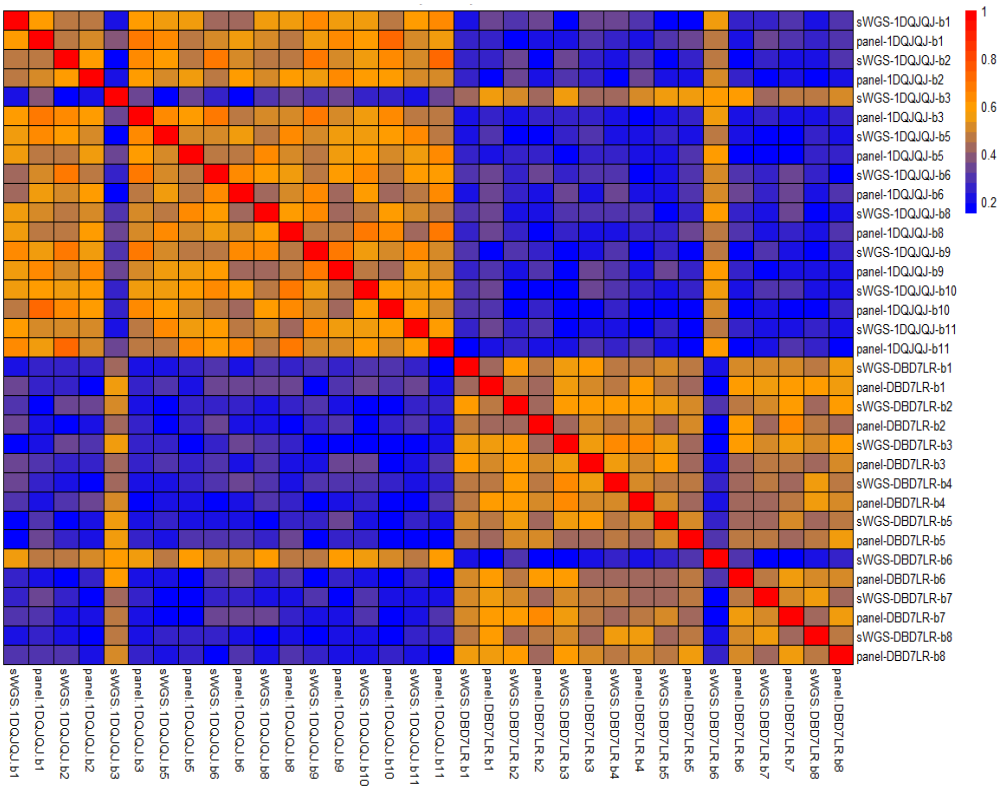


Figure 2-11 The correlation coefficients between K34R-1DQJQJ and K34R-DBD7LR. Samples from the same patient at different times have a higher correlation coefficient. Samples from the same patient are grouped in a square matrix. The correlation coefficients between different samples are low, showing blue.

2.3.3 Tumor DNA fraction

The tumor DNA fraction of lcWGS samples was estimated using ichorCNA, and the tumor DNA fraction of panel samples was estimated by mutation-based method.

The tumor DNA fraction of each lcWGS sample was obtained by selecting 1 Mb as bin size using default parameter by ichorCNA. The tumor DNA fraction of 340 samples was 0, and the mean and median of the remaining 76 samples were 0.108 and 0.083. Among them, 49 samples had tumor DNA fraction less than 0.1, and 27 samples had tumor DNA fraction

greater than 0.1. The distribution of the tumor DNA fraction of these samples was shown in Figure 2-12.

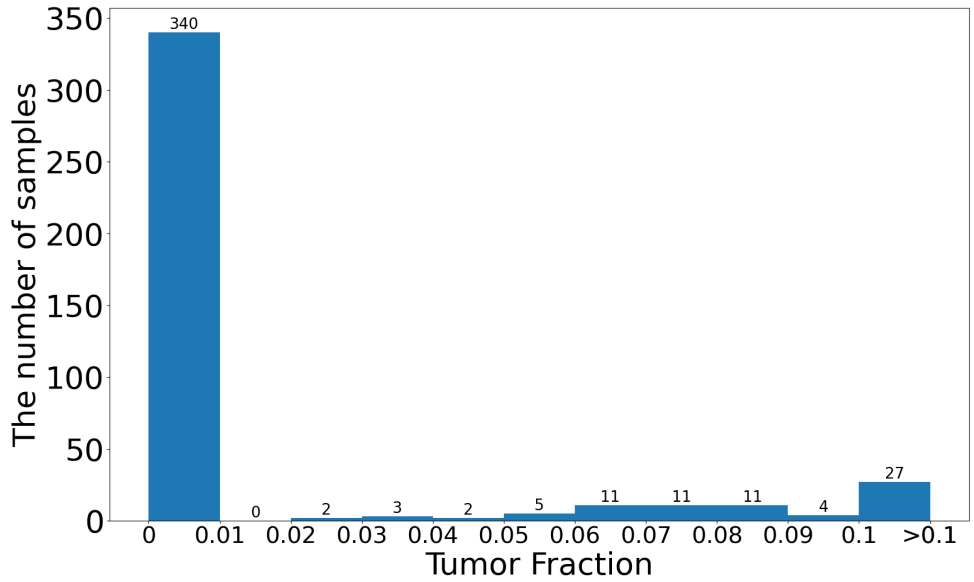


Figure 2-12: Tumor DNA fraction of all samples detected by ichorCNA with default parameters.

The tumor DNA fraction of 389 samples was less than 0.1.

On the other hand, the mutation-based method was used to calculate the tumor DNA fraction of the panel sequencing samples. The distribution of the tumor DNA fraction of the panel samples is shown in Figure 2-13. No mutaiton was found in 193 samples. Of the remaining 209 samples, the mean and median tumor fraction were 0.0245 and 0.0077. For 121 samples, the tumor fraction was between 0 and 0.01. Additionally, 79 samples had tumor fraction less than 0.1, and 9 samples had tumor fraction greater than 0.1.

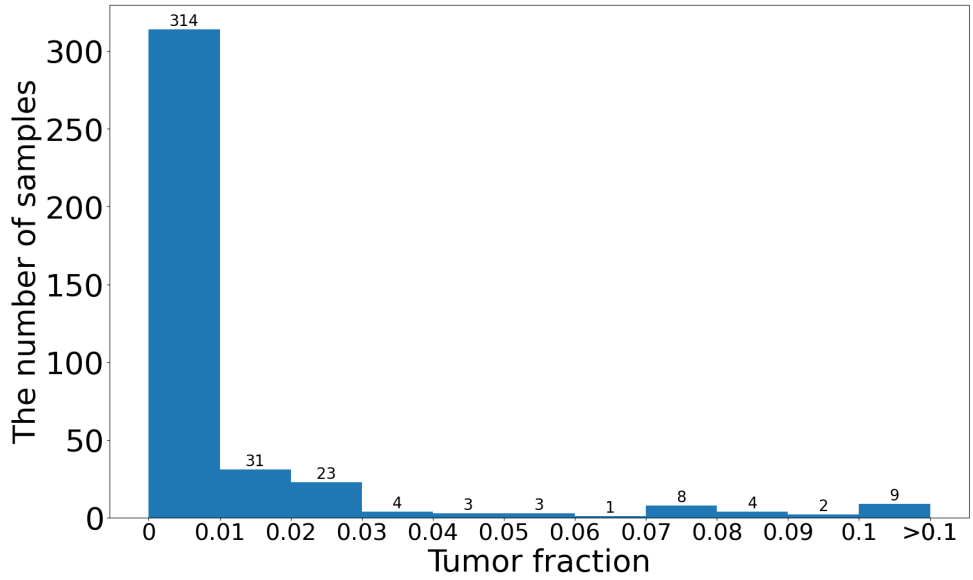


Figure 2-13 The tumor DNA fraction of all panel samples. *The tumor DNA fraction of 314 samples was less than 0.1.*

2.3.4 The consistency of the results of the two methods

It is necessary to analyze the consistency of the results to test the two methods' performance. Germline variants cannot be used to estimate tumor DNA fraction, so I excluded sites with MAF value above 0.4. The consistency of the two methods is shown in Figure 2-14. Each point represents a sample. The X-axis value is the tumor DNA fraction estimate based on MAF and the Y-axis value is the tumor DNA fraction estimate CNV method (ichorCNA). 142 samples had a tumor DNA fraction of 0 by both methods. For 13 samples, the tumor DNA fraction detected by MAF based method was 0, but by ichorCNA was not 0. For 186 samples, the tumor DNA fraction detected by ichorCNA was 0, but by MAF based method was not 0. The consistency was high only in a small number of samples, and a larger number of samples were detected with tumor in only one result.

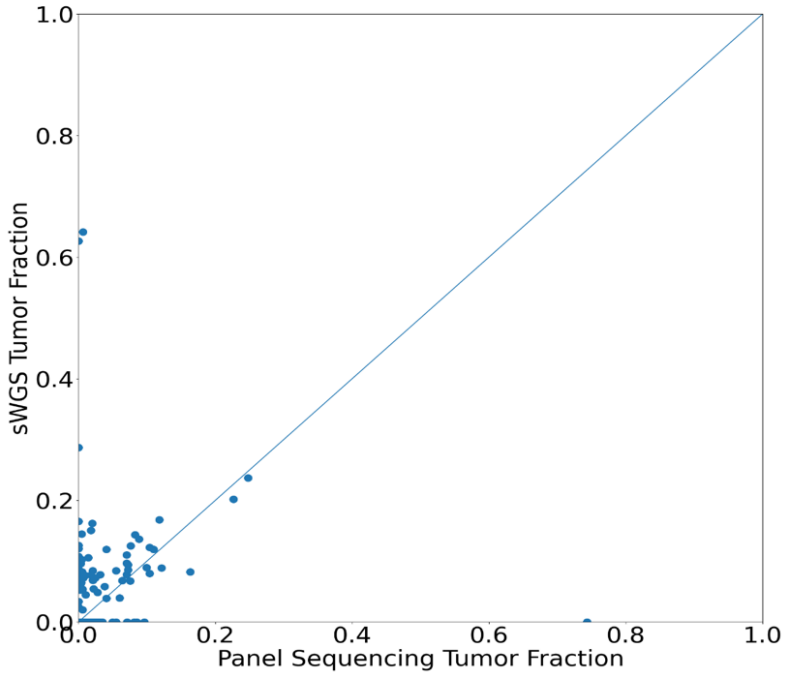


Figure 2-14 The consistency of the results of the two methods. For most samples, the two methods show highly inconsistent results. Each point represents a sample. The X-axis value is the tumor DNA fraction estimate based on MAF and the Y-axis value is the tumor DNA fraction predicted based on CNV method (ichorCNA). The Spearman Correlation is 0.31 and the P-value is 1.69e-10.

As for the samples with low quality, sample K34R-4UWF2Y_tumor2-b4 did not have a matched panel sequencing data. The tumor DNA fractions for samples K34R-S77QY2_tumor1-b2 and K34R-QBKEXL_tumor1-b5, as predicted by both sWGS and MAF value methods, were 0. The tumor fractions predicted by CNV-based method and MAF-based method for the samples with extreme low coverage (K34R-4UWF2Y_tumor1-b2, K34R-PKN8UU_tumor1-b2, and K34R-PKN8UU_tumor1-b8) were 0.0776 and 0.0350109, 0.07702 and 0, and 0.2869 and 0, respectively, showing poor consistency.

2.4 Discussion

2.4.1 The selection of MAF value

For the method based on MAF values, the MAF values of mutation sites may be affected by CNV and subclones, thus overestimating or underestimating the tumor DNA fraction. In theory, sites in the copy number neutral region should be selected to estimate tumor DNA fraction, and the number of sites should not be too small[99]. To eliminate the effect of subclonality, enough sites should also be used to plot the histogram to find the highest frequency MAF value[118]. If only a few variations are observed, it is difficult to find a reliable MAF value. The panel used in this project included only 17 genes which may contain not enough variants. A total of 432 variants were reported in 402 panel sequencing samples, with an average of 1.07 variants per sample. This makes MAF based estimation challenging for our research. In addition, sites suspected of germline variation (MAF values around 0.5 or 1) should also be excluded. Therefore, in this project, after excluding germline variations, I selected the highest MAF to carry out estimation of tumor DNA fraction.

2.4.2 The inconsistency between the two estimation methods

As shown in Figure 2-14, the results of some samples were highly consistent. However, a large number of samples showed a tumor DNA fraction of 0 by ichorCNA and a none 0 tumor DNA fraction by the MAF-based method. In another subset of samples, the situation was exactly the opposite. The reason for the inconsistency between the two methods may be that they have different judgment basis.

When the sample doesn't contain enough CNVs, ichorCNA could not accurately determine the tumor DNA fraction. This was also mentioned by Polski A et al in their research using

ichorCNA to estimate tumor DNA fraction of retinoblastoma patients[119]. It is worth noting that in addition to the optimal solution given by ichorCNA, for some samples they manually selected better solutions from other solutions given by ichorCNA. The selection of solutions affects the judgment of tumor DNA fraction. In addition, the operating parameters of ichorCNA also have an impact on the tumor DNA fraction. This will be discussed in detail in the next chapter.

When the MAF-based method is used to estimate tumor DNA fraction, the number of mutations and the change of copy number at these mutations could greatly affect the results. Newman A. M. et al. performed CAPP-Seq on NSCLC patients and confirmed that CAPP-Seq could achieve a reliable assessment of tumor burden[97]. Remarkably, the median number of mutations in their sample was 4, while in our samples the average number was 1. For samples without mutations detected, their tumor DNA fraction was considered as 0 based on the MAF-based method. However, it should be noted that since the patients in this project received TKIs treatment, a lower tumor DNA fraction may affect the detection of mutations. Researchers suggested that a higher sequencing depth and a lower detection limit can be utilized for samples taken after treatment to enhance the accuracy of detection[120].

In general, both approaches have their own limitations. The combination of the two methods should be adopted, and the parameters of CNV detection tools should be optimized to improve the accuracy of tumor DNA fraction estimation.

3. Benchmark of CNV detection tools using simulated cfDNA data

The CNV-based method plays an important role in tumor DNA fraction estimation. In addition, CNV detection itself has important implications for the diagnosis and treatment of cancer. Therefore, different CNV detection tools need to be benchmarked. The parameters of the detection tools can affect the results, so in this chapter, the parameters of ichorCNA were adjusted and ACE was modified. In addition, ground truth is very important for benchmarking the tools, but this ground truth for actual samples is not available. So simulated data with known tumor DNA fractions were used in this chapter to benchmark the performance of WisecondorX[121], modified ACE[100,] and ichorCNA with two different sets of parameters.

3.1 Introduction

In general, when performing CNV analysis on tissue samples, the corresponding blood samples can be used as references to improve the accuracy of CNV analysis. Moreover, tissue samples have a higher tumor DNA fraction than cfDNA and are usually sequenced with higher coverage. The tumor DNA fraction and CNVs can be determined by using CNV detection tools such as ACEseq[122]. But for cfDNA samples, the analysis of tumor DNA fraction and CNV is difficult. To benchmark the tools, simulated data with known tumor DNA fractions and CNVs should be generated. To ensure the accuracy of the benchmark, the coverage of the simulation data should be set to the average coverage (0.5X) of cfDNA samples in the K34R project.

In this chapter, the ability of ichorCNA, ACE, and WisecondorX to detect CNVs was evaluated using simulated lcWGS data. ACE is a tool for absolute copy number estimation. It provides alternative solutions in addition to the optimal predicted solution selected by the tool. The dynamic data visualization function allows user to visually check the absolute copy number estimated and select the best fit. Unlike the currently commonly used tools that estimate absolute copy number based on SNP array[123], WES[124], or high coverage WGS data[125], ACE does not need matched normal samples and can be applied to lcWGS. Segmentation data can be obtained from lcWGS through the QDNAseq pipeline, which integrates mapping, mapping correction, GC content analysis[126], and segmentation. The function of genome segmentation is realized by integrating DNAcopy[127]. WisecondorX was designed to detect copy number variation in lcWGS data. It divides the genome into bins of equal size and calculates the reads number in each bin. After adjusting GC bias and mappability issues, the Hidden Markov model was used to estimate the copy number status of each bin. Finally, CNVs are detected by considering the copy number status of adjacent bins, and statistical tests are used to filter out false positives. The predecessor of WisecondorX is WISECONDOR. Compared with WISECONDOR[128], WisecondorX can better deal with low coverage areas and improve the sensitivity and specificity of CNV detection by adjusting HMM parameters and adding statistical tests.

3.2 Methods

3.2.1 Preparation of simulated data for benchmark

To benchmark the performance of three tools, WGS data which has low coverage (around 0.5X) and low tumor DNA fraction (<0.1) was necessary. Therefore, I generated simulated data which has a known tumor DNA fraction and copy number profile (Figure 3-1). These simulated datasets were generated from 5 paired (tumor and control) samples (PID: 4117030X, 4139483X, 4122063, 4170577, and 4144633) with high coverage (30X) bulk sequencing data

and higher tumor DNA fraction (average 0.62) from the ICGC MMMLseq project (Molecular Mechanisms in Malignant Lymphoma). The tumor DNA fraction and CNVs of the original samples were detected by ACEseq. The tumor DNA fractions of the simulated data were from 0.5% to 10% with steps of 0.5%.

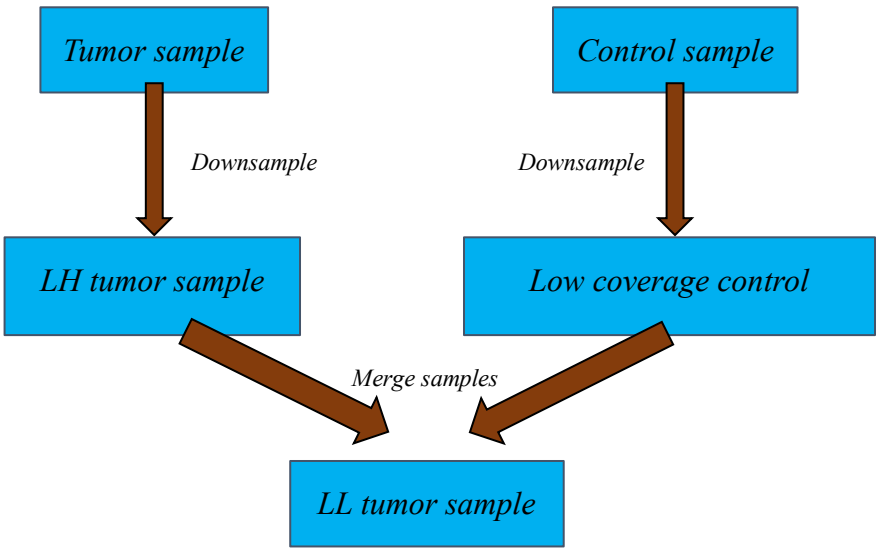


Figure 3-1 The workflow of generating simulated data. *The original paired WGS samples (tumor and control) are sequenced to around 30X coverage. After downsampling, resulting LH (low coverage and high tumor DNA fraction) tumor samples were merged with low coverage control samples to generate LL (low coverage and low tumor DNA fraction) tumor samples.*

To generate samples with each specific tumor DNA fraction, the required read counts from tumor and control samples were computed and the samples were downsampled accordingly. The tumor and control samples were downsampled into LH (low coverage and high tumor DNA fraction) samples and LL (low coverage and low tumor DNA fraction) samples. I used GATK (Genome Analysis Toolkit, version 4.0.9.0) which invoked the DownsampleSam tool in Picard (version 1.125). To ensure that coverage of the generated LL tumor sample was around 0.5X, I checked the total reads number (Rtot) of the corresponding 0.5X sample and determined that Rtot was 17,900,000.

The probability of keeping any individual read (Pt) of the required LH tumor sample was calculated by the following formula, in which tf represents the tumor DNA fraction to be

reached by the new sample, and Rt represents the total reads number of the original tumor sample, tp represents the tumor purity of the original tumor sample.

$$Pt = (tf * Rtot)/(Rt * tp)$$

The probability of keeping any individual read (Pc) of the low coverage control sample was calculated by the following formula, in which Rc represents the total reads number of the control sample corresponding to the original tumor sample.

$$Pc = (Rtot - Rt * Pt)/Rc$$

Finally, subsamples from tumor and control were merged. To produce more random samples to see the stability of the assessment, the procedure was repeated 100 times for each tumor DNA fraction.

3.2.2 Tools adjustment

To find the most suitable parameters, I performed parameter optimization for ichorCNA (OPT parameters). IchorCNA uses EM step to find the optimal solution, that is, it looks for the local optimal solution from each initial normal contamination and looks for the global optima among these local optima. Because the content of ctDNA is low, setting higher initial normal contamination and a higher number of initial normal contaminations is more likely to find the optimal solution. I changed the initial normal contamination from the default 0.5, 0.6, 0.7, 0.8 and 0.9 to 0.8, 0.83, 0.86, 0.89, 0.90, 0.93, 0.96 and 0.99. The initial ploidy was changed from 2, 3, 4, and 5 to 1.5-4.0 with steps of 0.1. Because ACE iterated the purity from 5% to 100% with steps of 1%, it only reports a purity larger than 5%. I modified the code of ACE to enlarge the iteration range from 1% to 100%. WisecondorX needs a reference file to run the analysis for samples. For 5 different PIDs, their own 100 low coverage control samples were used to

create the corresponding references, respectively. Therefore, I benchmarked WisecondorX, modified ACE, and ichorCNA with default parameters and OPT parameters.

3.2.3 Tumor DNA fraction benchmark

Because WisecondorX does not report tumor DNA fraction, the performance of modified ACE, ichorCNA with default parameters and opt parameters on tumor DNA fraction prediction was compared in this section. For every PID, I performed the analysis separately. Each PID had 2000 low coverage and low tumor DNA fraction (LL) tumor samples which included 20 groups that had different tumor DNA fractions from 0.5% to 10% step by 0.5%. Each group contained 100 LL tumor samples. Theoretically, the tumor DNA fraction predicted by the tools should match the theoretical tumor DNA fraction of each group.

3.2.4 CNV event benchmark

In this section, the performance of WisecondorX, modified ACE, and ichorCNA with default parameters and opt parameters on CNV event detection was benchmarked. First, the tools split the whole genome into 1MB bins and segment the genome. For each segment, the tool predicted a copy number. If a bin's copy number was greater than 2.5, I defined it as gain event, less than 1.5 as loss event, and 1.5-2.5 as neutral event. ACEseq's copy number prediction of each PID's original high-coverage (30x) sample was used as ground truth. For each bin, if the CNV predicted by the tool was consistent with ground truth, the prediction score for this bin was 1, otherwise, it was 0. Finally, I summed up the prediction scores of all bins of the LL tumor sample and divided them by the total number of bins in the whole genome as the tool’s CNV event prediction accuracy.

3.2.5 Tumor DNA-free sample detection

ACE returns a most likely tumor DNA fraction for each sample. However, due to the limitations of the tool algorithm, for the tumor DNA-free sample, the result returned by ACE is still a tumor DNA fraction greater than 1%. Therefore, another software, ichorCNA, was needed to determine whether a sample is a tumor DNA-free sample.

To benchmark the performance of tumor DNA-free samples prediction, 500 tumor samples that come from 5 groups of simulated data (0.5%, 1.0%, 1.5%, 2.0%, and 2.5% tumor DNA fraction) and 500 tumor DNA-free samples which were downsampled from the control sample (0.5x coverage) were chosen for the benchmark. IchorCNA with default parameter and OPT parameter were used to detect the 1000 samples, and the more appropriate parameter was selected by comparing the prediction accuracy.

3.3 Results

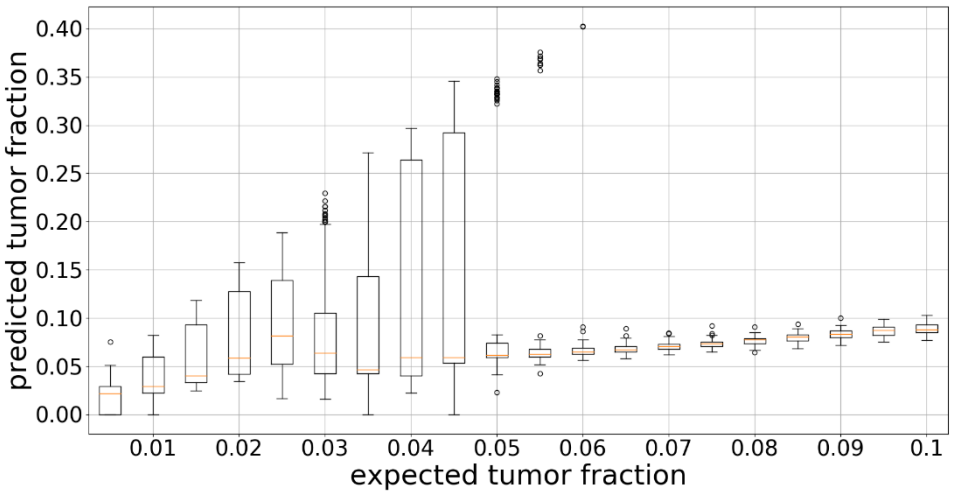
In this section, I chose a typical example (PID: 4117030X) to describe ACE and ichorCNA with default and OPT parameters performance. Since WisecondorX does not report tumor DNA fraction it was not included in the tumor DNA fraction benchmark section.

3.3.1 Tumor DNA fraction benchmark

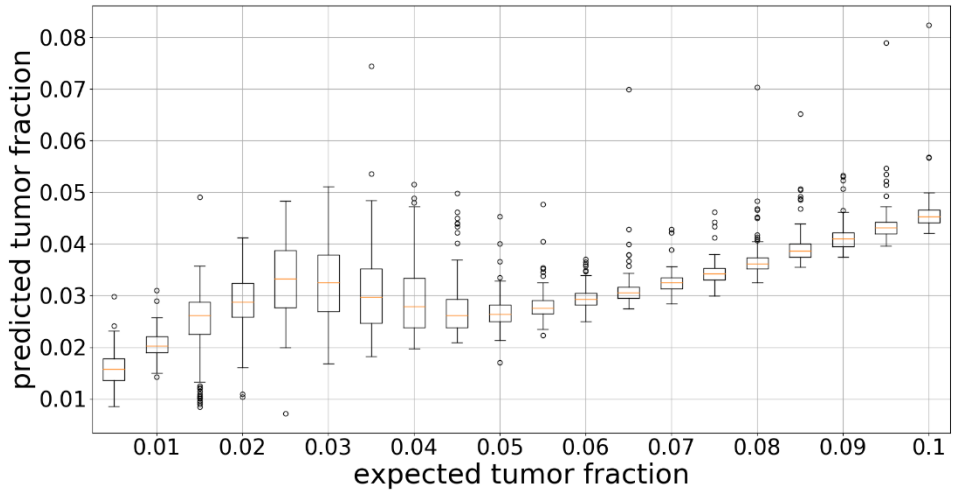
This section shows the performance of ACE and ichorCNA with default and OPT parameters to estimate tumor DNA fraction. In Figure 3-2, the X axis shows the 20 groups which have different tumor DNA fractions from 0.5% to 10% step by 0.5% and the Y axis shows the predicted tumor DNA fractions by the respective tool. In the performance plot for ichorCNA

with default parameters (Figure 3-2a), the prediction results of samples with expected tumor DNA fractions less than 5% were particularly unstable. There were many outliers with too high tumor DNA fractions (>20%). However, when the expected tumor DNA fraction was greater than 5%, the results became stable and fit a linear regression model with the expected tumor DNA fraction (R^2=0.96). However, the absolute predicted tumor DNA fractions were lower than the expected tumor DNA fractions.

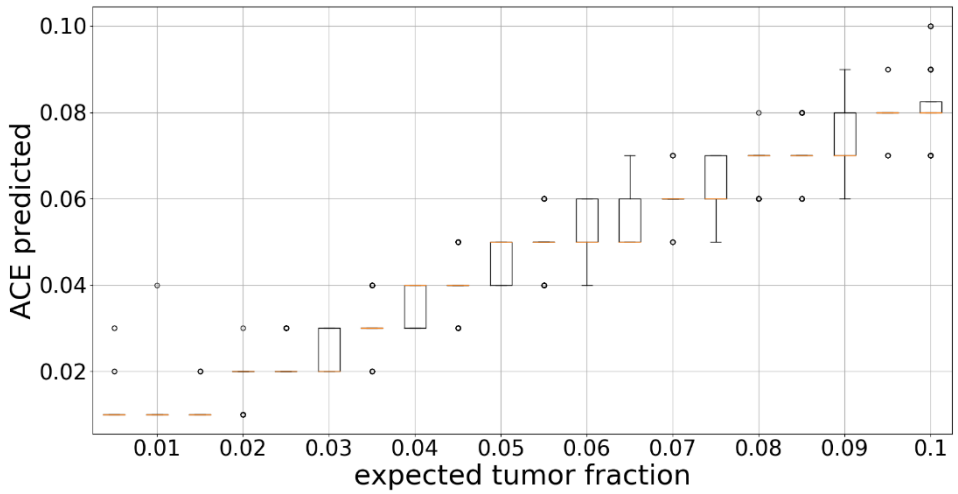
In the performance plot of ichorCNA with OPT parameters (Figure 3-2b), the prediction results of samples with expected tumor DNA fractions less than 3% were more stable and accurate than in Figure 3-2a. However, when the expected tumor DNA fraction was greater than 3%, the results were far away from the expectation. In contrast, outlier values (single points) were much closer to the expectation. Compared to ichorCNA, ACE's tumor DNA fraction prediction was accurate and stable. The only disadvantage was that the prediction accuracy of ACE can only reach 0.01 (Figure 3-2c).



(a)



(b)



(c)

Figure 3-2 Performance of PID 4117030X tumor DNA fraction prediction by tools. (a) Performance of default parameter by ichorCNA. (b) Performance of OPT parameters by ichorCNA. (c) Performance of modified ACE

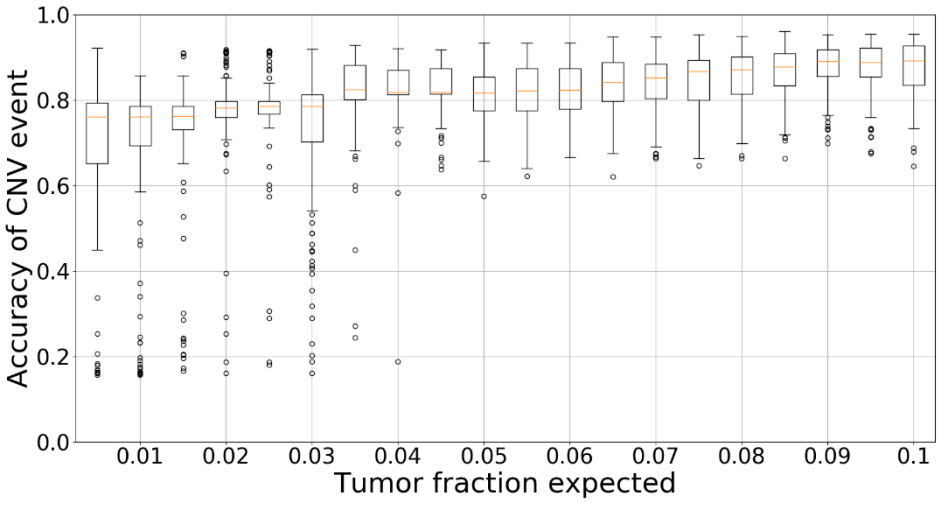
3.3.2 CNV event benchmark

This section shows the performance of ACE, ichorCNA with default and OPT parameters, and

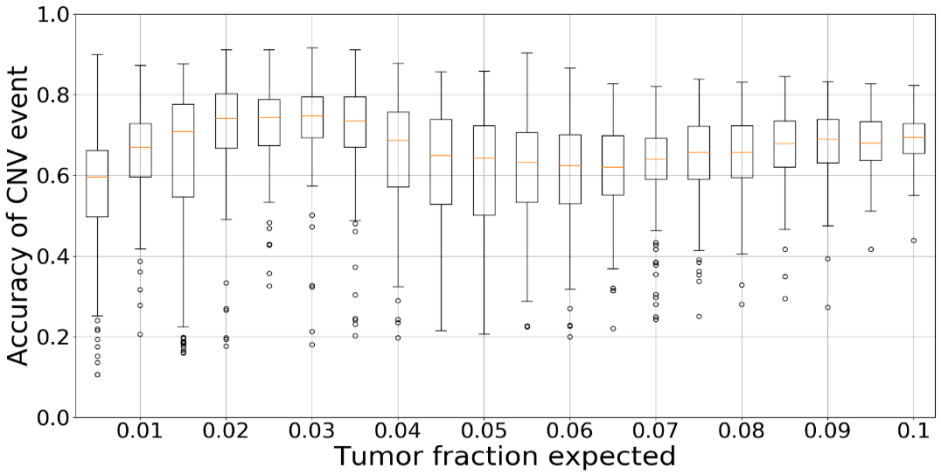
WisecondorX to detect CNV events. In Figure 3-3, the X axis shows the 20 groups which have different tumor DNA fractions from 0.5% to 10% step by 0.5% and the Y axis shows the accuracy of predicted CNV event by the respective tool. In the performance plot for ichorCNA with default parameters (Figure 3-3a), the accuracy of prediction results with expected tumor DNA fractions less than 4% was particularly unstable. They included many outliers with too low accuracy (<50%). However, when the expected tumor DNA fraction was greater than 4%, the accuracy became stable and higher.

In the performance plot of ichorCNA with OPT parameters (Figure 3-3b), the prediction results had an overall low accuracy. They included many outliers with too low accuracy (<50%).

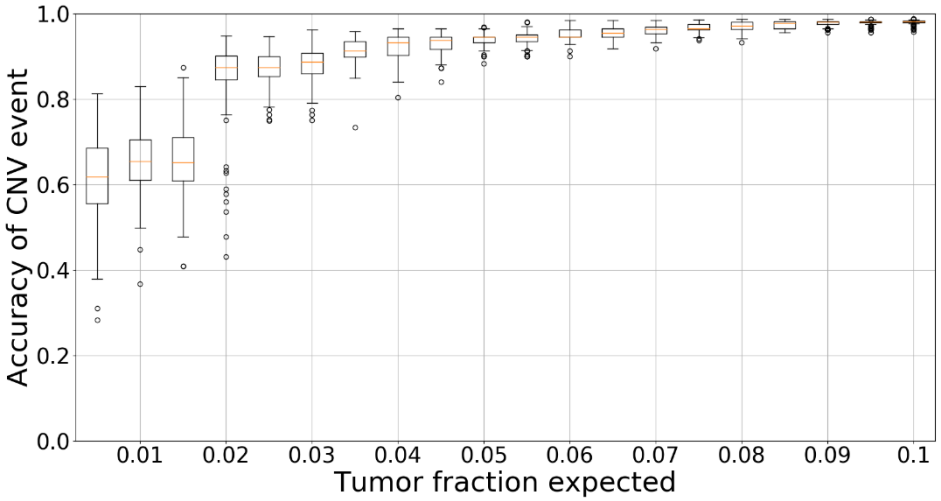
Compared to ichorCNA, ACE and WisecondorX had better performance on CNV event prediction. In the performance plot for ACE (Figure 3-3c), the prediction results with expected tumor DNA fractions larger than 3% had fewer outliers and high accuracy (>90%). In the performance plot for WisecondorX (Figure 3-3d), the prediction results with expected tumor DNA fractions larger than 5% had an accuracy higher than 95%.



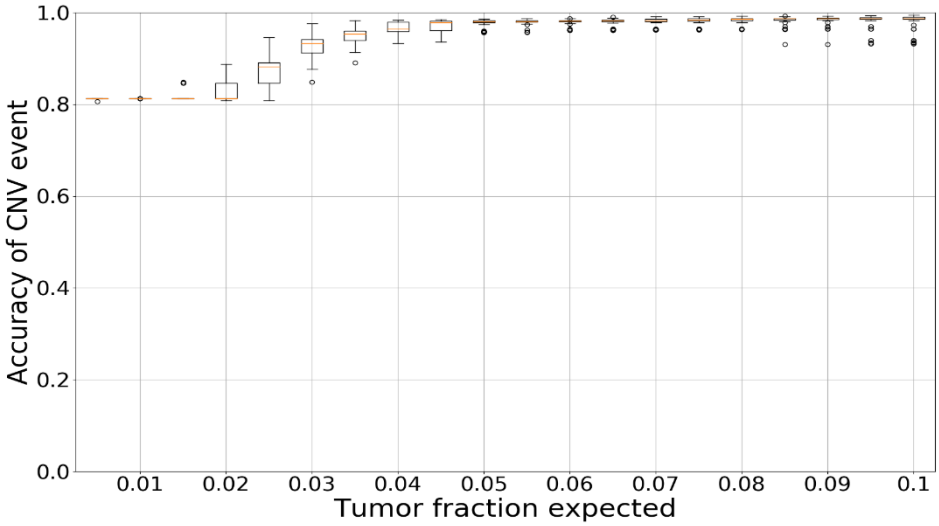
(a)


Accuracy of CNV event
1.0
0.8
0.6
0.4
0.2
0.0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
Tumor fraction expected

(b)


Accuracy of CNV event
1.0
0.8
0.6
0.4
0.2
0.0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
Tumor fraction expected

(c)



(d)

Figure 3-3 Performance of PID 4117030X CNV event prediction by tools. *(a) Performance of default parameter by ichorCNA. (b) Performance of OPT parameters by ichorCNA. (c) Performance of modified ACE. (d) Performance of WisecondorX.*

3.3.3 Tumor DNA-free sample detection

IchorCNA with default parameters and OPT parameters were used to benchmark. For 500 tumor DNA-free samples, default parameters predicted that all 500 samples do not contain tumor DNA, while OPT parameters only predicted that 63 samples do not contain tumor DNA. For 500 tumor samples, default parameters predicted 412 samples were tumor samples and OPT parameters predict 491 samples were tumor samples. (Table 3-1)

IchorCNA with default parameters had higher specificity than ichorCNA with OPT parameters but lower sensitivity. Considering specificity and sensitivity by calculating F-score, ichorCNA with default parameters were more suitable for determining whether a sample contains tumor DNA or not.

Table 3-1 Performance of PID 4117030X tumor DNA-free sample detection.

	IchorCNA with default parameters	IchorCNA with OPT parameters
Specificity	500/500	63/500
Sensitivity	412/500	491/500
Precision	1	0.529
Recall	0.824	0.982
F-score	0.904	0.688

3.4 Discussion

3.4.1 The ichorCNA parameters

The parameters of CNV detection tools have a great influence on the results. IchorCNA divides the sample into fragments with a length of 1MB, and there are about 3000 bins in the whole genome. The Expectation-Maximization (EM) algorithm is carried out to estimate the tumor DNA fraction and ploidy of the sample based on reads number of each bin. That is, ichorCNA initially uses the preset tumor DNA fraction and ploidy to calculate the discrepancy between the solution for each bin and the actual situation. Then, it attempts to fine-tune the parameters to reduce the discrepancy to find the optimal tumor DNA fraction and ploidy. Therefore, appropriate initial parameters are important to accurately predict the tumor DNA fraction and ploidy of the samples.

The default parameter of ichorCNA for initial normal contamination is from 0.5 to 0.9 which means the tumor DNA fraction is from 0.1 to 0.5. Therefore, when the tumor DNA fraction is much less than 0.1, there is no particularly close initial normal contamination and there is a higher probability to pick an outlier. From Figure 3-2a, it can be observed that as the tumor DNA fraction decreases compared to 0.1, an increasing number of outliers appears. When

the tumor fraction is below 5%, the growing number of outliers results in an enlargement of the box in the boxplot. This explains why the results are very unstable and inaccurate when the tumor DNA fraction is low. However, as the initial tumor DNA fraction approaches the real tumor DNA fraction, the percentage of outliers decreases. In this case, ichorCNA can find a more appropriate solution. The ctDNA in the blood of cancer patients is only a small fraction of the total cfDNA, usually between 0.01% to 2%[129], so initial normal contamination needs to be set to a higher value.

3.4.2 The performance of ichorCNA

The ichorCNA with default parameter had a higher performance in predicting tumor DNA fraction and detecting CNV when the tumor DNA fraction of the sample is above 5%. However, when the tumor DNA fraction was below 0.05, the estimated tumor DNA fraction fluctuated greatly, and higher outliers were observed. Adalsteinsson V A et al. also tested the performance of ichorCNA using simulated data[94]. In their research, simulated data with exact tumor DNA fraction was generated by mixing the cfDNA from cancer patients and healthy donors. According to the data provided by them, when the tumor DNA fractions of the simulated data were below 0.05, the predicted tumor DNA fractions were all below 0.1. Their data performed better when tumor DNA fractions were below 0.05, possibly because the CNV patterns of the raw data we used were different. The complexity of CNVs and the number of subclones would affect the results[130, 131]. Adalsteinsson V A et al. used cfDNA from patients with breast or prostate cancer, while I used samples of patients with malignant lymphoma to generate simulated data. In addition, the setting of parameters is a key factor affecting the performance of ichorCNA. In their study, lcWGS samples had matched WES samples with higher coverage, and the ploidy generated by ABSOLUTE/TITAN from these WES samples was used as the initial ploidy for ichorCNA. Using appropriate initial ploidy can improve the analytical performance of ichorCNA.

With the OPT parameters, ichorCNA improved its performance in estimating tumor DNA fraction below 5%, but its ability to detect CNV was not as good as the default parameter. For detecting tumor DNA free samples, ichorCNA with default parameter showed 100% specificity, which was better than ichorCNA with OPT parameter.

3.4.3 The reference for CNV detection tools

The reference plays a very important role in detecting CNVs of ctDNA. It can be found in section 3.3.2, WisecondorX had the best ability to detect CNVs. This may be because WisecondorX provides reference parameters, allowing users to customize the input reference to eliminate bias. In section 3.3.2, I used the control sample which was used to generate the simulated data as a reference. This may be the reason why WisecondorX performed best.

The concentration of ctDNA in cfDNA is low, so a high noise background can interfere with the detection of CNV[132]. The reference can eliminate bias from lab and sequencing, it can improve the accuracy of CNV detection. One difficulty in detecting ctDNA is that ctDNA is mixed with DNA from normal cells. The reference can be used to clarify the pattern of normal DNA fragments, to better identify tumor signals[133].

In section 3.3, ichorCNA used its own reference to calibrate the sequencing data. Due to sequencing instruments and batch effect, the built-in reference may have different bias from the sequencing data in our project. To address this, I'll build our own reference in the next chapter. Although ACE performs well in the detection of tumor DNA fraction and CNV, it does not use reference to correct bias. Adding a reference correction step might further improve its performance. This will also be discussed in detail in the next chapter. In short, the reference correction is very important for the CNV detection tools, and I needed to optimize it to improve the tools' performance.

4. CNV detection tools benchmark based on short fragment read samples

In this chapter, the analysis was based on the data from project HIPO-K34. Firstly, I established a PoN (Panel of Normals) dataset. Secondly, based on the PoN dataset, a reference was created for each tool. During this process, ACE was optimized to enhance its performance. Thirdly, I generated simulated data using enriched short fragments. Finally, I benchmarked the performance of tools based on the simulated data.

4.1 Introduction

Since researchers first reported the existence of cfDNA in human plasma in 1948, it has become an attractive research topic as a non-invasive disease biomarker. cfDNA can be present in serum, plasma, and other body fluids such as urine or saliva[134]. In addition to the large abundance of cfDNA, there may also be a small amount of ctDNA in the plasma of cancer patients. The presence of this ctDNA makes it possible for early screening of cancer or a more convenient and less traumatic concomitant diagnosis. However, an issue that needs to be noted is that accurate tumor information can be obtained only when the abundance of ctDNA in cfDNA is high enough[135]. Except for some advanced cancers, this abundance is not easy to reach in most patients. At present, the method to improve the sensitivity and accuracy of ctDNA detection is to increase the sequencing depth. However, increasing sequencing depth may lead to a higher false-positive rate, as DNA of non-tumor derived may also carry various tumor associated mutations[136]. This problem has limited the application of liquid biopsy. However, some research has brought new ideas to the detection of ctDNA. Previous studies have shown that the length of cfDNA released into the plasma by different cells is different. The length of cfDNA is about 167bp in general, which is similar to that of a

nucleosome, which may be related to caspase dependent DNA cleavage during apoptosis[137]. cfDNA from different cell sources will show a unique pattern in length, for example, the cfDNA of infants is significantly shorter than that of mothers[138]. In cancer patients, the fragment lengths of cfDNA from normal cells and ctDNA from tumor cells also have different distributions. Nitzan Rosenfeld et al.[139] found that the distribution of cfDNA between healthy people and cancer patients was different between 90-150bp, 180-220bp, and 250-320bp. ctDNA fragments with cancer mutations are generally 20-40bp shorter than the 167bp of nucleosome DNA fragments and are enriched in the range of 90-150bp (Figure 4-1). In addition, some ctDNA fragments are enriched in the range of 250-320bp, which is considered to be binuclear body fragments of tumor cells. At present, the reason for this biological difference has not been clearly explained[101]. However, the method of fragment length screening can significantly improve the abundance of ctDNA and thus improve the sensitivity and accuracy of detection.

As shown in the results in Chapter 3, for samples with tumor DNA fractions greater than 5%, both ichorCNA and ACE can accurately measure tumor DNA fractions. When a higher tumor DNA fraction is obtained by filtering the reads, the accuracy of tumor fraction measurement may be improved.

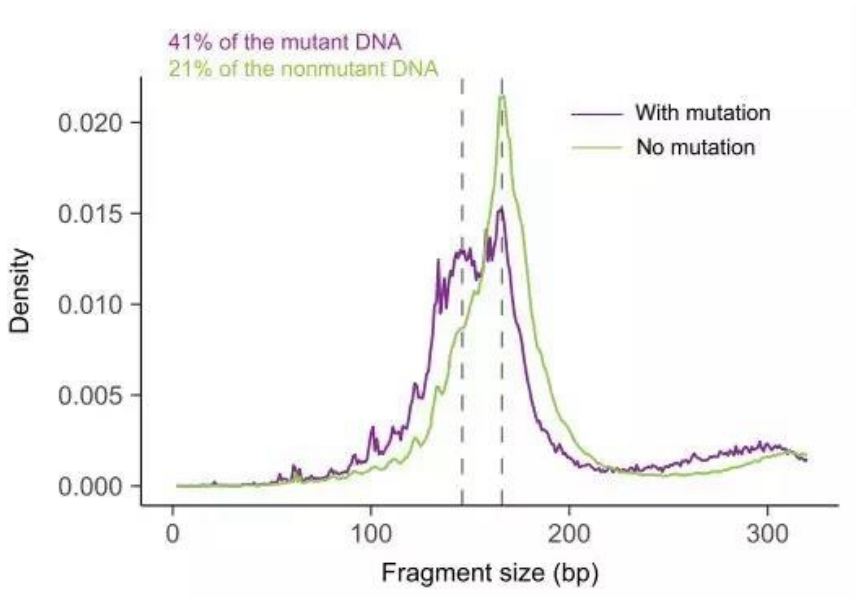


Figure 4-1 The distribution of cfDNA fragments with mutation and without mutation. *In the range*

of 90-150bp and 250-320bp, cfDNA with mutation has higher enrichment than cfDNA without mutation (from Nitzan Rosenfeld et al., 2018).

4.2 Methods

4.2.1 Establishing the PoN dataset

Using a PoN dataset as a reference can reduce systematic biases arising from library construction, sequencing platform, and cfDNA-specific artifacts. In this case, 11 samples from project HIPO-K34 which have not detected any SNV were selected to form the PoN dataset. After establishing the PoN dataset, I validated its effectiveness using the following methods:

1. Firstly, using a bin size of 1Mb, the reads number in each bin of the test samples was calculated.
2. Secondly, loess regression was applied to correct for GC bias for the reads number within each bin. To enhance comparability, the corrected reads numbers were normalized to a range of 0-1 (normalized coverage score), where 0 represents minimum coverage and 1 represents maximum coverage.
3. Thirdly, for each bin in the test samples, the normalized coverage score was divided by the corresponding normalized coverage score in the PoN dataset. The resulting ratio was then normalized again to a range of 0-1.
4. Finally, a heatmap was generated for all samples to observe if there are any specific biases present after PoN correction.

4.2.2 Reference creation and ACE modification

In this section, I created a corresponding reference for each CNV detection tool based on the

PoN dataset established in the previous section. However, since ACE did not have a built-in functionality for reference correction, I made modifications to ACE.

4.2.2.1 ichorCNA

The default reference of ichorCNA is derived from 27 healthy donors. For the three default bin sizes (10kb, 500kb, and 1MB), ichorCNA has the corresponding reference files. In addition, ichorCNA also offers users the method to create their own reference file. I followed ichorCNA's documentation to create the PoN_reference file. The process included the following steps:

1. The WIG files were created. For each sample in the PoN dataset, a WIG file was generated by the following command.

```
/path/to/HMMcopy/bin/readCounter --window 1000000 --quality 20 \
            --
chromosome "1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22" \
/path/to/tumor.bam > /path/to/tumor.wig
```

2. The 'createPanelOfNormals.R' script provided by ichorCNA was used to generate the reference file. Where '*--filelist*' was the file containing the path to all normal sample WIG files, '*--gcWig*' was the GC Wig file of the reference genome, '*--mapWig*' was the mappability Wig file of the reference genome, and '*--centromere*' was the file containing the centromere location. The GC Wig file, mappability Wig file and centromere files were provided by ichorCNA.
3. '*--normalPanel*' was used to reduce the systematic biases.

4.2.2.2 WisecondorX

WisecondorX also has the ability to generate the reference file from the PoN dataset. The details were as follows.

1. The BAM files of the PoN dataset were converted to NPZ files.

```
WisecondorX convert input.bam --binsize 1000000 output.npz
```

2. The generated NPZ files were used to create the reference file.

```
WisecondorX newref reference_input_dir/*.npz reference_output.npz –
binsize 1000000
```

4.2.2.3 ACE

The reference was merged from the PoN bams by samtools (version 1.9)[140]. After the preparation of the reference, it was placed at the beginning of the sample list for ACE analysis. ACE's built-in process will downsample all input samples to 1Gb, so that all bam files have a similar coverage. When the sample size is less than 1Gb, the downsample will not be performed.

ACE integrates functions from QDNAseq to perform tasks including obtain the reads number of each bin, removing the blacklist areas, GC bias correction, read counts normalization, and segmentation. However, ACE does not include a function to do reference correction. Therefore, prior to segmentation, I made a modification the ACE's code to implement reference correction. The modified code section was as follows:

```
for (b in binsizes) {
  currentdir <- file.path(outputdir, paste0(b, "kbp"))
  dir.create(currentdir)
  bins <- QDNAseq::getBinAnnotations(binSize = b, genome = genome)
```

```
  readCounts <- QDNAseq::binReadCounts(bins, path = inputdir)
  readCountsFiltered <- QDNAseq::applyFilters(readCounts, residual = TRUE, blacklist =
TRUE)
  readCountsFiltered <- QDNAseq::estimateCorrection(readCountsFiltered)
  copyNumbers <- QDNAseq::correctBins(readCountsFiltered)
  copyNumbers <- QDNAseq::normalizeBins(copyNumbers)
  copyNumbers <- QDNAseq::smoothOutlierBins(copyNumbers)
  ################################
  #  old code
  #  copyNumbersSegmented <- QDNAseq::segmentBins(copyNumbers,transformFun =
 "sqrt")
  ################################
  #  modification code
  tumorVsNormal<-
QDNAseq::compareToReference(copyNumbers,c(FALSE,rep(1,length(copyNumbers[[1]])
-1)))
  copyNumbersSegmented <- QDNAseq::segmentBins(tumorVsNormal,  transformFun
= "sqrt")
  #  modification end
  ################################
  copyNumbersSegmented <- QDNAseq::normalizeSegmentedBins(copyNumbersSegm
ented)
  saveRDS(copyNumbersSegmented, file = file.path(outputdir,  paste0(b, "kbp.rds")))
  ploidyplotloop_lowrange(copyNumbersSegmented, currentdir, ploidies, imagetype, me
thod, penalty, cap, bottom, trncname, printsummaries, autopick)
}
```

The original code of ACE directly uses the copy number data obtained by 'QDNAseq::smoothOutlierBins(copyNumbers)' to perform segmentation. I added 'QDNAseq: : compareToReference' to perform reference correction for tumor samples. This function takes copy number data as input, along with a vector indicating which samples need to do reference correction. In this vector, ‘False’ represents the reference, and ‘1’ represents a tumor sample. Since I placed the reference bam file at the beginning of the sample list during the preprocessing process, the first element of the vector was ‘False’. The copy number data corrected by reference was then used for segmentation.

4.2.3 Generating simulated data

To benchmark the performance of the tools, I used simulated data. The advantage of these simulated samples over real samples is that they have known tumor DNA fraction and CNV events. The tumor DNA fraction of sample K34R-XDDFED_tumor1-b3 was detected to be 10% by both ichorCNA and ACE. I filtered the reads with a short fragment size (90-150bp) in this sample and treated these reads as a new sample. The tumor DNA fraction of this new sample was 0.13 (by ACE). Using the same method in Chapter 3, I downsampled the new sample obtained in the previous step and mixed it with healthy donor samples to form a test sample set. The tumor DNA fraction of this test sample set was 0.5% to 10% with steps of 0.5%. For each tumor DNA fraction, this process was repeated 100 times.

4.2.4 Tumor DNA fraction and CNV event benchmark

The performance of modified ACE, ichorCNA with default parameters and opt parameters and WisecondorX were benchmarked in this section.

Because WisecondorX does not report tumor DNA fraction, the performance of modified ACE, ichorCNA with default parameters and opt parameters on tumor DNA fraction prediction was compared for tumor DNA fraction prediction. According to the tumor DNA fraction from 0.5% to 10%, the samples were divided into 20 groups, each group containing 100 samples. Theoretically, the tumor DNA fraction predicted by the tools should match the theoretical tumor DNA fraction of each group.

In the CNV event section, the performance of WisecondorX, modified ACE, and ichorCNA with default parameters and opt parameters on CNV event detection was benchmarked. Consistent with the method in Chapter 3, for each segment, the tool predicts a copy number.

If a bin's copy number was greater than 2.5, it was defined as gain event, less than 1.5 as loss event, and 1.5-2.5 as neutral event. ACEseq's copy number prediction of the original sample K34R-XDDFED_tumor1-b3 was used as ground truth. For each bin, if the CNV predicted by the tool was consistent with ground truth, the prediction score for this bin was 1, otherwise, it was 0. Finally, I sum the prediction scores of all bins of the LL tumor sample and divided them by the total number of bins in the whole genome as the tool's CNV event prediction accuracy.

4.3 Results

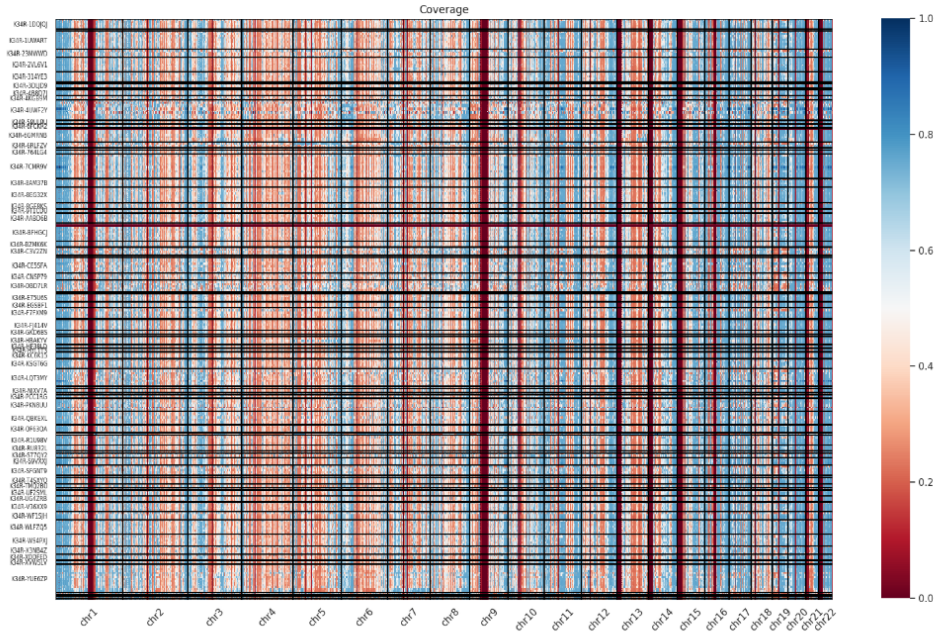
4.3.1 The ability of PoN dataset to remove bias

This section shows the original coverage distribution of all samples and the coverage distribution of all samples after removing systematic biases by the selected PoN dataset. Figure 4-2 provides an overview of the normalized reads number for each bin of all samples. The x-axis represents all bins arranged by chromosome position, and the y-axis represents all the samples. Samples from the same patients were grouped together and sorted according to sampling time. Different patients were separated by a horizontal black line. Red indicates lower reads number and blue indicates higher reads number. Figure 4-2a shows the original coverage distribution of all samples. It can be seen that almost all of the samples exhibited a similar coverage at the same x-axis positions (visible as blue or red vertical lines). This indicates that reads number bias was present at specific genomic locations. In order to remove the bias, a PoN dataset was selected. This PoN dataset were from 11 samples in this project which have not detected any SNV.

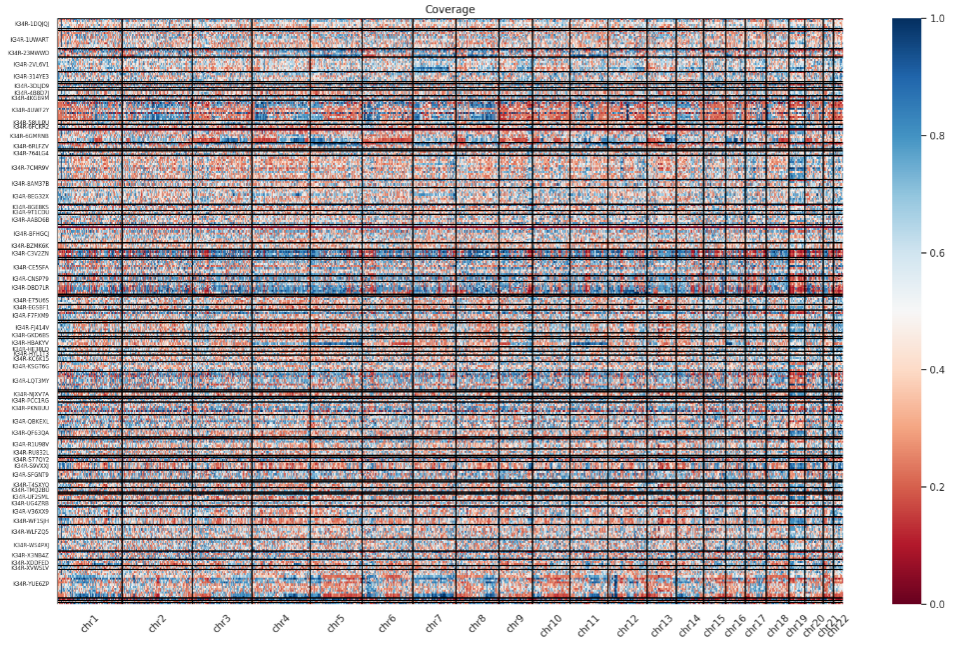
In the process of removing coverage bias using reference, the ratio of sample coverage and the average coverage of the reference was used to be the new coverage. For comparability between samples, the ratio was then normalized. As shown in Figure 4-2b, there were no

apparent vertical lines observed, indicating that the bias was effectively removed.

For IchorCNA and WisecondorX, the PoN dataset was used as a reference to eliminate bias. ACE does not have a reference correction process, so I achieved this function by modifying the ACE code, as described in the methods section.



(a)



(b)

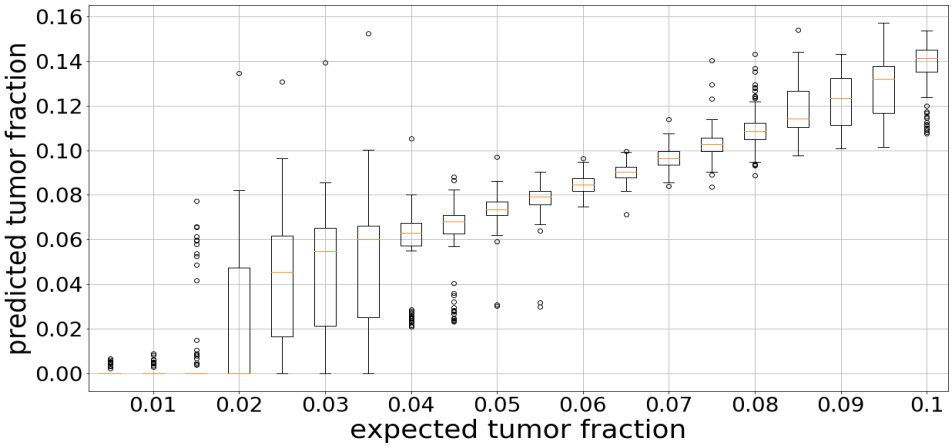
Figure 4-2 Coverage distribution plot of K34R samples. *(a) Coverage distribution before removing bias. (b) Coverage distribution after removing bias.* The x-axis represents all the bins arranged according to their chromosomal positions, while the y-axis represents all the samples. The color scale indicates the reads number of each bin, with red indicating a lower reads number and blue indicating a higher reads number. Distinct vertical lines were observed in (a), indicating that different samples had similar coverage distributions at the same genomic positions. No vertical lines were observed in (b), indicating an improvement in coverage bias.

4.3.2 Tumor DNA fraction benchmark

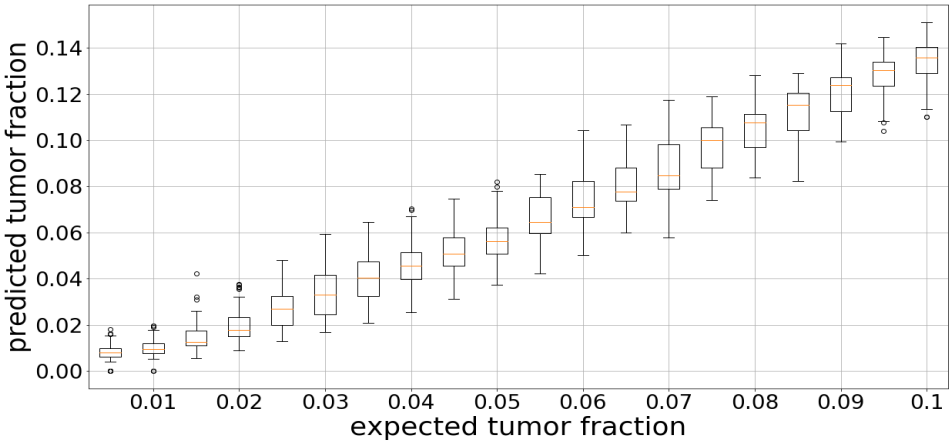
In Figure 4-3, the X axis shows the 20 groups which have different tumor DNA fractions from 0.5% to 10% step by 0.5% and the Y axis shows the predicted tumor DNA fractions by the respective tool.

In the performance plot for ichorCNA with default parameters (Figure 4-3a), it can be seen that when the expected tumor DNA fraction was lower than 5% and higher than 8%, the result was unstable and there were more outliers. The performance of ichorCNA with opt parameters was significantly better than that of ichorCNA with default parameters (Figure 4-3b). In all 20 groups ranging from 0.5% to 10%, their predicted tumor DNA fractions were relatively stable and had a linear relationship with the expected tumor DNA fractions. However, these predicted tumor DNA fractions were generally slightly larger than the corresponding expected tumor DNA fractions.

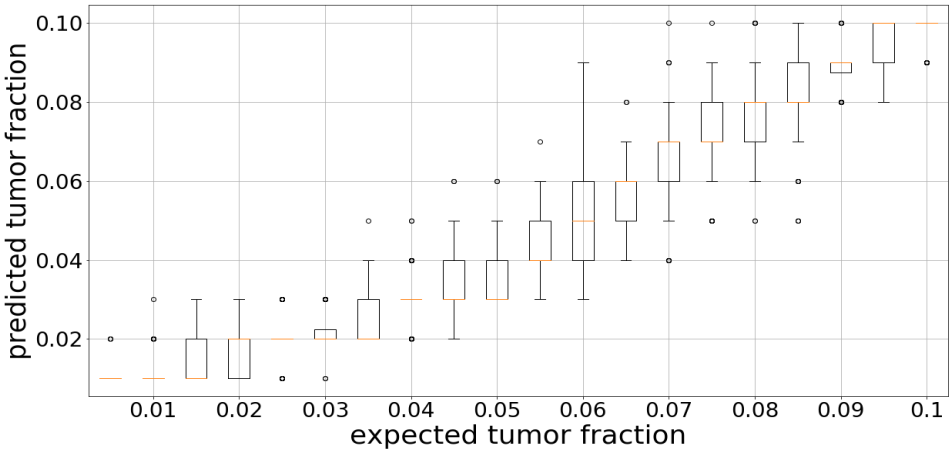
ACE's results performed best among the three tools. It can be seen from Figure 4-3c that the predicted tumor DNA fractions and the expected tumor DNA fractions had a higher consistency in the entire interval from 0.5% to 10%. And for each tumor DNA fraction, the distribution of prediction results was relatively concentrated.



(a)



(b)



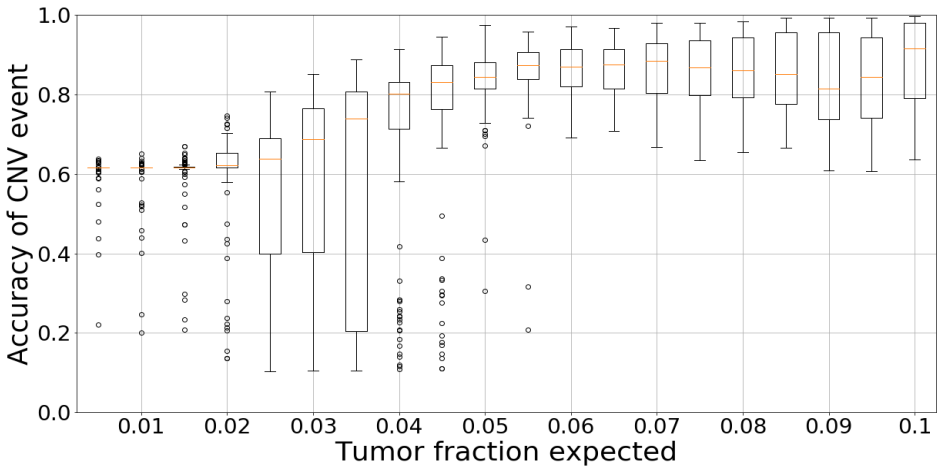
(c)

Figure 4-3 Performance of tumor DNA fraction prediction by tools. (a) Performance of default parameter by ichorCNA. (b) Performance of OPT parameters by ichorCNA. (c) Performance of modified ACE.

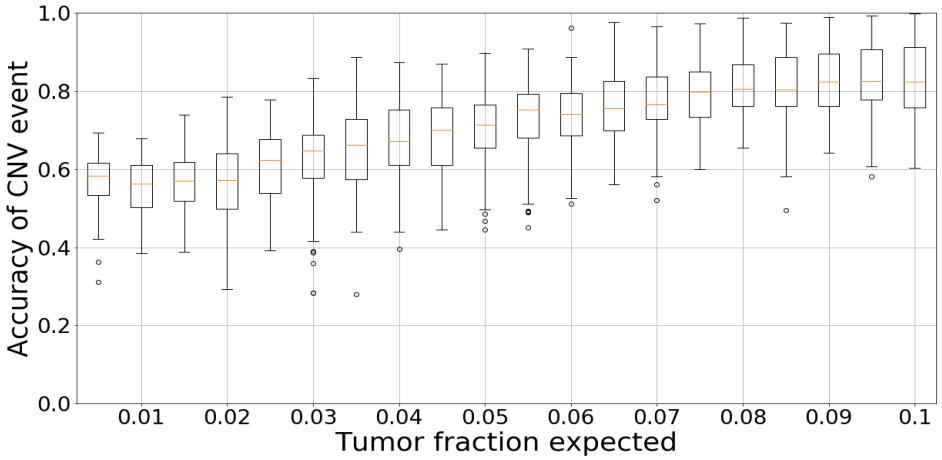
4.3.3 CNV event benchmark

In Figure 4-4, the X axis shows the 20 groups which have different tumor DNA fractions from 0.5% to 10% step by 0.5% and the Y axis shows the accuracy of predicted CNV event by the

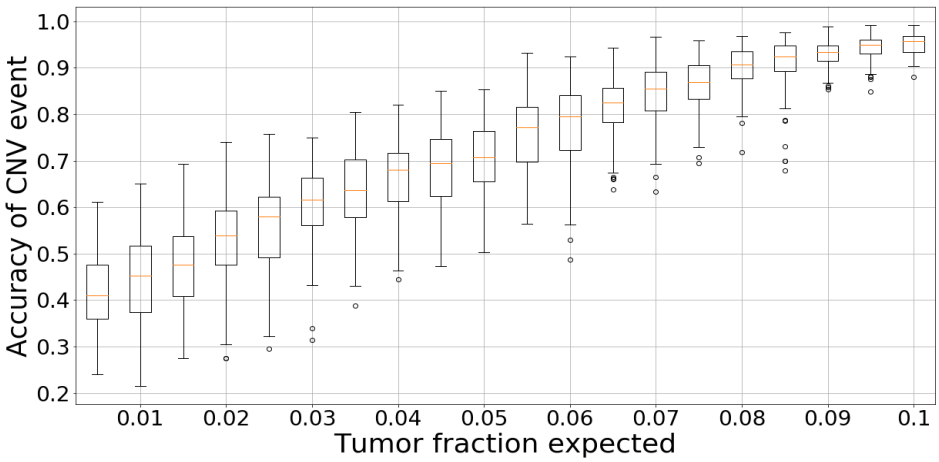
respective tool. In the performance plot for ichorCNA with default parameters (Figure 4-4a), it can be observed that when the expected tumor DNA fraction was greater than 5%, the accuracy of the CNV event was higher (>80%). When the expected tumor DNA fraction was less than 5%, the result was very unstable and contained a lot of outliers. The performance plot for ichorCNA with opt parameters (Figure 4-4b) showed relatively stable results. But only when the expected tumor DNA fraction was greater than 8%, the accuracy was higher (>80%). In The performance plot for ACE (Figure 4-4c), the accuracy of prediction exceeded 80% when the expected tumor DNA fraction was greater than 6.5%, and when the expected tumor DNA fraction was less than 6.5%, the accuracy decreased as the tumor DNA fraction decreased. The accuracy of WisecondorX for predicting CNV events did not change much with different tumor DNA fractions, and stayed always around 60% (Figure 4-4d).



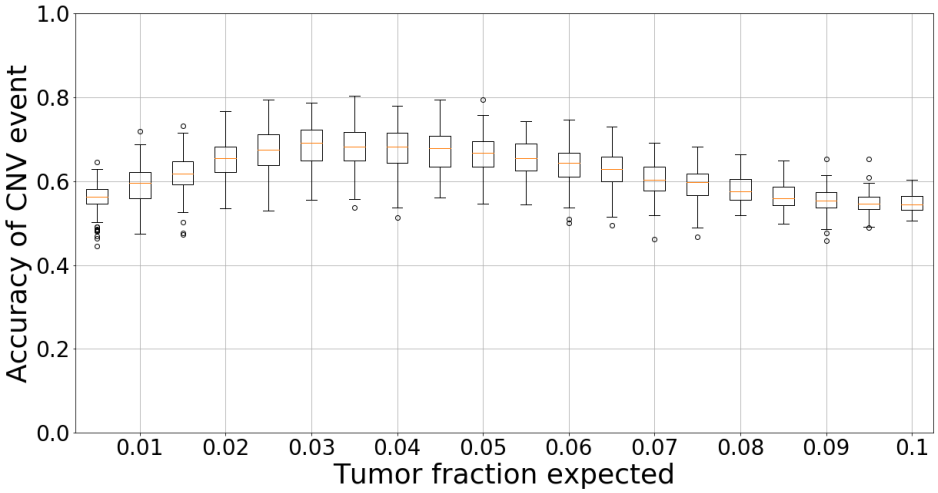
(a)



(b)



(c)



(d)

Figure 4-4 Performance of CNV event prediction by tools. *(a) Performance of default parameter by ichorCNA. (b) Performance of OPT parameters by ichorCNA. (c) Performance of modified ACE. (d) Performance of WisecondorX.*

4.4 Discussion

4.4.1 PoN dataset to reduce bias

Experimental procedures, such as PCR, library preparation, target capture, and sequencing can introduce biases into NGS data[141]. At present, the conventional method is to use the sequencing data of patients' white blood cells or the sequencing data of healthy donors as control to remove the bias[142, 143]. The PoN data I selected in this section came from 11 samples in project HIPO-K34. Because they came from the same experimental conditions as other liquid biopsy samples, it can effectively reduce bias. However, it should be noted that these 11 samples were derived from cfDNA of patients and were selected for the PoN dataset because no SNV was detected. It should be noted that the panel region represents only a small portion of the genome. While the samples were from NSCLC patients, and the genes included in the panel are common mutation genes for this type of patients, it does not imply that the mutations in these patients occur only within the panel region[144, 145]. Although Avenio reported that no SNVs were detected in these samples, the presence of ctDNA in these 11 samples cannot be ruled out. If using cfDNA from healthy donors in the same batch of experiment as PoN dataset, better results may be obtained.

4.4.2 Enrichment of short fragments in NSCLC

In this chapter, I obtained a sample with higher tumor DNA fraction (13%) by enriching fragments of 90-150bp in length and used this sample to generate simulated data of different tumor DNA fraction. The differences in fragment lengths between ctDNA and cfDNA have been demonstrated in several studies[139, 146, 147]. However, there are still aspects in this field that require further investigation and exploration. For example, in the study by Jiang P et al.[148], it was mentioned that hepatocellular carcinoma patients with lower concentrations of tumor

DNA fractions in plasma had significantly longer size distributions than healthy controls. These longer fragments may be derived from necrosis rather than apoptosis. As for lung cancer, in the study of Underhill HR et al., it was confirmed that the length of cfDNA fragments in lung cancer patients was generally shorter than that in healthy people. However, there was overlap in the distribution of cfDNA fragment length between patient samples and healthy people.[Error! Bookmark not defined.] This suggests that in lung cancer, perhaps short fragments of ctDNA are present only in part of the samples. The differences in fragment lengths may be related to the mechanisms of ctDNA formation, which still require extensive research.

4.4.3 The performance of tumor DNA fraction estimation

The reference can affect tumor DNA fraction estimation. Since the simulated data in this chapter was generated using one sample from project K34, I selected samples from project K34 that did not detect any SNVs as the PON dataset. After bias removal, the performance of ichorCNA with default parameters did not show significant improvement compared to the results in Chapter 3. This is because the default parameters of ichorCNA set the initial tumor DNA fraction from 0.1 to 0.5. Since ichorCNA uses the EM algorithm to find local optima, the initial tumor DNA fraction can affect the resulting tumor DNA fraction, as discussed in detail in Chapter 3. The results of ichorCNA with optimized parameters showed improvement compared to the results in Chapter 3. The tumor DNA fraction estimated by ichorCNA with optimized parameters and the expected tumor DNA fraction had a strong linear relationship. However, these predicted tumor DNA fractions were generally slightly larger than the corresponding expected tumor DNA fractions. ACE also showed a strong linear relationship between the estimated tumor DNA fraction and the expected tumor DNA fraction. Overall, after bias removal, both ichorCNA with optimized parameters and ACE demonstrate good performance in tumor DNA fraction estimation.

4.4.4 The performance of CNV detection

As for the prediction of CNV events, the performance of ichorCNA with optimized parameters was improved compared to the result in Chapter 3. Using samples from the same batch as reference might eliminate coverage bias specific to experimental and sequencing processes[149], potentially leading to the observed improvement. However, the performance of ACE in CNV detection declined compared to Chapter 3. In Chapter 3, the simulated data was generated from tissue samples, while in this chapter, the simulated data was generated from cfDNA enriched by short fragment size. This difference may contribute to the discrepancy in performance. Currently, although there are several studies indicating that ctDNA fragments are relatively shorter compared to cfDNA, it is still not clear whether the short DNA fragments are uniformly distributed throughout the entire genome. In this chapter, I enriched short fragments to generate the simulated data. This enrichment process may result in locally increased or decreased coverage, affecting the ability of PoN datasets to effectively remove bias. The performance of WisecondorX in CNV detection significantly declined compared to Chapter 3. This is because, in Chapter 3, WisecondorX used the normal sample for which the simulated data was generated, enabling it to effectively eliminate background noise. In this chapter, the reference was replaced by the PoN dataset, resulting in a degradation of the performance to detect CNVs.

In summary, when the expected tumor DNA fraction is greater than 8%, ACE achieves a prediction accuracy of over 90% for CNV events. Additionally, both ichorCNA with default parameters and ichorCNA with optimized parameters achieve a prediction accuracy of around 80%. However, when the tumor DNA fraction is low, the accuracy of these four tools is not ideal. When the tumor DNA fraction is extremely low, the noise in the sample can be easily mistaken for CNV events, thereby affecting the accuracy of the CNV detection. Therefore, these tools are more suitable for samples with higher tumor DNA fractions when detecting CNV events.

5. CNV detection tool benchmark based on liquid biopsy samples

In the previous sections, the CNV detection tools were evaluated using simulated data. The advantage of simulated data is that the sample's tumor DNA fraction and CNV events are already known and can be used as ground truth to evaluate the tools. However, for real samples, the accurate tumor DNA fraction and CNV events are not known. In this chapter, to evaluate the tools' ability to detect CNV using real samples, I compared the CNV results of blood samples and tissue samples from the same patients over the same time period to assess the consistency of the results.

5.1 Introduction

For NGS samples, the commonly used CNV detection method is based on read depth (RD). This method indicates copy number amplification and deletion through the read depth difference between tumor sample and control sample in sliding windows. The core principle of the RD method is based on a linear relationship between the RD and CNV. Through methods based on statistical models and machine learning, such as Hidden Markov model and circular binary segmentation (CBS), RD is processed to find the copy number variation region. In theory, the sequencing process is uniform, and the RD in sliding windows on the chromosome should be subject to Gaussian distribution.[150] An increase or decrease in the RD indicates that a CNV has occurred. However, the deviation of GC content, mapping affinity, and the background noise introduced during the experimental procedures and sequencing process make the relationship between RD and CNV not linear, so the accuracy of CNV detection will be affected. Current CNV detection tools often include correction sections for

GC and mapping affinity bias. As for the background noise introduced in the process of experiment and sequencing, a segmentation quality score can be introduced to evaluate the CNVs' accuracy.

The data utilized in this chapter was sourced from the INFORM (Individualized Therapy for Relapsed Malignancies in Childhood) project, which was initiated by the Society for Pediatric Oncology and Hematology (GPOH) in collaboration with the German Cancer Consortium (DKTK).[151] NGS was employed to acquire the biological attributes of every patient, and a skilled panel of specialists then evaluated and categorized the identified abnormalities in each patient, considering their clinical significance. The advantage of the INFORM project is that it contains tissue samples and blood samples from the same patient in the same timeline. Because of the advantages of higher coverage, and the reference from the same patient to remove bias, the CNVs detected in tissue samples can be used as ground truth to evaluate the performance of tools to detect CNV events in real liquid biopsy samples. The tissue samples in INFORM are WES samples, CNVkit was used to detecte CNVs for these WES samples. CNVkit is a CNV calling tool published in PLOS computational biology in 2016. It is characterized by the fact that copy number variation analysis can be performed on specified regions.[87]

5.2 Methods

To benchmark the effectiveness of different CNV detection tools for liquid biopsy, patients with both liquid biopsy samples and tissue samples were selected. Since the liquid biopsy sample and tissue sample come from the same patient at the same sampling time, they should have a similar CNV profile. By comparing the CNV results of liquid biopsy samples with the CNV results of tissue samples, it is possible to determine which tool is more accurate for liquid biopsy samples (Figure 5-1).

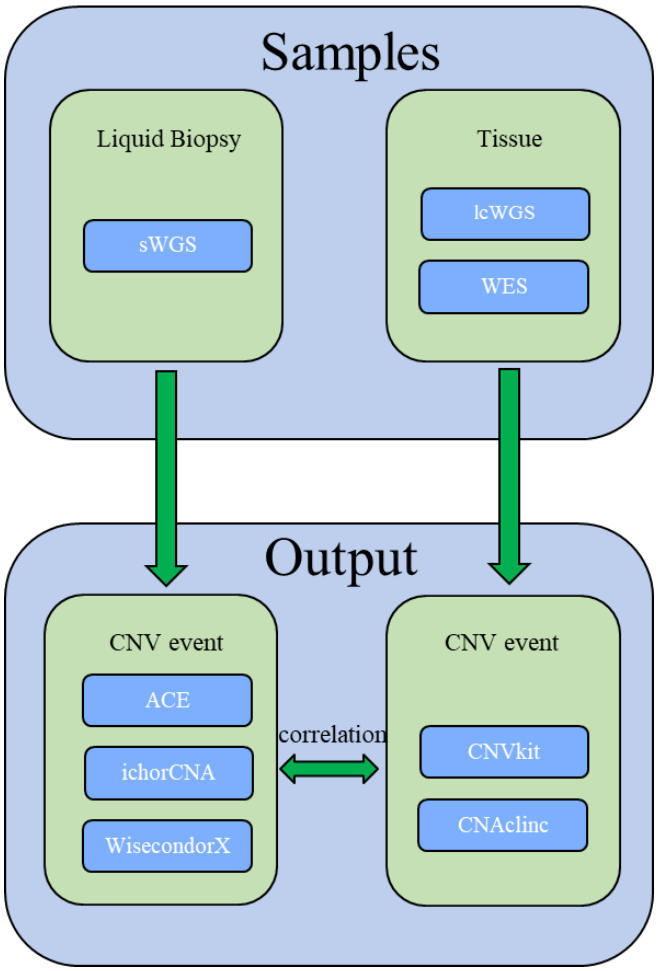


Figure 5-1 The analysis pipeline in this chapter.

5.2.1 Data preparation

In the INFORM project, the tissue samples include two types of sequencing data, namely lcWGS and WES. The lcWGS data provides coverage across the entire genome, enabling the detection of CNVs throughout the whole genome, including repeats and non-coding regions. However, the lcWGS data has a lower coverage of 5X. On the other hand, the WES samples have a higher coverage of around 200X, but they can only identify variations within the exons, limiting their ability to detect CNVs across the entire genome. INFORM also has sWGS (shallow Whole Genome Sequencing) data of blood samples from the same patient in the same timeline, with a coverage of about 0.5X.

To judge the accuracy of CNV results, patients with both tissue and blood samples were enrolled in the dataset in this section. In the previous chapter, I have already evaluated that ichorCNA with default parameter was the most accurate to determine whether a sample contains tumor or not. So one criterion for determining whether the sample can be enrolled here was to use the ichorCNA with default parameter to check whether the tumor DNA fraction of the sample is greater than 0. Finally, 30 sets of samples were collected in this project.

5.2.2 Tools

In this chapter, I used CNVkit, CNAclinic[152], ichorCNA, ACE and WisecondorX for CNV detection. The WES samples were detected by CNVkit, while the lcWGS samples were analyzed by both CNAclinic and CNVkit. As for the liquid biopsy samples, I used ichorCNA (default parameter), ACE modified with reference, and WisecondorX to detect their CNV.

5.2.3 Estimate CNV of tissue samples (CNVkit, CNAclinic)

CNVkit and CNAclinic were used to estimate the CNV events of tissue samples. Among them, CNVkit was used to estimate the CNV event of WES and lcWGS samples, while CNAclinic was used to estimate the CNV of lcWGS samples.

The workflow of CNVkit was as follows.

1. Identification the target regions and add gene annotation information. This step was achieved by guess_baits.py of CNVkit.

```
guess_baits.py -g access.hg19.bed Sample1.bam Sample2.bam -o baits.bed
```

2. The sequencing depth calculation. Two subcommands ‘coverage’ and ‘autobin’ were used in this step.

```
1. cnvkit.py autobin *.bam -t baits.bed -g access.hg19.bed
2. cnvkit.py coverage Sample.bam baits.target.bed -
   o Sample.targetcoverage.cnn
3. cnvkit.py coverage Sample.bam baits.antitarget.bed -
   o Sample.antitargetcoverage.cnn
```

3. The normal genome sequencing distribution model construction. This was achieved by the ‘reference’ subcommand.

```
1. cnvkit.py reference *Normal.{,anti}targetcoverage.cnn --fasta hg19.fa -
   o my_reference.cnn
```

4. Systematic biases correction and log2 ratio calculation.

```
1. cnvkit.py fix Sample.targetcoverage.cnn Sample.antitargetcoverage.cnn my_ref
   erence.cnn -o Sample.cnr
```

5. Segmentation.

```
1. cnvkit.py segment Sample.cnr -o Sample.cns
```

CNVkit reports only the log2 ratio but not the CNV events of each segment, so the threshold of CNV events needs to be determined. In this project, the same reference was used for all samples, but the tumor purity of each sample was different. Hence it is inappropriate to use a fixed log2 ratio as the threshold for determining the CNV event. To give each sample a suitable threshold for their own, the log2 ratio was employed to indicate the status of CNV. CNVkit recalibrated each log2 ratio by subtracting the median value derived from all log2 ratios. Ideally, a log2 ratio near 0 signifies diploid status. As for the thresholds of gain and loss, they can be obtained by the iterative method. The initial thresholds were set to 0.2 (gain) and -0.2 (loss), the distance score was calculated by Equation 1. Then, the gain threshold (gt) was increased to 0.5 by step 0.01, while the loss threshold (lt) decreased to -0.5 by step 0.01. Finally, gt and lt with the minimal distance score were used as the final threshold.

$$score = \sum_{<lt}(distance\ to\ lt)^2 + \sum_{>lt\ and<gt}(distance\ to\ 0)^2 + \sum_{>gt}(distance\ to\ gt)^2 \qquad \textit{Equation 1}$$

CNAclinic was specifically developed for CNV detection of lcWGS samples. The workflow involved several steps. Initially, both the tumor sample and control sample were downsampled using the subsetData function, resulting in files containing 10,000,000 reads each. Subsequently, the data underwent processing and preparation using the processForSegmentation function. In the next step, the data was segmented using the runSegmentation function. Finally, CNV detection and calculation were performed on the segmented results to derive the final outcome.

5.2.4 Estimate CNV of liquid biopsy samples

Liquid biopsy samples were analyzed using ichorCNA, WisecondorX, and ACE to detect CNV events. In the previous section, the PoN was used to remove biases from the samples. Both ichorCNA and WisecondorX allow the creation of a specialized reference file using PoN samples. For ACE, I modified the code to include a reference comparison process (as described in section 4.2.2.3), enabling all three software tools to remove biases using PoN datasets. Following bias correction, default parameters were used to detect CNVs in the samples using ichorCNA, WisecondorX, and ACE.

5.2.5 Correlation between tissue samples and ctDNA samples

In this project, the tissue samples and ctDNA samples were from the same patient with the same timeline. It can be considered that CNVs detected in ctDNA samples should have a high consistency with CNVs detected in tumor tissue samples. For the purpose of evaluating the accuracy of CNV detection of liquid biopsy tools, CNV results of tissue samples was regarded

as ground truth since tissue samples had higher coverage and higher tumor load. By comparing the results of different liquid biopsy CNV detection tools with the results of tissue samples, it was possible to determine which tools were more accurate.

The correlation coefficient between the predicted results of different tools in the same sample can reflect the similarity of the predicted results. The whole genome was divided into 2880 bins of 1MB bin size. There were three types of CNV events for each bin, gain, loss, and neutral. The CNV events (gain or loss) predicted by different tools on the same bin were compared. If the events given by two different tools on the same bin were both gain or loss, this bin was marked as 1. If the events were different, this bin was marked as 0. Finally, the consistency of the two tools for the same sample was obtained by using Pearson correlation. The higher the correlation coefficient, the closer the two predicted results were.

5.2.6 Segmentation quality score

Segmentation is an important part of CNV detection, the quality of segmentation directly affects the CNV detection results. After GC bias removal and reference correction, the CNV detection tools segment the samples. If there is a significant fluctuation in the read numbers between neighboring bins within the same segment, it is difficult to find a copy number close to the true value, then we can conclude that the quality of the segmentation is not high. A segmentation quality score was introduced to determine whether the segmentation quality was good or not. The segmentation distance score is a sum score of the distance square between the bins' log2 Ratio and their segment's log2 Ratio. To obtain the segmentation quality score, I subtracted the normalized segmentation distance score from 1. Normalization was implemented using the sklearn package's QuantileTransformer function with parameters n_quantiles=5 and random_state=0.

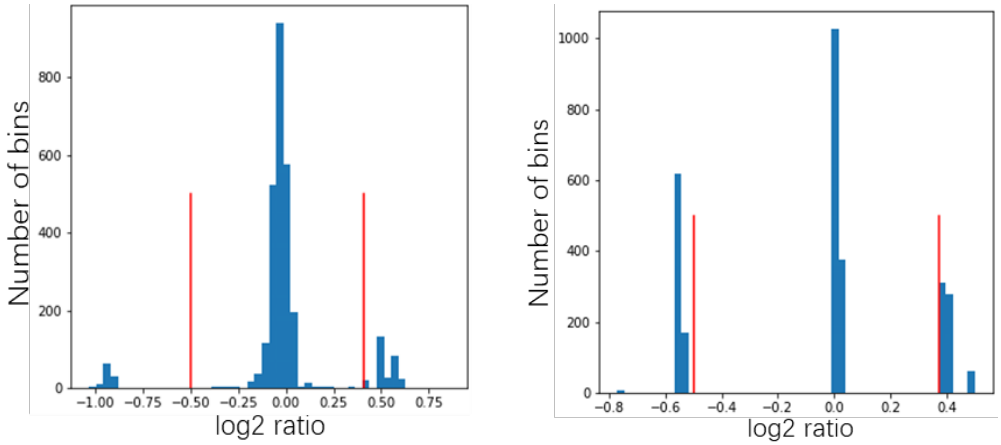
$$Segmentation\ quality\ score = 1 - \text{Normalizaion}(\sum(log2Ratio\ of\ bin - log2Ratio\ of\ segement)^2$$

Equation 2

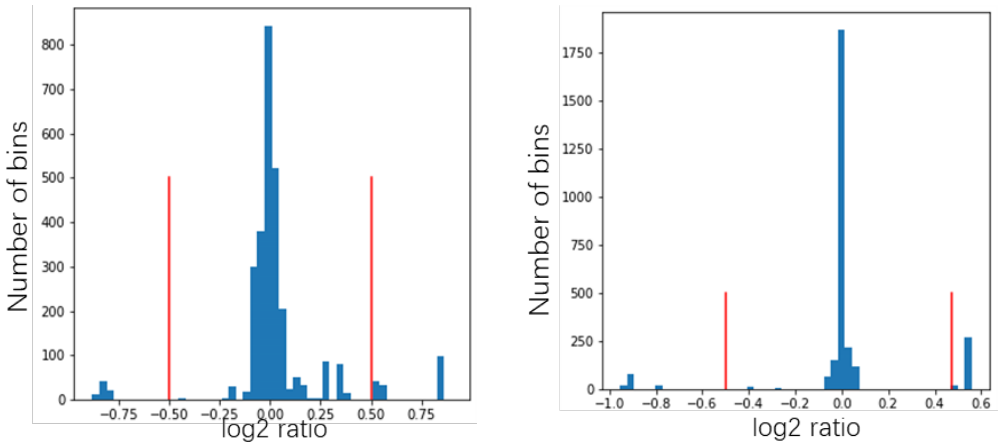
5.3 Results

5.3.1 Estimate CNVs of tissue samples

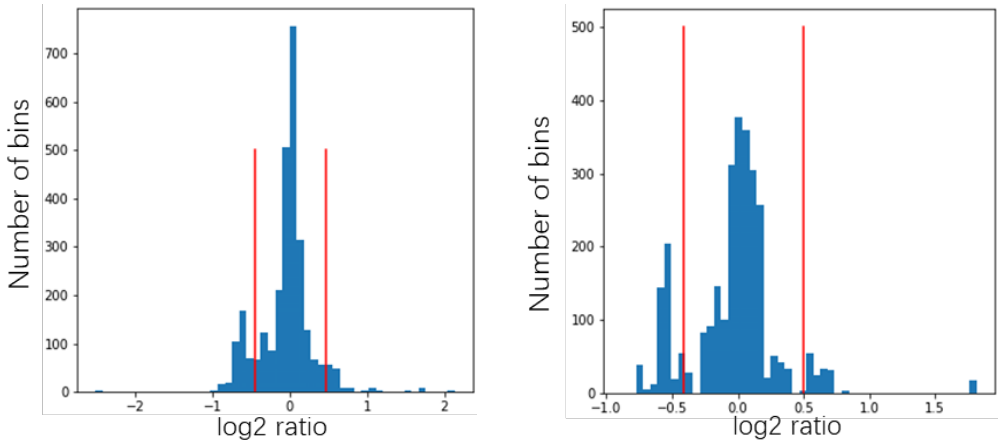
It is necessary to determine a specific threshold of CNV event for each sample. Due to variations in coverage, tumor purity and quality among samples, the log2 ratio distribution of each sample was different. According to different log2 ratio distributions, the thresholds of gain event and loss event should also be changed. Figure 5-2 shows the effect of using different thresholds to distinguish CNVs, using the histogram of log2 ratios for six samples as an example. The X axis represents the bins' log2 ratio value, and the y axis represents the number of bins. By using the method described in section 5.2.3 to determine the gain and loss event thresholds (shown as red lines), all bins in the samples were divided into three clusters (loss event region, neutral region, and gain event region). Bins falling into the loss event region were determined to have undergone loss events, those in the gain event region were determined to have undergone gain events, and those in the neutral region were determined to have not undergone any CNV event. It can be observed that in (a) to (d), the samples were distinctly categorized into three regions by the red lines, indicating that the threshold determination method aligned with expectations. In (e) and (f), the boundary between gain and loss was not particularly clear. This may be due to the complex CNV pattern in the sample or the presence of subclones. For these samples, multiple thresholds were theoretically required to determine the copy number. In order to simplify the problem, I chose only one threshold for the same sample to distinguish gain or loss, and the red line in the graph was the threshold that minimizes the distance score.



(a)5LB-022 WES (b)5LB-022 lcWGS



(c)2LB-033 WES (d)2LB-033 lcWGS



(e) 2LB-022 WES (f) 2LB-022 lcWGS

Figure 5-2 Histograms of read number's log2ratio per bin for 6 samples. The samples were (a)5LB-022_WES, (b)5LB-022_lcWGS, (c)2LB-033_WES, (d)2LB-033_lcWGS, (e)2LB-022_WES and (f) 2LB-022_lcWGS. The X axis represents the bins' log2ratio, and the y axis represents the number of bins. The red line on the left is the loss threshold, and the red line on the right is the gain

threshold.

5.3.2 Correlation between tissue samples and ctDNA samples

I compared the consistency of CNV results obtained from 30 tissue samples with different sequencing methods and different CNV detection tools. The CNV results by lcWGS-based CNAclinic, lcWGS-based CNVkit, and WES-based CNVkit were compared pairwise. As shown in Figure 5-5, the pairwise correlation coefficients between lcWGS-based CNAclinic, lcWGS-based CNVkit, and WES-based CNVkit were much higher. Among them, the median correlation coefficient of lcWGS-based CNVkit and CNAclinic reached 0.950, while the median correlation coefficient of WES-based CNVkit and lcWGS-based CNVkit was 0.935. The median correlation coefficient between WES-based CNVkit and lcWGS-based CNAclinic was 0.898. In general, the results of tissue samples with different sequencing methods and CNV detection tools were consistent.

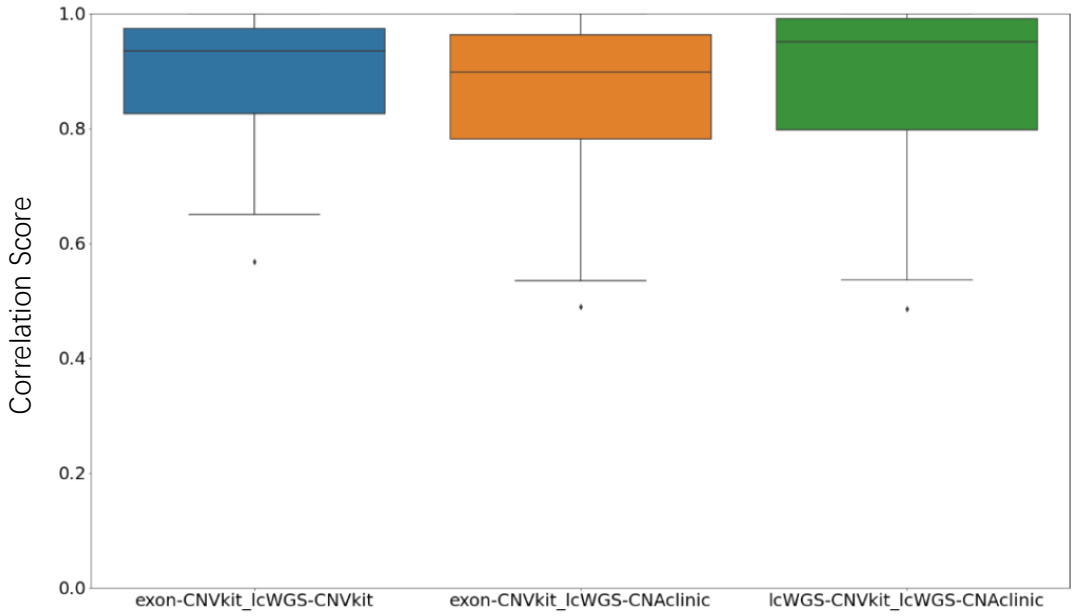


Figure 5-3 The correlation coefficient between tissue CNV results by lcWGS-based CNAclinic, lcWGS-based CNVkit, and WES-based CNVkit. The pairwise correlations among the three tools are relatively high.

Afterwards, I compared the consistency between the CNV results obtained from 30 sWGS samples using ACE, ichorCNA, and WisecondorX with the results obtained from the tissue samples. The median consistency of ACE results of 30 sWGS samples and the results of corresponding tissue samples was 0.32. For ichorCNA, the median consistency with tissue samples was 0.28. For WisecondorX, the median consistency with tissue samples was 0.23. ACE exhibited the highest degree of correlation, followed by IchorCNA. Conversely, WisecondorX demonstrated poor performance in comparison. In general, CNV results from ctDNA and tissue were poorly consistent.

In the subsequent analysis, I selected ACE with the best performance to observe the impact of segmentation quality on the consistency of CNV results. In Figure 5-4, the x-axis represents 30 patients, each of whom has three types of sequencing data: ctDNA by sWGS, tissue by lcWGS, and WES. The lcWGS and WES samples have matched normal sequences that can help eliminate errors caused by sequencing technologies. To evaluate various tools for detecting CNVs in ctDNA, lcWGS and WES samples were used as ground truth to calculate the correlation between the CNV detection results of tissue samples and liquid biopsy samples. The y-axis represents the data type used for comparison and the corresponding CNV detection tool. As for the color scale, white represents the highest correlation and black represents no correlation. The y-axis also includes the segmentation quality of ACE (for ctDNA), where white represents high quality and black represents low quality.

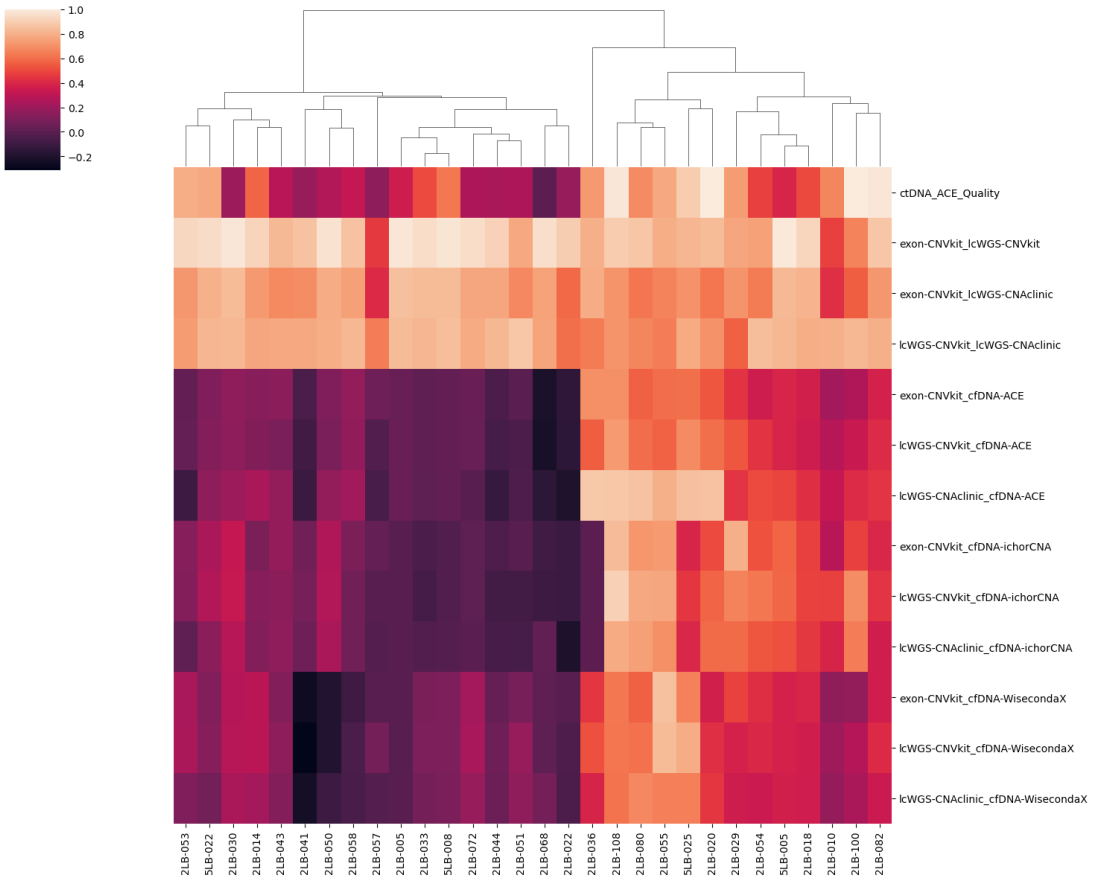


Figure 5-4 Correlation of CNV detection tools for tissue samples and for liquid biopsy samples. *The X axis represents 30 samples, and the Y axis* represents the data type used for comparison and the corresponding CNV detection tool*. White indicates high correlation or high segmentation quality, black indicates low correlation or low segmentation quality.*

As shown in Figure 5-4, ACE, ichorCNA, WisecondorX, and CNV results of tissue samples showed high correlations for samples with high segmentation quality (the right region of the figure), while the performance of these three tools was poor for samples with low segmentation quality (the left region of the figure).

In Figure 5-5, the x-axis is the segmentation quality score of ctDNA samples analyzed by ACE, the y-axis is the mean value of correlation between the liquid biopsy samples detected with ACE and the tissue samples detected by different methods. It is observed that when the segmentation quality score increased, there was a higher correlation between CNVs of ctDNA samples and CNVs of tissue samples. When the samples were divided into two groups of 15

samples each according to the segmentation quality score (Figure 5-6), there was a significant difference in the correlation between the CNV results detected by the best performing ACE and the tissue sample detection results (p-value=0.002). The median correlation coefficients of the two groups of samples were 0.41 and 0.04, respectively.

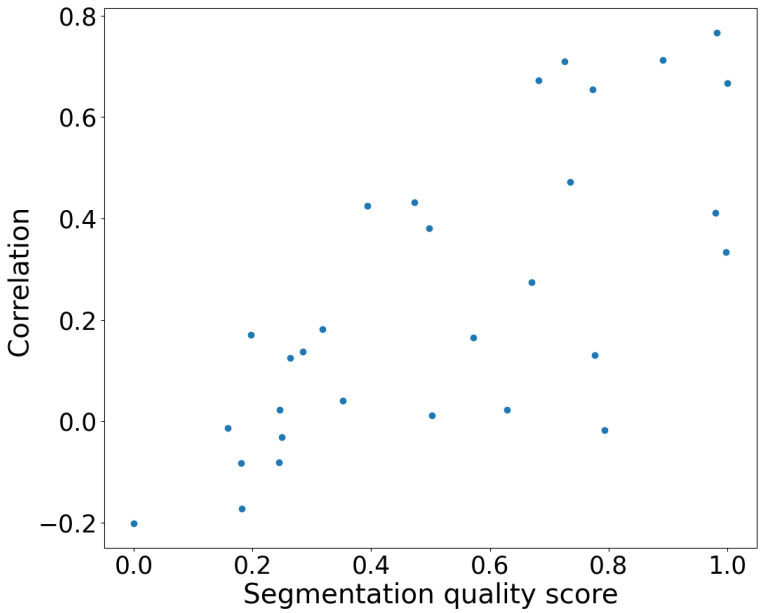


Figure 5-5 The segmentation quality affects the correlation. *The x-axis is the segmentation quality score of ctDNA samples analyzed by ACE, the y-axis is the mean value of correlation between the liquid biopsy samples detected with ACE and the tissue samples detected by different methods.*

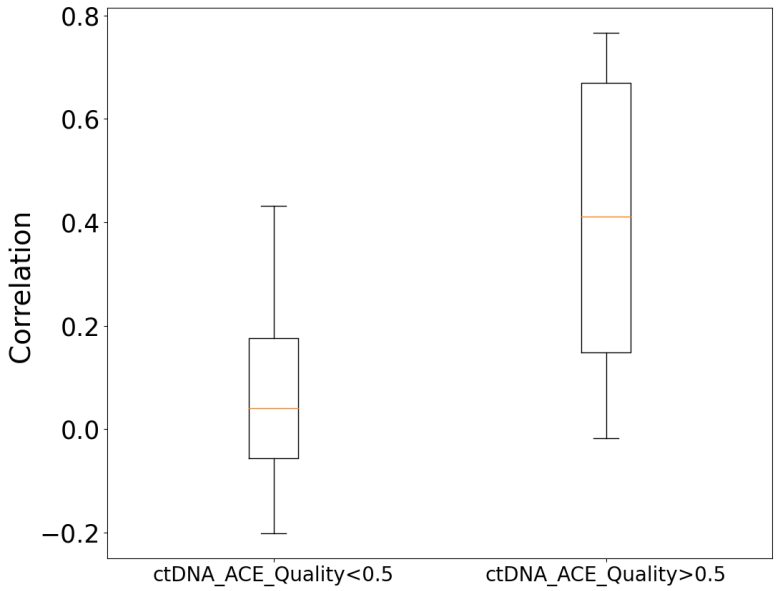


Figure 5-6 Boxplot shows the difference between the segmentation quality smaller than 0.5 and the segmentation quality larger than 0.5. *The samples were divided into two groups, depending on the segmentation quality obtained from ACE. When the segmentation quality was higher than 0.5, the consistency of CNV results obtained from ctDNA and tissue was improved.*

5.4 Discussion

In this chapter, the consistency of CNV results analyzed by different tools and different sample types is discussed. I introduced the segmentation quality score to evaluate the accuracy of CNV results of liquid biopsy samples.

5.4.1 The consistency of CNV among different tools and sample types

The study aimed to investigate the consistency of CNV detection results among different tools for both tissue and liquid biopsy samples.

5.4.1.1 The tissue samples showed high consistency in CNV detection

This section addresses the consistency of CNV results between lcWGS and WES samples. As CNVkit does not establish a fixed threshold for CNV events, a dynamic threshold was proposed to better capture the CNV events in diverse samples compared to the default threshold, ensuring more accurate results.

The lcWGS samples encompass the entire genome, albeit with a low depth. Conversely, WES samples exhibit greater depth, but they do not provide complete genome coverage. The CNV results of lcWGS tissue samples and WES tissue samples reveal a high level of consistency. This can be attributed to two factors. Firstly, the tissue samples have corresponding controls to eliminate the noise caused by sequencing. Secondly, noise has minimal impact on CNV detection when the tumor purity is high. Researchers have demonstrated that despite the poor DNA quality and increased noise observed in formalin-fixed and paraffin-embedded tissues, CNV detection results remain reliable when the tumor proportion exceeds 20%[153]. Overall, there is a high consistency in the CNV results of tissue lcWGS samples and WES samples when using CNVkit and CNAclinic.

5.4.1.2 The liquid biopsy samples showed lower consistency with tissue samples in CNV detection

In comparison to tissue samples, there is a lack of high consistency in CNV event detection results among different tools for liquid biopsy samples. The findings reveal that the

consistency between ACE and tissue samples is notably higher than that between the other tools and tissue samples. However, even the best-performing ACE exhibits a correlation of only 0.317.

Previous studies have also indicated a lower level of consistency in CNV detection results between ctDNA and tissue samples. Research conducted by Molparia et al. highlighted that the copy number and length of CNV regions can impact detection sensitivity. When there is minimal variation in CNV copy numbers and shorter CNV regions, a higher sequencing depth is required for accurate CNV detection[154]. In the study by Chae Y K et al., CNV results from tissue and liquid biopsies of 86 Breast Cancer samples were only 3.5% consistent[155]. The study of R Wang et al. showed that in Aggressive variant Prostate Cancer, the CNV consistency of tumor tissue and ctDNA was 20.2%[156]. From the above studies, it can be found that tumor type may affect the detection of CNV in ctDNA, but even in more aggressive cancer species, the consistency of results between tissue and ctDNA is still low.

The consistency of CNV results between ctDNA and tissue is low, which may be due to tumor heterogeneity. ctDNA contains the genomes of all cancer cells in the body, while tissue samples contain only the genomes of the tissues from which they were extracted. Some samples (2LB-053, for example) have high segmentation quality, but the correlation with tissue samples is low. This suggests that the CNVs in these samples may be different in tissue and ctDNA samples, possibly due to tumor heterogeneity. On the other hand, CNV results of ctDNA may be influenced by low sample quality. As mentioned in the previous section, when tumor proportion is high enough, poorer DNA quality and increased noise have less effect on CNV detection. However, the tumor DNA fraction of liquid biopsy samples is low, and poor DNA quality may lead to increased noise and affect CNV detection.

5.4.2 The segmentation quality affects the CNV detection

In this study, the segmentation quality score was introduced as a measure of the disparity in sequencing depth between neighboring regions. When this disparity is significant, it can lead to inaccurate segmentation during CNV detection. Various factors, such as sample quality and sequencing technology, can contribute to the substantial differences in sequencing depth observed in adjacent regions. For instance, when sample quality is compromised due to severe DNA degradation or the presence of numerous impurities, it can result in substantial variations in sequencing depth between neighboring locations[157]. Second, the experimental process or sequencing platform may lead to uneven sequencing depth. When sample uniformity is poor, it is difficult to reliably identify CNV[158, 159]. In addition, when the CNV detection tools show poor segmentation performance due to its own algorithm, the accuracy of CNV detection will also be reduced.

The segmentation quality score can serve as an indicator of the accuracy of CNV results to a certain extent. The findings presented in section 5.3.2 demonstrated that when liquid biopsy samples with higher segmentation quality were chosen, the detected CNV results exhibited greater consistency with the CNV results obtained from tissue samples.

6. Analysis Pipeline for liquid biopsy samples

In this section, I summarized all previous evaluation results and developed a data analysis pipeline specifically designed for liquid biopsy samples, which includes both sWGS and panel sequencing data. This pipeline was used to reanalyze the samples in Chapter 2.

6.1 Introduction

Avenio panel sequencing allows gene sequencing analysis in the region including 17 genes (as shown in Table 6-1). With high coverage sequencing (average 5000X), it can detect SNVs of the genes in the table as well as Indels of ALK, APC, BRAF, EGFR, ERBB2, KIT, MET, and TP53. It can also detect fusions of ALK, RET, and ROS1, as well as CNVs of EGFR, ERBB2, and MET. In the detection of CNVs, panel sequencing only provides information on whether there are gain events in the three genes of MET, EGFR, and ERBB2. In fusion detection, the fusion score of each sample is obtained by calculating the proportion of the number of reads where fusion occurs in all reads.

Table 6-1 The 17 gene regions included in panel sequencing. Green represents the type of variation the gene can be detected for.

Gene	SNV	Indel	Fusion	CNV
ALK				
APC				
BRAF				
BRCA1				
BRCA2				
DPYD				
EGFR				

ERBB2				
KIT				
KRAS				
MET				
NRAS				
PDGFRA				
RET				
ROS1				
TP53				
UGT1A1				

sWGS samples can determine the CNVs and tumor purity using low coverage sequencing technology. The critical loci of certain key genes in tumor cells can be used to monitor the cancer progression of patients and determine the degree of resistance to tumor drugs. However, due to the low coverage, the variations of these critical loci cannot be detected by sWGS and can only be detected in samples with high coverage such as panel sequencing. Similarly, the detection of gene fusion requires samples with high coverage using panel sequencing. For the detection of CNVs, both panel sequencing and sWGS can detect CNVs to some extent. In this chapter, the analysis results of the two types of sequencing data will be considered comprehensively.

6.2 Methods

6.2.1 Overview of the pipeline

To facilitate the efficient and automated analysis of paired samples of bulk ctDNA sWGS and panel sequencing, a tailored pipeline was developed, drawing from the evaluation results discussed in the preceding chapters. The workflow for this pipeline is outlined as follows (Figure 6-1).

Firstly, the PoN dataset selection and reference establishment were conducted. A subset of relatively tumor DNA free samples was chosen as the PoN dataset, which was used to establish the reference.

Secondly, quality control was performed on both panel sequencing and sWGS samples. The panel sequencing data was generated by Avenio, and SNVs with a MAF value exceeding 40% were filtered out. This is because, in liquid biopsy samples, the tumor DNA fraction is typically low, and SNVs with a MAF value above 40% are more likely to originate from the germline rather than the tumor. Fingerprinting was employed to verify the consistency between the sWGS and panel sequencing data, ensuring that the samples originated from the same patient.

Thirdly, the tumor DNA fraction and CNVs were detected. Two tools, ichorCNA and ACE, were utilized in this process. As described in previous chapters, modifications were made to these tools to improve the detection of tumor DNA fraction and CNVs. In addition, a segmentation quality score was calculated using the method described in Chapter 5 to assess the sample's segmentation quality.

Finally, a comprehensive evaluation of the results was conducted. The reliability of the CNV predictions was assessed based on all the aforementioned information. The samples were assessed to determine if they contained tumors. For samples that were identified as containing tumors, the tumor DNA fraction and segmentation quality score were examined to determine the tumor burden and identify CNVs.

The above is an overview of the entire pipeline. The details of the pipeline will be described in the following sections.

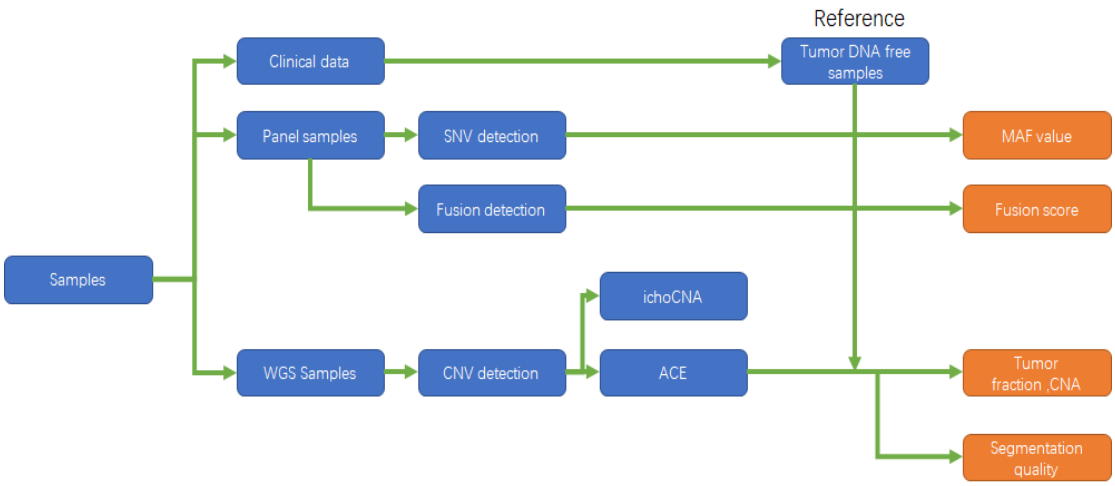


Figure 6-1 ***The workflow of the data analysis pipeline*** *for liquid biopsy samples.*

6.2.2 PoN dataset selection and reference establish

In this project, 11 samples were selected as the PoN dataset, depending on three criteria: first, no SNVs were detected in the Panel sequencing results; second, the samples were determined as clean by ichorCNA; third, the samples were assessed based on clinical information to determine if they were in the early stage. As described in section 4.2.2.1, the createPanelOfNormals function of ichorCNA was used to get the reference file for ichorCNA. This reference file can help to normalize the cancer patient cfDNA to correct for systematic biases arising from library construction, sequencing platform, and cfDNA-specific artifacts. Due to the lack of reference correction function in ACE, I made modifications to ACE, as described in section 4.2.2.3.

6.2.3 QC

Quality control is an important step in the pipeline. All samples were fingerprint tested to determine whether there was a high correlation between samples from the same patient. As described in section 2.2.2.3, the genotypes at the selected characteristic SNVs were listed in

a matrix, and the correlation coefficients between different samples were calculated. Then, the samples with high correlation were gathered through clustering, to test whether the pairing information of samples was correct. Among them, samples from the same patient with different sequencing techniques (Panel sequencing and sWGS) and samples from the same patient with different timelines were tested simultaneously, thus improving the accuracy and credibility of fingerprinting.

Fastqc was used to test the sequencing quality of all samples. After checking the basic statistics, per base sequence quality, per sequence quality scores, per base sequence content, sequence length distribution, sequence duplication level, etc., the samples with poor quality were marked.

In addition, the GC content per sequence of each sample was checked. The GC content of the sample is also an important factor that affects the quality of the sample. GC bias can be caused by the preference of sequencing technology. The commonly used method to eliminate bias is to obtain the real reads number by loess regression. After the removal GC bias, the GC content distribution graph of the sample should approximate a horizontal line (as described in Chapter 2).

6.2.4 Tools adjustments

In this pipeline, in order to better detect CNVs, the tools need parameter optimization. For ichorCNA and ACE, the adjustment of parameters was critical. Since there were no controls for liquid biopsy samples, the most important step was to find a suitable reference for noise cancellation so that CNV detection tools do not detect sequencing noise as CNV events. In addition, some other parameters need to be adjusted for each tool, details were as follows.

6.2.4.1 ichorCNA

For ichorCNA, the default parameter was employed to determine whether the sample contains tumor or not. The default parameters adopted initial normal contamination (0.5-0.9) and initial ploidy (2, 3, 4, 5). In addition, I added functionality to output reads number per bin after removal of GC bias for filtering low quality samples.

6.2.4.2 ACE

For ACE, I first expanded the detection range of ACE for tumor DNA fraction from the original 5%-100% to 1%-100% to accommodate low-tumor-fraction samples. Secondly, the bias correction function was added and the PoN data set was used as a reference to eliminate the impact of coverage bias on CNV detection. When calculating the copy number of each bin, the log2 value of reads number of bin in sample divided by the reads number of the same bin in reference was used, replacing the log2 value of the reads number of bin in sample.

6.2.5 Output

The output of the entire pipeline contains three parts: the output of panel sequencing, the output of sWGS, and the data analysis of the integrated results of both sequencing methods.

6.2.5.1 Panel sequencing output

The result of panel sequencing was provided by Avenio, providing information on SNVs, CNVs,

and fusions present within the targeted region. According to the manufacturer, Avenio demonstrates high sensitivity and positive predictive value (PPV) exceeding 99% for all types of mutations. In terms of SNV detection, panel sequencing can identify SNVs with an allele frequency greater than 0.5% accurately within the targeted region. The detection limit for fusion is 1%. For CNV detection, the detection limits for MET, EGFR, and ERBB2 are 2.3, 2.3, and 2.6 copies, respectively[160].

6.2.5.2 sWGS output

For every sWGS sample, ichorCNA was used to determine whether the sample was tumor DNA free. For the sample containing tumor, its whole-genome CNV results were exported by ACE, including all the solutions of tumor DNA fraction ranging from 1% to 10%. Of course, the tumor DNA fraction of the best solution was also shown. The detection limit of tumor DNA fraction was 1% to 10%, while the detection range of CNV was 0 to 5 copies. The segmentation quality scores of sWGS samples were also reported to evaluate the segmentation quality of sample segmentation and the accuracy of their CNV results.

6.3 Results

6.3.1 Comparison of CNV detection by panel sequencing and sWGS

Panel sequencing reports whether there are gain events in MET, EGFR, ERBB2, while sWGS reports CNV events in the whole genome. In order to verify the consistency of the results of the two sequencing methods, the CNVs detection results of sWGS on the three genes of MET, EGFR, and ERBB2 were compared with panel sequencing.

A total of 395 samples were sequenced by panel sequencing and sWGS simultaneously (Table 6-2). Among them, for the EGFR gene, 18 samples were predicted to have gain events by both panel sequencing and sWGS. 10 samples were only predicted to have gain events by panel sequencing, and 22 samples were only predicted to have gain events by sWGS. For the remaining 345 samples, panel sequencing and sWGS both reported that there was no gain event. For the MET gene, 15 samples were predicted to have gain events by both panel sequencing and sWGS. 8 samples were predicted to have gain events only by panel sequencing, and 26 samples were predicted to have gain events only by sWGS. 346 samples were predicted by panel sequencing and sWGS as no gain event. For the ERBB2 gene, only 1 sample was predicted to have gain events by both panel sequencing and sWGS. 14 samples were reported to have gain events in sWGS but not in panel sequencing. For the remaining 380 samples, no gain event was reported in panel sequencing and sWGS.

CNVs are a crucial factor in tumor load evaluation. Various sequencing methods can be used to detect CNVs, but it is essential to determine the consistency of results between different methods. By analyzing the CNV detection results of sWGS and panel sequencing methods for the above 395 samples, it can be found that 27 samples were detected to contain CNV patterns in both panel sequencing and sWGS. 6 samples were detected to contain CNV patterns in panel sequencing but not in sWGS. 50 samples contained CNV patterns in sWGS test but not in panel sequencing. In the remaining 312 samples, the CNV pattern was detected neither in panel sequencing nor in sWGS. In addition, among the six samples that were detected as negative by sWGS but positive by panel sequencing, 5 of them have a score of less than 5 in panel sequencing, which means that they have low confidence in the CNV detection results.

It can be observed in Table 6-2, most of the samples were not detected with gain events by either method. For the remaining samples, more gain events were detected by sWGS.

Table 6-2 CNV events were detected by both panel and sWGS. (a), (b) and (c) represent the CNV events in the EGFR, MET, and ERBB2 respectively. (d) shows whether CNV exists in the sample.

	EGFR sWGS with CNV gain	EGFR sWGS without CNV gain
EGFR Panel with CNV gain	18	10
EGFR Panel without CNV gain	22	345

(a)

	MET sWGS with CNV gain	MET sWGS without CNV gain
MET Panel with CNV gain	15	8
MET Panel without CNV gain	26	346

(b)

	ERBB2 sWGS with CNV gain	ERBB2 sWGS without CNV gain
ERBB2 Panel with CNV gain	1	0
ERBB2 Panel without CNV gain	14	380

(c)

	sWGS with CNV	sWGS without CNV
Panel with CNV	27	6
Panel without CNV	50	312

(d)

6.3.2 Detection of ctDNA as biomarker for tumor samples

In addition to CNVs, SNVs and gene fusions can also be present in tumor samples. The panel sequencing samples was determined to contain tumor by detecting the presence of SNV, CNV, or fusion. As for sWGS samples, ichorCNA was used to determine if they contain tumor. A confusion matrix between the results of panel sequencing and the results of sWGS was created (Table 6-3), it can be found that 66 samples were detected to contain tumor by both panel sequencing and sWGS. 152 samples were detected to contain tumor by panel

sequencing but not by sWGS. 11 samples were detected by sWGS to contain tumor but not by panel sequencing. 166 samples contained no tumor by panel sequencing and sWGS. Through comparison, it can be found that for tumor detection, the consistency of the two sequencing methods was poor. However, it is worth noting that for the 152 samples that were detected as positive by panel but negative by sWGS, 148 of them had a fusion allele frequency less than 0.01, indicating a low tumor DNA fraction in these samples.

Table 6-3 Confusion matrix of the detection results between panel sequencing samples and sWGS samples from the same patient in the same timeline.

	sWGS with tumor	sWGS without tumor
Panel with tumor	66	152
Panel without tumor	11	166

It can be speculated that the low allele frequency (AF) of SNV and fusion may be a factor that affects the consistency of detection. In order to verify this hypothesis, samples were compared respectively by SNV AF and fusion AF as biomarkers. Among them, the two groups were further divided into four levels according to the allele frequency of their panel sequencing, respectively AF=0, AF between 0-0.01, AF between 0.01-0.05, and AF greater than 0.05. The AF value here represents the maximum AF value among all SNVs in a sample.

First, the SNV AF in panel sequencing was used as a biomarker to evaluate the consistency of detection (Table 6-4a). Among the 188 samples reported by panel sequencing that did not contain SNV, 16 samples were detected as containing tumor by sWGS. Among 119 samples with AF between 0-0.01, sWGS detected 19 samples with tumor. Among 61 samples with AF between 0.01-0.05, sWGS detected 19 samples with tumor. Among 27 samples with AF greater than 0.05, sWGS detected 23 samples containing tumor.

Among the 35 samples with an SNV AF value less than 0.01 and tumors detected by sWGS, 12 samples were detected by panel sequencing as containing fusion, 10 samples contained

CNV, and 8 samples only contained SNV. Among the 16 samples that did not contain SNV, a total of 11 samples did not contain any mutations according to the panel sequencing, that is, did not contain any SNV, CNV, or fusion.

For all 61 samples with SNV AF between 0.01-0.05, tumor was detected in only 19 sWGS samples, resulting in a low consistency of 31.1% when compared to the panel results. Among 27 samples with SNV AF greater than 0.05, 23 were detected to contain tumor by sWGS, resulting in a consistency of 85%.

Second, the fusion AF was used as the biomarker for evaluation (Table 6-4b). By grouping the samples according to fusion AF, 35 samples with no fusion were detected by sWGS to contain tumor. Among 42 samples with a fusion AF between 0 to 0.01, sWGS detected 10 samples containing tumor. Among 30 samples with a fusion AF between 0.01 to 0.05, 26 samples were detected by sWGS to contain tumor. The 6 samples with a fusion AF greater than 0.05 were all detected by sWGS to contain tumor.

For the 35 samples with no fusion but had tumor detected by sWGS, 7 samples contained SNV and CNV, 15 samples contained only SNV, 2 samples contained only CNV, and the remaining 11 samples did not contain any mutations.

For all 30 samples of fusion AF between 0.01-0.05, tumor was detected in sWGS of 26 samples, resulting in a consistency of 86.7%. And tumor was detected in sWGS of 6 samples with fusion AF greater than 0.05, resulting in a consistency of 100%. When the fusion AF exceeded 0.01, only 4 out of 36 sWGS samples did not detect the tumor. Compared with the SNV results, maybe fusion events are more suitable as a biomarker for tumor burden detection.

Table 6-4 The number of sWGS samples that contain tumor according to different ranges of AF, using SNV and fusion as biomarkers, respectively. (a) The detection results of sWGS are classified according to the highest AF of SNV in the panel; (b) the detection results of sWGS are classified

according to the fusion AF in the panel.

	sWGS contains tumor	Total samples
Panel SNV AF =0	16	188
Panel SNV AF 0-0.01	19	119
Panel SNV AF 0.01-0.05	19	61
Panel SNV AF >0.05	23	27
Sum	77	395

(a)

	sWGS contains tumor	Total samples
Panel fusion AF 0	35	317
Panel fusion AF 0-0.01	10	42
Panel fusion AF 0.01-0.05	26	30
Panel fusion AF >0.05	6	6
Sum	77	395

(b)

6.4 Discussion

6.4.1 Panel sequencing combined with sWGS for tumor detection in liquid biopsy

At present, in the field of tumor detection, panel sequencing is widely used due to its low price and high sensitivity and specificity[161]. Panel sequencing is a good choice for detecting SNV and fusion of specific genes. In addition to SNV and fusion, CNV is also a potential biomarker or prognostic factor for tumor therapy. However, panel sequencing can detect a limited range of CNVs. For example, Avenio, used in this project, only reports the CNV of three genes: EGFR, ERBB2 and MET. Although these genes are commonly altered in lung

cancer, there are still other genes that are common in lung cancer that need to be tested. For example, for ALK fusion-positive tumors, amplification of MYC and MDM2 is common[162]. MDM2 amplification is associated with poor clinical outcomes and significantly increases tumor growth rates in PD-1 /PD-L1 immunotherapy[163]. In other cancers, such as Acute Lymphoblastic Leukemia, the absence of CDKN2A and CDKN2B are independent prognostic markers[164].

In addition, segmentation is one important step of CNV detection. This step uses statistical methods such as Hidden Markov models or circular binary segmentation to merge regions with similar read counts to estimate the CNV events within the region[131.] This means that CNV detection requires that the detection area is long enough and has a stable coverage. For panel sequencing, the detection area is limited, only specific genes can be detected, and the coverage heterogeneity caused by the hybridization capture step may affect the accuracy of CNV detection[158, 159]. Therefore, the ability of panel sequencing to detect CNVs has certain limitations. The ability of using sWGS for CNV detection has been demonstrated. For example, the researchers demonstrated that CNVs associated with glioma diagnosis can be detected using sWGS samples from glioma patients, and other glioma-associated abnormalities can also be revealed, such as EGFR amplification and homozygous loss of CDKN2A/B[165]. In one study of urothelial bladder carcinoma, although the average depth of the sWGS samples was 0.6X, amplification of MDM2, ERBB2, CCND1, and CCNE1 and deletion of CDKN2A, PTEN, and RB1 were observed. These are all known to change frequently in urothelial bladder carcinoma, and the CNV patterns of cfDNA showed similar patterns to tumor samples[166]. However, due to the low coverage of sWGS and the lower tumor content in cfDNA, the resolution of CNV detection is low, and it is more susceptible to noise, resulting in reduced accuracy[167]. Therefore, a comprehensive analysis of a cancer sample is required from the perspectives of SNVs, fusions, CNVs, etc. In this project, a combination of panel sequencing and sWGS methods was employed to better analyze the relevant tumor characteristics of the samples.

6.4.2 Consistency between panel sequencing and sWGS

As indicated in Table 6-3, 152 samples tested tumor positive by Panel sequencing, but were negative by sWGS. It is worth mentioning that among these 152 samples, 148 of them exhibited fusion AF below 1%. According to Avieno's documentation, when the fusion score exceeds 1%, the sensitivity and positive predictive value (PPV) are both above 99%. However, for samples with fusion AF below 1%, the detection accuracy may be compromised, leading to false positives. Another potential explanation is that although some samples may have SNVs or fusion events, there might be a low occurrence of CNV events, which may not be detected by CNV analysis tools.

As shown in Table 6-4, the consistency between the SNV results by panel sequencing and sWGS, as well as fusion detection results and sWGS, were analyzed separately. It can be observed that in this project, the consistency between sWGS results and fusion detection results is higher compared to SNV detection results. Especially when the fusion AF exceeded 0.01, only in 4 out of 36 samples sWGS analysis did not detect tumor DNA. It can be inferred that in this project, fusion events are more suitable as a biomarker for tumor burden detection. There are several possible reasons for this observation. Firstly, since all the samples analyzed in this project were cfDNA samples from tumor patients without corresponding control samples, it is difficult to determine whether the higher positive rate of SNV detected by panel sequencing is influenced by clonal hematopoiesis. Secondly, all the samples in this project were derived from ALK fusion-positive patients, indicating that fusion events may better reflect the tumor characteristics compared to SNVs in these specific cases.

6.4.3 Further works

In addition to using comprehensive detection methods based on SNVs, fusions, and CNVs as

mentioned in this project to detect tumors and monitor tumor progression, literature also suggests the use of t-mad score (Trimmed Median Absolute Deviation from copy number neutrality) to detect circulating tumor DNA[139]. The article mentioned that through the analysis of 97 samples, a strong correlation (Pearson correlation coefficient r=0.80) was found between t-mad and VAF in high ctDNA cancer types. Furthermore, another study showed that the t-mad scores of cfDNA at the 6th and 8th weeks after treatment in metastatic breast cancer patients are correlated with subsequent RECIST response on imaging[168]. There is one study demonstrated that t-mad score exhibits higher sensitivity and lower specificity than the mean value of VAF in NSCLC patients, and the t-mad score may be more suitable for use in the early stages of the disease[111.] In conclusion, the t-mad score can serve as a potential biomarker for detecting ctDNA.

Additionally, in SNV detection of liquid biopsy samples, panel sequencing of tumor-only samples has certain limitations. Due to the influence of clonal hematopoiesis, it is difficult to determine the origin of SNVs. To ascertain whether the SNVs identified in the sequencing are derived from circulating tumor DNA or from white blood cells, leukocyte separation sequencing is necessary[169]. However, in this project, only liquid biopsy blood samples were available, which poses challenges for accurate detection of tumor SNVs. In addition, a study by Sun J X et al. mentioned how to differentiate somatic mutations from germline mutations in tumor tissue samples[170]. Another research suggested the use of population frequency to remove common variants[171]. This may provide some ideas to improve the accuracy of SNV detection in cfDNA.

7. Literature

[1] Alexandrov L B, Nik-Zainal S, Wedge D C, et al. Signatures of mutational processes in human cancer[J]. Nature, 2013, 500(7463): 415-421.

[2] Goldfarb M, Shimizu K, Perucho M, et al. Isolation and preliminary characterization of a human transforming gene from T24 bladder carcinoma cells[J]. Nature, 1982, 296(5856): 404-409.

[3] Coelho M A, de Carné Trécesson S, Rana S, et al. Oncogenic RAS signaling promotes tumor immunoresistance by stabilizing PD-L1 mRNA[J]. Immunity, 2017, 47(6): 1083-1099. e6.

[4] Lalande M, Dryja T P, Schreck R R, et al. Isolation of human chromosome 13-specific DNA sequences cloned from flow sorted chromosomes and potentially linked to the retinoblastoma locus[J]. Cancer genetics and cytogenetics, 1984, 13(4): 283-295.

[5] Chinnam M, Goodrich D W. RB1, development, and cancer[J]. Current topics in developmental biology, 2011, 94: 129-169.

[6] Zapatka M, Borozan I, Brewer D S, et al. The landscape of viral associations in human cancers[J]. Nature genetics, 2020, 52(3): 320-330.

[7] Grønbaek K, Hother C, Jones P A. Epigenetic changes in cancer[J]. Apmis, 2007, 115(10): 1039-1059.

[8] Falzone L, Salomone S, Libra M. Evolution of cancer pharmacological treatments at the turn of the third millennium[J]. Frontiers in pharmacology, 2018, 9: 1300.

[9] Stratton M R, Campbell P J, Futreal P A. The cancer genome[J]. Nature, 2009, 458(7239): 719-724.

[10] Downing J R, Wilson R K, Zhang J, et al. The pediatric cancer genome project[J]. Nature genetics, 2012, 44(6): 619-622.

[11] Kandoth C, McLellan M D, Vandin F, et al. Mutational landscape and significance across 12 major cancer types[J]. Nature, 2013, 502(7471): 333-339.

[12] Chen J M, Cooper D N, Férec C, et al. Genomic rearrangements in inherited disease and cancer[C]//Seminars in cancer biology. Academic Press, 2010, 20(4): 222-233.

[13] Oshi M, Murthy V, Takahashi H, et al. Urine as a source of liquid biopsy for cancer[J]. Cancers, 2021, 13(11): 2652.

[14] Ilié M, Hofman P. Pros: Can tissue biopsy be replaced by liquid biopsy?[J]. Translational lung cancer research, 2016, 5(4): 420.

[15] Larson N B, Fridley B L. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data[J]. Bioinformatics, 2013, 29(15): 1888-1889.

[16] Underwood J J, Quadri R S, Kalva S P, et al. Liquid biopsy for cancer: review and implications

for the radiologist[J]. Radiology, 2020, 294(1): 5-17.

[17] Li G, Tang W, Yang F. Cancer liquid biopsy using integrated microfluidic exosome analysis platforms[J]. Biotechnology journal, 2020, 15(5): 1900225.

[18] Gao X H, Li J, Gong H F, et al. Comparison of fresh frozen tissue with formalin-fixed paraffin-embedded tissue for mutation analysis using a multi-gene panel in patients with colorectal cancer[J]. Frontiers in oncology, 2020, 10: 310.

[19] Diaz J L A, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA[J]. Journal of clinical oncology: official journal of the American Society of Clinical Oncology, 2014, 32(6): 579-586..

[20] Lone S N, Nisar S, Masoodi T, et al. Liquid biopsy: A step closer to transform diagnosis, prognosis and future of cancer treatments[J]. Molecular cancer, 2022, 21(1): 79.

[21] Fujihara J, Takinami Y, Ueki M, et al. Circulating cell-free DNA fragment analysis by microchip electrophoresis and its relationship with DNase I in cardiac diseases[J]. Clinica Chimica Acta, 2019, 497: 61-66.

[22] Zhang L, Liang Y, Li S, et al. The interplay of circulating tumor DNA and chromatin modification, therapeutic resistance, and metastasis[J]. Molecular cancer, 2019, 18(1): 1-20.

[23] Delgado P O, Alves B C A, de Sousa Gehrke F, et al. Characterization of cell-free circulating DNA in plasma in patients with prostate cancer[J]. Tumor Biology, 2013, 34: 983-986.

[24] Fernandez-Garcia D, Hills A, Page K, et al. Plasma cell-free DNA (cfDNA) as a predictive and prognostic marker in patients with metastatic breast cancer[J]. Breast Cancer Research, 2019, 21(1): 1-13.

[25] Lubotzky A, Zemmour H, Neiman D, et al. Liquid biopsy reveals collateral tissue damage in cancer[J]. JCI insight, 2022, 7(2).

[26] Schwaederle M, Husain H, Fanta P T, et al. Detection rate of actionable mutations in diverse cancers using a biopsy-free (blood) circulating tumor cell DNA assay[J]. Oncotarget, 2016, 7(9): 9707-9717.

[27] Wang R, Li X, Zhang H, et al. Cell-free circulating tumor DNA analysis for breast cancer and its clinical utilization as a biomarker[J]. Oncotarget, 2017, 8(43): 75742.

[28] Van Gelder R N. Molecular diagnostics for ocular infectious diseases: LXXVIII Edward Jackson memorial lecture[J]. American journal of ophthalmology, 2022, 235: 300-312.

[29] Horak P, Fröhling S, Glimm H. Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls[J]. ESMO open, 2016, 1(5): e000094.

[30] Kruglyak K M, Lin E, Ong F S. Next-generation sequencing and applications to the diagnosis and treatment of lung cancer[J]. Lung Cancer and Personalized Medicine: Novel Therapies and Clinical Management, 2016: 123-136.

[31] Shyr D, Liu Q. Next generation sequencing in cancer research and clinical application[J]. Biological procedures online, 2013, 15: 1-11.

[32] Sanger F, Thompson E O P. The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates[J]. Biochemical Journal, 1953, 53(3): 353.

[33] Niall H D. [36] Automated edman degradation: The protein sequenator[M]//Methods in enzymology. Academic Press, 1973, 27: 942-1010.

[34] Sanger F, Brownlee G G, Barrell B G. A two-dimensional fractionation procedure for radioactive nucleotides[J]. Journal of molecular biology, 1965, 13(2): 373-IN4.

[35] Holley R W, Everett G A, Madison J T, et al. Nucleotide sequences in the yeast alanine transfer ribonucleic acid[J]. Journal of Biological Chemistry, 1965, 240(5): 2122-2128.

[36] Sanger F, Coulson A R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase[J]. Journal of molecular biology, 1975, 94(3): 441-448.

[37] Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors[J]. Proceedings of the national academy of sciences, 1977, 74(12): 5463-5467.

[38] Maxam A M, Gilbert W. A new method for sequencing DNA[J]. Proceedings of the National Academy of Sciences, 1977, 74(2): 560-564.

[39] Sanger F, Air G M, Barrell B G, et al. Nucleotide sequence of bacteriophage φX174 DNA[J]. nature, 1977, 265(5596): 687-695.

[40] Zhang L, Chen F X, Zeng Z, et al. Advances in metagenomics and its application in environmental microorganisms[J]. Frontiers in microbiology, 2021, 12: 766364.

[41] Van Dijk E L, Auger H, Jaszczyszyn Y, et al. Ten years of next-generation sequencing technology[J]. Trends in genetics, 2014, 30(9): 418-426.

[42] Rothberg J M, Leamon J H. The development and impact of 454 sequencing[J]. Nature biotechnology, 2008, 26(10): 1117-1124.

[43] Smith L M, Sanders J Z, Kaiser R J, et al. Fluorescence detection in automated DNA sequence analysis[J]. Nature, 1986, 321(6071): 674-679.

[44] Bentley D R, Balasubramanian S, Swerdlow H P, et al. Accurate whole human genome sequencing using reversible terminator chemistry[J]. nature, 2008, 456(7218): 53-59.

[45] Kircher M, Heyn P, Kelso J. Addressing challenges in the production and analysis of illumina sequencing data[J]. BMC genomics, 2011, 12: 1-14.

[46] Rhoads A, Au K F. PacBio sequencing and its applications[J]. Genomics, proteomics & bioinformatics, 2015, 13(5): 278-289.

[47] Choi M, Scholl U I, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing[J]. Proceedings of the National Academy of Sciences, 2009, 106(45): 19096-19101.

[48] Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics[J]. Forensic Science International: Genetics, 2015, 18: 78-89.

[49] Nakagawa H, Wardell C P, Furuta M, et al. Cancer whole-genome sequencing: present and future[J]. Oncogene, 2015, 34(49): 5943-5950.

[50] Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data[J]. Frontiers in bioengineering and biotechnology, 2015, 3: 92.

[51] Tran B, Dancey J E, Kamel-Reid S, et al. Cancer genomics: technology, discovery, and translation[J]. J Clin Oncol, 2012, 30(6): 647-660.

[52] Teer J K, Mullikin J C. Exome sequencing: the sweet spot before whole genomes[J]. Human molecular genetics, 2010, 19(R2): R145-R151.

[53] Singleton A B. Exome sequencing: a transformative technology[J]. The Lancet Neurology, 2011, 10(10): 942-946.

[54] Serratì S, De Summa S, Pilato B, et al. Next-generation sequencing: advances and applications in cancer diagnosis[J]. OncoTargets and therapy, 2016: 7355-7365.

[55] Rehm H L. Disease-targeted sequencing: a cornerstone in the clinic[J]. Nature reviews genetics, 2013, 14(4): 295-300.

[56] Easton D F, Pharoah P D P, Antoniou A C, et al. Gene-panel sequencing and the prediction of breast-cancer risk[J]. New England Journal of Medicine, 2015, 372(23): 2243-2257.

[57] Sung J S, Chong H Y, Kwon N J, et al. Detection of somatic variants and EGFR mutations in cell-free DNA from non-small cell lung cancer patients by ultra-deep sequencing using the ion ampliseq cancer hotspot panel and droplet digital polymerase chain reaction[J]. Oncotarget, 2017, 8(63): 106901.

[58] Welch J S, Petti A A, Miller C A, et al. TP53 and decitabine in acute myeloid leukemia and myelodysplastic syndromes[J]. New England Journal of Medicine, 2016, 375(21): 2023-2036.

[59] Tano K, Akimitsu N. Long non-coding RNAs in cancer progression[J]. Frontiers in genetics, 2012, 3: 33822.

[60] Oh J M, Venters C C, Di C, et al. U1 snRNP regulates cancer cell migration and invasion in vitro[J]. Nature communications, 2020, 11(1): 1.

[61] Goodall G J, Wickramasinghe V O. RNA in cancer[J]. Nature Reviews Cancer, 2021, 21(1): 22-36.

[62] Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing[J]. Nature Reviews Genetics, 2010, 11(10): 685-696.

[63] Hong M, Tao S, Zhang L, et al. RNA sequencing: new technologies and applications in cancer research[J]. Journal of hematology & oncology, 2020, 13(1): 1-16.

[64] Cvekl A, Duncan M K. Genetic and epigenetic mechanisms of gene regulation during lens development[J]. Progress in retinal and eye research, 2007, 26(6): 555-597.

[65] Kulis M, Esteller M. DNA methylation and cancer[J]. Advances in genetics, 2010, 70: 27-56.

[66] Daskalos A, Nikolaidis G, Xinarianos G, et al. Hypomethylation of retrotransposable elements

correlates with genomic instability in non-small cell lung cancer[J]. International journal of cancer, 2009, 124(1): 81-87.

[67] Bau S, Schracke N, Kränzle M, et al. Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays[J]. Analytical and bioanalytical chemistry, 2009, 393: 171-175.

[68] Chakraborty S, Hosen M I, Ahmed M, et al. Onco-multi-OMICS approach: a new frontier in cancer research[J]. BioMed research international, 2018, 2018.

[69] Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine[J]. Cancer science, 2018, 109(3): 513-522.

[70] Bugter J M, Fenderico N, Maurice M M. Mutations and mechanisms of WNT pathway tumour suppressors in cancer[J]. Nature Reviews Cancer, 2021, 21(1): 5-21.

[71] Heeke A L, Pishvaian M J, Lynce F, et al. Prevalence of homologous recombination–related gene mutations across multiple cancer types[J]. JCO precision oncology, 2018, 2: 1-13.

[72] Ottaviani D, LeCain M, Sheer D. The role of microhomology in genomic structural variation[J]. Trends in Genetics, 2014, 30(3): 85-94.

[73] Hastings P J, Lupski J R, Rosenberg S M, et al. Mechanisms of change in gene copy number[J]. Nature Reviews Genetics, 2009, 10(8): 551-564.

[74] Li Y, Roberts N D, Wala J A, et al. Patterns of somatic structural variation in human cancer genomes[J]. Nature, 2020, 578(7793): 112-121.

[75] Okunade K S. Human papillomavirus and cervical cancer[J]. Journal of Obstetrics and Gynaecology, 2020, 40(5): 602-608.

[76] Farrell P J. Epstein–Barr virus and cancer[J]. Annual Review of Pathology: Mechanisms of Disease, 2019, 14: 29-53.

[77] Esteller M. Epigenetics provides a new generation of oncogenes and tumour-suppressor genes[J]. British journal of cancer, 2006, 94(2): 179-183.

[78] Parikh A R, Van Seventer E E, Siravegna G, et al. Minimal Residual Disease Detection using a Plasma-only Circulating Tumor DNA Assay in Patients with Colorectal CancerPlasma-only ctDNA-guided MRD Detection in Patients with CRC[J]. Clinical Cancer Research, 2021, 27(20): 5586-5594.

[79] Locke W J, Guanzon D, Ma C, et al. DNA methylation cancer biomarkers: translation to the clinic[J]. Frontiers in genetics, 2019, 10: 1150.

[80] Malla M, Loree J M, Kasi P M, et al. Using circulating tumor DNA in colorectal cancer: current and evolving practices[J]. Journal of Clinical Oncology, 2022, 40(24): 2846.

[81] Peng Y, Mei W, Ma K, et al. Circulating tumor DNA and minimal residual disease (MRD) in solid tumors: current horizons and future perspectives[J]. Frontiers in Oncology, 2021, 11: 763790.

[82] Lennon A M, Buchanan A H, Kinde I, et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention[J]. Science, 2020, 369(6499): eabb9601.

[83] Davis A A, Iams W T, Chan D, et al. Early assessment of molecular progression and response by whole-genome circulating tumor DNA in advanced solid tumors[J]. Molecular cancer therapeutics, 2020, 19(7): 1486-1496.

[84] Jee J, Lebow E S, Yeh R, et al. Overall survival with circulating tumor DNA-guided therapy in advanced non-small-cell lung cancer[J]. Nature Medicine, 2022, 28(11): 2353-2363.

[85] Lockwood W W, Chari R, Coe B P, et al. DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers[J]. Oncogene, 2008, 27(33): 4615-4624.

[86] Junttila T T, Parsons K, Olsson C, et al. Superior in vivo efficacy of afucosylated trastuzumab in the treatment of HER2-amplified breast cancer[J]. Cancer research, 2010, 70(11): 4481-4489.

[87] Talevich E, Shain A H, Botton T, et al. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing[J]. PLoS computational biology, 2016, 12(4): e1004873.

[88] Benjamini Y, Speed T P. Summarizing and correcting the GC content bias in high-throughput sequencing[J]. Nucleic acids research, 2012, 40(10): e72-e72.

[89] Zhao M, Wang Q, Wang Q, et al. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives[J]. BMC bioinformatics, 2013, 14(11): 1-16.

[90] Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer[J]. Nature biotechnology, 2012, 30(5): 413-421.

[91] Bao L, Pu M, Messer K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data[J]. Bioinformatics, 2014, 30(8): 1056-1063.

[92] Shen R, Seshan V E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing[J]. Nucleic acids research, 2016, 44(16): e131-e131.

[93] Favero F, Joshi T, Marquard A M, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data[J]. Annals of Oncology, 2015, 26(1): 64-70.

[94] Adalsteinsson V A, Ha G, Freeman S S, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors, 8(1): 1324.

[95] Lai D, Ha G, Shah S, et al. Package ‘HMMcopy’[J]. 2011.

[96] Zhou X, Cheng Z, Dong M, et al. Tumor fractions deciphered from circulating cell-free DNA methylation for cancer early diagnosis[J]. Nature Communications, 2022, 13(1): 7694.

[97] Newman A M, Bratman S V, To J, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage[J]. Nature medicine, 2014, 20(5): 548-554.

[98] Newman A M, Lovejoy A F, Klass D M, et al. Integrated digital error suppression for improved detection of circulating tumor DNA[J]. Nature biotechnology, 2016, 34(5): 547-555.

[99] Locallo A, Prandi D, Fedrizzi T, et al. TPES: tumor purity estimation from SNVs[J]. Bioinformatics, 2019, 35(21): 4433-4435.

[100] Poell J B, Mendeville M, Sie D, et al. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data[J]. Bioinformatics, 2019, 35(16): 2847-2849.

[101] Underhill H R, Kitzman J O, Hellwig S, et al. Fragment length of circulating tumor DNA[J]. PLoS genetics, 2016, 12(7): e1006162.

[102] Chabon J J, Hamilton E G, Kurtz D M, et al. Integrating genomic features for non-invasive early lung cancer detection[J]. Nature, 2020, 580(7802): 245-251.

[103] Lebofsky R, Decraene C, Bernard V, et al. Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types[J]. Molecular oncology, 2015, 9(4): 783-790.

[104] Pessoa L S, Heringer M, Ferrer V P. ctDNA as a cancer biomarker: A broad overview[J]. Critical reviews in oncology/hematology, 2020, 155: 103109.

[105] Mao X, Zhang Z, Zheng X, et al. Capture-based targeted ultradeep sequencing in paired tissue and plasma samples demonstrates differential subclonal ctDNA-releasing capability in advanced lung cancer[J]. Journal of Thoracic Oncology, 2017, 12(4): 663-672.

[106] Choi Y L, Soda M, Yamashita Y, et al. EML4-ALK mutations in lung cancer that confer resistance to ALK inhibitors[J]. New England Journal of Medicine, 2010, 363(18): 1734-1739.

[107] Lin J J, Riely G J, Shaw A T. Targeting ALK: precision medicine takes on drug resistance[J]. Cancer discovery, 2017, 7(2): 137-155.

[108] Pan Y, Deng C, Qiu Z, et al. The resistance mechanisms and treatment strategies for ALK-rearranged non-small cell lung cancer[J]. Frontiers in oncology, 2021, 11: 713530.

[109] Soda M, Choi Y L, Enomoto M, et al. Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer[J]. Nature, 2007, 448(7153): 561-566.

[110] Maia M C, Salgia M, Pal S K. Harnessing cell-free DNA: plasma circulating tumour DNA for liquid biopsy in genitourinary cancers[J]. Nature Reviews Urology, 2020, 17(5): 271-291.

[111] Angeles A K, Christopoulos P, Yuan Z, et al. Early identification of disease progression in ALK-rearranged lung cancer using circulating tumor DNA analysis[J]. NPJ Precision Oncology, 2021, 5(1): 100.

[112] Benner P, Vingron M. ModHMM: a modular supra-Bayesian genome segmentation method[J]. Journal of Computational Biology, 2020, 27(4): 442-457.

[113] Andrews S. FastQC: a quality control tool for high throughput sequence data[J]. 2010.

[114] Ewels P, Magnusson M, Lundin S, et al. MultiQC: summarize analysis results for multiple tools and samples in a single report[J]. Bioinformatics, 2016, 32(19): 3047-3048.

[115] Aird D, Ross M G, Chen W S, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries[J]. Genome biology, 2011, 12: 1-14.

[116] Hinrichs A S, Karolchik D, Baertsch R, et al. The UCSC genome browser database: update 2006[J]. Nucleic acids research, 2006, 34(suppl_1): D590-D598.

[117] Molparia B, Oliveira G, Wagner J L, et al. A feasibility study of colorectal cancer diagnosis via circulating tumor DNA derived CNV detection[J]. PloS one, 2018, 13(5): e0196826.

[118] Morrison C D, Liu P, Woloszynska-Read A, et al. Whole-genome sequencing identifies genomic heterogeneity at a nucleotide and chromosomal level in bladder cancer[J]. Proceedings of the National Academy of Sciences, 2014, 111(6): E672-E681.

[119] Polski A, Xu L, Prabakar R K, et al. Cell-free DNA tumor fraction in the aqueous humor is associated with therapeutic response in retinoblastoma patients[J]. Translational Vision Science & Technology, 2020, 9(10): 30-30.

[120] Kim K, Shin D G, Park M K, et al. Circulating cell-free DNA as a promising biomarker in patients with gastric cancer: diagnostic validity and significant reduction of cfDNA after surgical resection[J]. Annals of surgical treatment and research, 2014, 86(3): 136-142.

[121] Raman L, Dheedene A, De Smet M, et al. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing[J]. Nucleic acids research, 2019, 47(4): 1605-1614.

[122] Kleinheinz K, Bludau I, Hübschmann D, et al. ACEseq–allele specific copy number estimation from whole genome sequencing[J]. BioRXiv, 2017: 210807.

[123] Van Loo P, Nordgard S H, Lingjærde O C, et al. Allele-specific copy number analysis of tumors[J]. Proceedings of the National Academy of Sciences, 2010, 107(39): 16910-16915.

[124] Linacre J M, Wright B D. Facets[J]. Computer Program for Many-faceted Rasch Measurement, 2014: 1998.

[125] Brito P L, Dos Santos A F, Chweih H, et al. Reduced blood pressure in sickle cell disease is associated with decreased angiotensin converting enzyme (ACE) activity and is not modulated by ACE inhibition[J]. Plos one, 2022, 17(2): e0263424.

[126] Scheinin I, Sie D, Bengtsson H, et al. 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[J]. Genome research, 2014, 24(12): 2022-2032.

[127] Venkatraman E S, Olshen A B. A faster circular binary segmentation algorithm for the analysis of array CGH data[J]. Bioinformatics, 2007, 23(6): 657-663.

[128] Straver R, Sistermans E A, Holstege H, et al. WISECONDOR: detection of fetal aberrations from shallow sequencing maternal plasma based on a within-sample comparison scheme[J]. Nucleic acids research, 2014, 42(5): e31-e31.

[129] Bryzgunova O E, Konoshenko M Y, Laktionov P P. Concentration of cell-free DNA in different tumor types[J]. Expert Review of Molecular Diagnostics, 2021, 21(1): 63-75.

[130] Liu B, Morrison C D, Johnson C S, et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges[J]. Oncotarget, 2013, 4(11): 1868.

[131] Zare F, Dow M, Monteleone N, et al. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data[J]. BMC bioinformatics, 2017, 18: 1-13.

[132] Zare F, Hosny A, Nabavi S. Noise cancellation using total variation for copy number variation detection[J]. BMC bioinformatics, 2018, 19: 1-12.

[133] Oh S, Geistlinger L, Ramos M, et al. Reliable analysis of clinical tumor-only whole-exome sequencing data[J]. JCO Clinical Cancer Informatics, 2020, 4: 321-335.

[134] Haber D A, Velculescu V E. Blood-Based Analyses of Cancer: Circulating Tumor Cells and Circulating Tumor DNABlood-Based Analysis of Cancer[J]. Cancer discovery, 2014, 4(6): 650-661.

[135] Murtaza M, Dawson S J, Tsui D W Y, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA[J]. Nature, 2013, 497(7447): 108-112.

[136] Bettegowda C, Sausen M, Leary R J, et al. Detection of circulating tumor DNA in early-and late-stage human malignancies[J]. Science translational medicine, 2014, 6(224): 224ra24-224ra24.

[137] Jahr S, Hentze H, Englisch S, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells[J]. Cancer research, 2001, 61(4): 1659-1665.

[138] Lo Y M D, Chan K C A, Sun H, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus[J]. Science translational medicine, 2010, 2(61): 61ra91-61ra91.

[139] Mouliere F, Chandrananda D, Piskorz A M, et al. Enhanced detection of circulating tumor DNA by fragment size analysis[J]. Science translational medicine, 2018, 10(466): eaat4921.

[140] Danecek P, Bonfield J K, Liddle J, et al. Twelve years of SAMtools and BCFtools[J]. Gigascience, 2021, 10(2): giab008.

[141] Van Dijk E L, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias[J]. Experimental cell research, 2014, 322(1): 12-20.

[142] Peng H, Lu L, Zhou Z, et al. CNV detection from circulating tumor DNA in late stage non-small cell lung cancer patients[J]. Genes, 2019, 10(11): 926.

[143] Zviran A, Schulman R C, Shah M, et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring[J]. Nature medicine, 2020, 26(7): 1114-1124.

[144] Ardeshir-Larijani F, Bhateja P, Lipka M B, et al. KMT2D mutation is associated with poor prognosis in non–small-cell lung cancer[J]. Clinical lung cancer, 2018, 19(4): e489-e501.

[145] Ricciardi G R R, Russo A, Franchina T, et al. NSCLC and HER2: between lights and

shadows[J]. Journal of Thoracic Oncology, 2014, 9(12): 1750-1762.

[146] Lapin M, Oltedal S, Tjensvoll K, et al. Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer[J]. Journal of translational medicine, 2018, 16: 1-10.

[147] Chen E, Cario C L, Leong L, et al. Cell-free DNA concentration and fragment size as a biomarker for prostate cancer[J]. Scientific reports, 2021, 11(1): 5040.

[148] Jiang P, Chan C W M, Chan K C A, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients[J]. Proceedings of the National Academy of Sciences, 2015, 112(11): E1317-E1325.

[149] Tom, J.A., Reeder, J., Forrest, W.F. et al. Identifying and mitigating batch effects in whole genome sequencing data. BMC Bioinformatics 18, 351 (2017).

[150] Abyzov A, Urban A E, Snyder M, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing[J]. Genome research, 2011, 21(6): 974-984.

[151] Worst B C, van Tilburg C M, Balasubramanian G P, et al. Next-generation personalised medicine for high-risk paediatric cancer patients–The INFORM pilot study[J]. European journal of cancer, 2016, 65: 91-101.

[152] Mouliere F, Mair R, Chandrananda D, et al. Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients[J]. EMBO molecular medicine, 2018, 10(12): e9323.

[153] Van der Linden M, Raman L, Vander Trappen A, et al. Detection of copy number alterations by shallow whole-genome sequencing of formalin-fixed, paraffin-embedded tumor tissue[J]. Archives of Pathology & Laboratory Medicine, 2020, 144(8): 974-981.

[154] Molparia B, Nichani E, Torkamani A. Assessment of circulating copy number variant detection for cancer screening[J]. PloS one, 2017, 12(7): e0180647.

[155] Chae Y K, Davis A A, Jain S, et al. Concordance of genomic alterations by next-generation sequencing in tumor tissue versus circulating tumor DNA in breast cancer[J]. Molecular cancer therapeutics, 2017, 16(7): 1412-1420.

[156] Wang R, Xu Q, Guo H, et al. Concordance and Clinical Significance of Genomic Alterations in Progressive Tumor Tissue and Matched Circulating Tumor DNA in Aggressive-variant Prostate Cancer[J]. Cancer Research Communications, 2023, 3(11): 2221-2232.

[157] Filia A, Droop A, Harland M, et al. High-resolution copy number patterns from clinically relevant FFPE material[J]. Scientific Reports, 2019, 9(1): 8908.

[158] Teo S M, Pawitan Y, Ku C S, et al. Statistical challenges associated with detecting copy number variations with next-generation sequencing[J]. Bioinformatics, 2012, 28(21): 2711-2718.

[159] Kim J, Park W Y, Kim N K D, et al. Good laboratory standards for clinical next-generation sequencing cancer panel tests[J]. Journal of pathology and translational medicine, 2017, 51(3):

191-204.

[160] Roche Sequencing and Life Science. AVENIO ctDNA Analysis Kits: Performance Across Illumina Sequencing Platforms. Available from: https://marketplace.clinicalomics.com/wp-content/uploads/2020/06/White-Paper_-AVENIO-ctDNA-Analysis-Kits-Performance-on-Illumina-Seq-Platforms.pdf

[161] Cainap C, Balacescu O, Cainap S S, et al. Next generation sequencing technology in lung cancer diagnosis[J]. Biology, 2021, 10(9): 864.

[162] Dai Y, Liu P, He W, et al. Genomic Features of Solid Tumor Patients Harboring ALK/ROS1/NTRK Gene Fusions[J]. Frontiers in Oncology, 2022, 12: 813158.

[163] Fang W, Zhou H, Shen J, et al. MDM2/4 amplification predicts poor response to immune checkpoint inhibitors: a pan-cancer analysis[J]. ESMO open, 2020, 5(1).

[164] Zhang W, Kuang P, Liu T. Prognostic significance of CDKN2A/B deletions in acute lymphoblastic leukaemia: a meta-analysis[J]. Annals of Medicine, 2019, 51(1): 28-40.

[165] Van der Eecken K, Van der Linden M, Raman L, et al. Shallow whole-genome sequencing: a useful, easy to apply molecular technique for CNA detection on FFPE tumor tissue—a glioma-driven study[J]. Virchows Archiv, 2022: 1-10.

[166] Lee D H, Yoon H, Park S, et al. Urinary exosomal and cell-free DNA detects somatic mutation and copy number alteration in urothelial carcinoma of bladder[J]. Scientific reports, 2018, 8(1): 14707.

[167] Stankunaite R. Molecular profiling of cell free DNA in patients with paediatric solid tumours[J]. 2023.

[168] Coombes R C, Badman P D, Lozano-Kuehne J P, et al. Results of the phase IIa RADICAL trial of the FGFR inhibitor AZD4547 in endocrine resistant breast cancer[J]. Nature Communications, 2022, 13(1): 3246.

[169] Razavi P, Li B T, Brown D N, et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants[J]. Nature medicine, 2019, 25(12): 1928-1937.

[170] Sun J X, He Y, Sanford E, et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal[J]. PLoS computational biology, 2018, 14(2): e1005965.

[171] Ptashkin R N, Mandelker D L, Coombs C C, et al. Prevalence of clonal hematopoiesis mutations in tumor-only clinical genomic profiling of solid tumors[J]. JAMA oncology, 2018, 4(11): 1589-1593.

Acknowledgement

First of all, I would like to express my sincere gratitude to Prof. Dr. Matthias Schlesner. It was an honor to be one of his doctoral students. The door to Matthias' office was always open and I miss our fun and enlightening discussions. During my PhD, he gave me very good advice on my thesis and I learned a lot from them. It was his guidance that supported me in completing my studies.

In addition to my supervisor, I would like to thank the other members of my thesis committee: Prof. Dr. Benedikt Brors, Prof. Dr. Holger Sültmann and Prof. Dr. Carl Herrmann, who provided insightful comments and encouragement during my studies. I am especially grateful to the B240 team members who have given me a lot of help and encouragement during my work and life in Heidelberg.

Finally, I would like to thank my family and friends for their meticulous care and companionship. Without them, this achievement would not have been possible. Thanks to them for keeping me optimistic and positive during some difficult times.