Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctoral of Natural Sciences

Presented by:

Pitithat Puranachot, Master of Science

Born in: Ratchaburi, Thailand

Oral examination: 16th December 2021

Next-Generation Sequencing Analysis of Cell-Free DNA Identifies Actionable Alterations and Genomic Features in Pediatric Cancers

Referees:

Prof. Dr. Benedikt Brors

Prof. Dr. Stefan Fröhling

“ Because one seeks, not because one waits

Because of expertise, not by chance

Because one’s competent, not one’s blessed

Thus, one’s perseverance directs one’s destiny ”

Zhuge Liang (181 - 234)

ABSTRACT

Pediatric cancer is the third leading course of death among children and adolescents in the USA despite its low incident and high survival rate. Next-generation sequencing technologies allow the profiling of tumor genetics and the prediction of disease progression and response to therapies. However, tumor temporal and spatial heterogeneity could complicate the success of the selected therapy. Serial sampling of tumors at multiple time-points can accurately track the dynamics of clonal evolution during treatment. Multiple sampling of tumors at different locations can reveal all clonal genetic structures of the tumor. Nevertheless, both strategies might post discomfort or critical risk to the patient. Liquid biopsy has become an attractive strategy for obtaining tumor biomarkers non-invasively. Sequencing of cell-free DNA (cfDNA), DNA fragments in the liquid sample such as blood, has become a strategy to detect tumor-derived genetic markers known as circulating tumor DNA. Recently, cfDNA has been extensively evaluated its clinical value with different high-throughput sequencing technology in many adult cancers. Hence, cfDNA could also have a potential benefit to the management of pediatric cancer patients.

In this thesis, we developed bioinformatics workflows for analyzing cfDNA derived from an extensive group of pediatric cancer patients. The workflow aims to detect genetic alterations from three sequencing strategies, including low-coverage whole-genome sequencing (lcWGS), whole-exome sequencing (WES), and deep gene-panel sequencing (Panel-seq). The capabilities of detecting copy-number aberrations and point mutations have been compared between those strategies. We also compared the detectability of plasma cfDNA across tumor entities, including brain tumors, sarcomas, and other pediatric cancers. Sequencing strategy and tumor location have influences on the success of cfDNA in detecting tumor genetic alterations. An R package, cfdnakit, was developed to extract the length of cfDNA fragments and perform genome-wide fragment-length analysis using lcWGS dataset. The fragment-length analysis shows that the enrichment of short-fragment cfDNA is correlating with copy-number aberrations. In addition, this package calculates a comprehensive copy-number aberration (CPA) score that combines copy-number aberration and short-fragmented cfDNA ratio. This CPA-score is correlating with a higher level of ctDNA and could suggest the use of subsequent detection methods such as WES to detect actionable mutations with more sensitivity. Moreover, we applied TelomereHunter, a telomeric DNA analysis tool. It showed that telomeric DNA exists which opens an opportunity to detect telomeric aberration in plasma cfDNA. Analyzing plasma cfDNA of the pediatric cohort has shown the declining of telomere content. However, elongation and integration of telomeric variant repeats were found among brain tumor and sarcoma patients.

Finally, we demonstrated the utility of liquid biopsy cfDNA in the management of pediatric cancer. cfDNA reveals heterogeneous mutations possibly shred by tumor at metastasis site in a child with bilateral nephroblastoma. This finding supports the utility of cfDNA as a comprehensive source of genetic information derived from the tumor population in the body without invasive multiple tumor biopsies. In addition, we found that cfDNA can detect tumor temporal heterogeneity in several sarcoma patients through serial biopsy. This finding supports the idea of utilizing cfDNA to follow-up patients during the course of therapy.

ZUSAMMENFASSUNG

Pädiatrische Krebserkrankungen sind trotz ihrer niedrigen Inzidenz und ihrer hohen Überlebensrate die dritthäufigste Todesursache bei Kindern und Jugendlichen in den USA. Next-Generation Sequencing Technologien ermöglichen die Erstellung eines genetischen Profils, welches hilft Vorhersagen zum Krankheitsverlauf sowie zum Behandlungserfolg zu treffen. Die zeitliche und räumliche Heterogenität des Tumors könnte jedoch den Erfolg der gewählten Therapie erschweren. Zum einen können Proben von Tumoren zu mehreren Zeitpunkten helfen die Dynamik der klonalen Entwicklung während der Behandlung genau zu verfolgen. Zum anderen können durch mehrere Proben des Tumors an verschiedenen Tumorstellen alle klonalen genetischen Strukturen des Tumors aufgedeckt werden. Nichtsdestotrotz sind beide o.g. Ansätze für den Patienten mit Beschwerden oder einem beachtlichem Risiko verbunden. Ein Bluttest, die Liquid Biopsy (Flüssigbiopsie) hat sich zu einer attraktiven Strategie zur nicht-invasiven Gewinnung von Tumorbiomarkern entwickelt. Durch Sequenzierung von zellfreier DNA (cfDNA) aus der Flüssigbiopsie-Probe kann vom Tumor abgesonderte DNA, der sogenannten zirkulierenden Tumor-DNA (ctDNA), nachgewiesen werden. Vor kurzem wurde der klinische Nutzen von cfDNA mit verschiedenen Hochdurchsatz-Sequenzierungstechnologien bei vielen Krebserkrankungen bei Erwachsenen umfassend untersucht. Daher könnte cfDNA auch einen potenziellen Nutzen für die Behandlung von pädiatrischen Krebspatienten haben.

In dieser Dissertation wurden bioinformatische Workflows zur Analyse von cfDNA entwickelt, welche aus einer umfangreichen Gruppe von pädiatrischen Krebspatienten gewonnen wurde. Der Workflow hat zum Ziel, genetische Veränderungen anhand von drei Sequenzierungsstrategien zu erkennen, darunter low-coverage whole-genome sequencing (lcWGS), whole-exome sequencing (WES), sowie deep gene-panel sequencing (Panel-seq). Die Fähigkeiten zur Erkennung von Kopienzahlaberrationen und Punktmutationen wurden zwischen diesen Strategien verglichen. Ebenso wurde auch die Nachweisbarkeit von Plasma cfDNA bei verschiedenen Tumorentitäten, einschließlich Hirntumoren, Sarkomen und anderen pädiatrischen Krebsarten verglichen. Die Sequenzierungsstrategie und Tumorlokalisation beinflussen die Nachweisbarkeit der tumorgenen Veränderungen mittels cfDNA. Das R-Paket cfdnakit wurde entwickelt, um die Länge von cfDNA-Fragmenten zu extrahieren und eine genomweite Fragmentlängenanalyse mittels lcWGS Daten durchzuführen. Die Fragmentlängenanalyse zeigt, dass die Anreicherung von kurzfragmentiger cfDNA mit der Kopienzahlaberration korreliert. Darüber hinaus berechnet dieses Paket einen umfassenden Kopienzahlaberrations-Score (CPA), der die Kopienzahlaberration und den Gehalt von kurzfragmentierten cfDNAs kombiniert. Dieser CPA-Score korreliert mit einem höheren ctDNA Gehalt und könnte die Verwendung nachfolgender sensitiver Nachweismethoden wie WES unterstützen. Darüber hinaus haben wir TelomereHunter, ein Telomer DNA-Analysetool, angewendet. Es zeigte sich, dass telomere DNA als Plasma cfDNA vorhanden ist, was eine Möglichkeit eröffnet, Telomeraberrationen zu detektieren . Die Analyse der Plasma cfDNA der pädiatrischen Kohorte hat eine Abnahme des Telomergehalts gezeigt. Bei der cfDNA von Hirntumor- und Sarkompatienten war jedoch eine Verlängerung sowie Integration von Telomer-Variantenwiederholungen vorhanden.

Schließlich demonstrierten wir die Verwendung von Flüssigbiopsie cfDNA bei der Behandlung von pädiatrischem Krebs. CfDNA deutete auf heterogene Mutationen hin, die möglicherweise durch einen metastasierendem Tumor bei einem Kind mit bilateralem Nephroblastom absondert werden. Dieser Befund unterstützt den Nutzen von cfDNA als umfassende Quelle genetischer Information der Tumorpopulation ohne mehrfache invasive Tumorbiopsien. Auch konnte gezeigt werden, dass cfDNA bei mehreren Sarkompatienten durch longitudinale Flüssigbiopsien die zeitliche Heterogenität des Tumors erkennen kann. Dieser Befund stützt die Idee cfDNA zur Nachsorge von Patienten im Therapieverlauf einzusetzen.

TABLE OF CONTENTS

List of Figures

List of Tables

LIST OF ABBREVIATIONS

ALL	acute lymphocytic leukemia
ALT	alternative lengthening of telomeres
BAF	B-allele frequency
BAM	binary alignment map
CEA	carcinoembryonic antigen
cfDNA	cell-free DNA
CNS	central nervous system
CSF	cerebrospinal fluid
CBS	Circular Binary Segmentation
CTC	circulating tumor cells
ctDNA	circulating tumor DNA, tumor-derived cfDNA
CPA score	copy number profile abnormality score
CNV	copy-number variation
ELISA	enzyme-linked immunosorbent assay
EGFR	epidermal growth factor receptor
EpCAM	epithelial cell adhesion molecule
EMT	epithelial-to-mesenchymal transmission
FDA	Food and Drug Administration
Panel-seq	gene panel-sequencing
HSP	heat shock protein
HIV	human immunodeficiency virus
INFORM	Individualized Therapy for Relapsed Malignancies in Childhood
INDEL	insertion/deletion
LOD	limit of detection
LOH	loss of heterozygosity
lcWGS	low-coverage whole-genome sequencing
MHC	major histocompatibility complex
MAD	median absolute deviation
MSI	microsatellite instability
MI	molecular index
NTRK	neurotrophic tyrosine kinase
NGS	next-generation sequencing
PCAWG	Pan-Cancer Analysis of Whole Genomes
PoN	Panel-of-Normal
PDX	patient-derived xenograft

PCR	polymerase chain reaction
PCA	principal component analysis
PC	principal component
RME	embryonal rhabdomyosarcoma
S.L.Ratio	short/long-fragment ratios
SNV	single nucleotide variant
SCNA	somatic copy-number aberration
SV	structural variant
TVR	telomeric variant repeat
TPA	tissue polypeptide antigen
TCN	total copy-number
TRK	tropomyosin-related kinase
TCC	tumor cell content
TF	tumor fraction
TME	tumor microenvironment
TMB	tumor mutational burden
TKI	tyrosine kinase inhibitor
UID	unique identifier
UMI	unique molecular identifier
ucfDNA	urinary cell-free DNA
VAF	variant allele frequency
WES	whole-exome sequencing
WHO	World Health Organization

1 INTRODUCTION

1.1 Tumors in Childhood and Their Genomic Landscape

1.1.1 Global incidence, mortality, and survival rate

Pediatric cancer is the third leading cause of death among children and adolescents aged 0-19 years in the USA despite its low overall incidence [1]. Approximately 300,000 children were diagnosed with cancer worldwide every year during the past decade [2, 3]. The incidence of tumors is different between patient's age at diagnosis (Figure 1). Among children aged 0-14 years, the most common tumors were leukemias, followed by brain and central nervous system (CNS) tumors, lymphoma, and neuroblastoma [2]. In young adults between 15 and 19 years old, lymphomas were the most common cancer followed by epithelial tumors and melanoma, leukemias, germ cell tumors, and sarcomas [2].
The mortality rate of cancer in childhood was low in comparison to adult tumors. During 2001 – 2016, the leading cause of cancer death in children was leukemia (28.5%) followed by brain and other nervous systems (26.9%) and bones and joints tumor (9%) [4, 5]. The overall death rate of pediatric tumors among children and adolescents aged 0 to 19 was approximately 25 per million in the USA [4, 5]. The death rates declined by 1.5% on average every year during 2002-2016 particularly among pediatric leukemia and lymphoma since the availability of advanced treatment and supportive care [6]. However, the death rate of brain, bone, and soft-tissue cancer remained stable. In 2011, the brain tumor has replaced leukemia and became the leading cause of tumor death [4, 5].

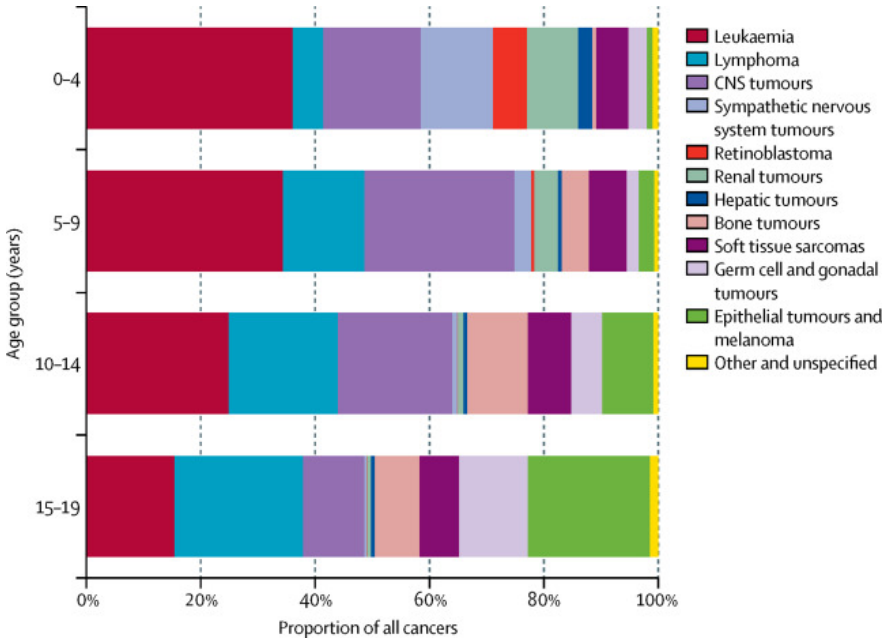


Figure 1: Global incidence of cancer in childhood (Reprinted from [2], Copyright © 2017 the World Health Organization, CC BY-NC-ND)

The overall survival rate of pediatric cancers was 83.5% during 2001 - 2015 [4, 5]. The improvement was significant among acute lymphocytic leukemia (ALL) and lymphoma, whereas stable among solid tissue tumors [4]. Minor improvements for pediatric brain tumors have been observed due to the development of neuroimaging, surgical technology, radiation technique, and supportive care. Soft tissue and bone cancers have no improvement in mortality and survival due to the lack of novel therapeutic agents and the limited development of existing agents during the past decade [7]. Overall, the important contribution toward the improvement of survival and decline in mortality rate has been related to accessibility to medical services where an early and accurate diagnosis is possible. Moreover, emerging innovative therapies and

palliative care reduce the late-effect of treatment and improve the outcome and quality of patient's life.

1.1.2 Improving outcomes of childhood tumor - early detection and accurate diagnosis

Early detection is a major key to improve outcomes of cancer care. The early identification of cancer leads to effective care that results in better survival, less intensive and suffering treatment [8, 9]. Late diagnosis leads to difficulty in having correct diagnosis due to complications and patients would suffer from late-effect from treatment. There are two early detection approaches previously described by the World Health Organization (WHO): (i) the recognition of symptomatic cancer in patients (early diagnosis); (ii) the identification of asymptomatic disease in a healthy target population (screening) [10]. Generally, it is not possible to screen for cancer in children because the cause of the majority of cancer in children is unknown [11]. Only very few cancers are caused by inherited genetic factors, environmental exposure, or chronic infections such as HIV, and hepatitis B [12, 13].
Early diagnosis is the most effective but requires awareness of warning symptoms by families and primary healthcare providers. Early and accurate clinical evaluation can help the medical doctor in deciding a specific treatment regimen that may include surgery, radiotherapy, and chemotherapy. The advance of high-throughput sequencing sheds light on personalized medicine and the development of new targeted agents. The genetic profile of the tumor allows the prediction of disease progression and response to therapies.

1.1.3 Genomic landscape of pediatric tumor

During past decade, comprehensive genomic studies have been focusing on cancers in adults, possibly due to their higher incidence, mortality rate and poor survival rate. They found that adult cancers usually developed multiple genetic alterations during life-time which together drive cancer progression. The genomes of adult cancers are mostly a mixture of small alterations of one or few of DNA bases, and larger structural alterations spanning more than 1,000 bases. The driver mutations are frequently shared across cancer types [14]. Recent pan-cancer genomic analyses have revealed the genomic landscape of tumors in children [15]. The results have increased our understanding of the genetic mechanisms that shape the genome cancer in children which is very essential for precision medicine.
The pediatric pan-cancer project has identified the major difference between genomes of pediatric cancer and adult cancer. A pediatric pan-cancer study analyzed nearly 400 whole-exome sequencings and 550 whole-genome sequencings across 24 tumor types, bias toward brain tumors, has reported a 14 times lower mutation rate than in adult cancers (Figure 2) [15]. The number of mutations significantly correlates with age — supporting the idea that cells accumulate mutations through a lifetime.

Second, childhood cancer is frequently driven by only a single cancer-driving mutation rather than multiple hits on cancer-driving genes. The driver mutations are likely preserved for specific cancer types. Half of the primary tumors harbor a potentially targetable genetic alteration. This finding emphasizes the need for personalized profiling to tailor more effective and less invasive therapies [15, 16]. Germline mutations, inherited from parents, have been identified as the causative factor in 7.6% of the cohort. Those germline mutations are enriched in DNA repair genes from mismatch (MSH2, MSH6, PMS2) and double-stranded break repair (TP53, BRACA2, CHEK2). Pediatric cancers are also characterized by a substantial degree of genomic instability which is strongly associated with somatics and germline TP53 mutations. Those unstable cancer genomes often display hyperploidy with a ploidy of four or more and are commonly found with chromothripsis.

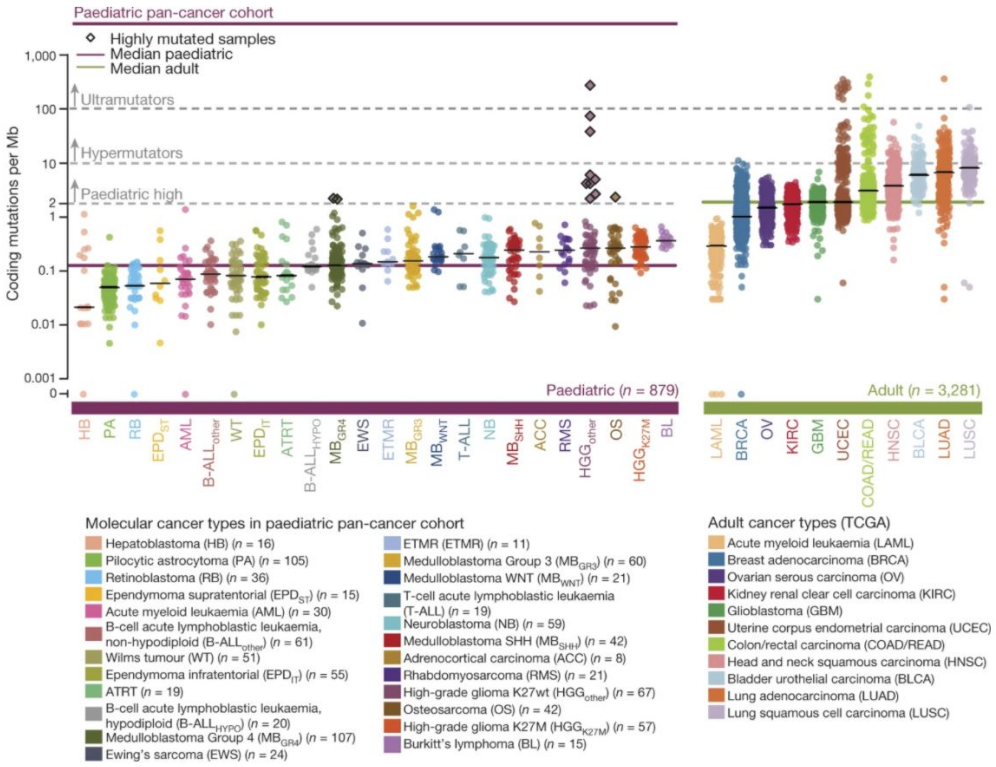


Figure 2: Somatic mutations in the pediatric pan-cancer cohort (Reprinted from [15], Copyright © 2018 by Macmillan Publishers Limited , CC BY)

1.2 Tumor Heterogeneity and Resisting Cell Death

Cancer is a complex disease. Multiple factors including germline genetic variations, somatic mutations, and environmental factors can dynamically shape the direction of evolution. This evolution supports the transformation of a non-malignant cell to a malignant cell through sequential mutations. Accumulation of mutations promotes the capabilities of self-sustaining proliferative signal, evading growth suppressors and cell death signals, induction of angiogenesis, and activation of tissue invasion and metastasis [17, 18]. These stochastic processes generate a genetically heterogeneous bulk of tumor where each cell harbors different molecular signatures. The difference in micro-environment and site-specific factors within and at different disease sites result in an uneven distribution of genetically diverse tumor subpopulations (spatial heterogeneity) (Figure 3A). Temporal heterogeneity refers to the genetic variation of a single tumor over time (Figure 3B). Heterogeneity within a bulk tumor result in different levels of sensitivity to cancer therapies. This section will point out tumor heterogeneity as a cause of tumor development against given therapy and resisting cell death.

1.2.1 Tumor temporal heterogeneity complicates the success of treatment

Both targeted therapies and nonspecific therapies apply dynamic selective pressure on the tumor population. This selective pressure influences the direction of clonal evolution depending on the administration schedule and specific choice of therapy. The resistant clone could emerge from the existing tumor population within 1-2 years during and after the treatment [19]. There are two mechanisms that drive resistance. Cells with resistant alterations are present at low allele frequency in the pretreatment tumor. This subpopulation could tolerate and expand under the therapeutic selective pressure (Figure 4A). Other findings support the alternative mechanism that cells could tolerate the therapy through adaptive activation of an alternative metabolic pathway, survival signals, and epigenetic programs. These cells

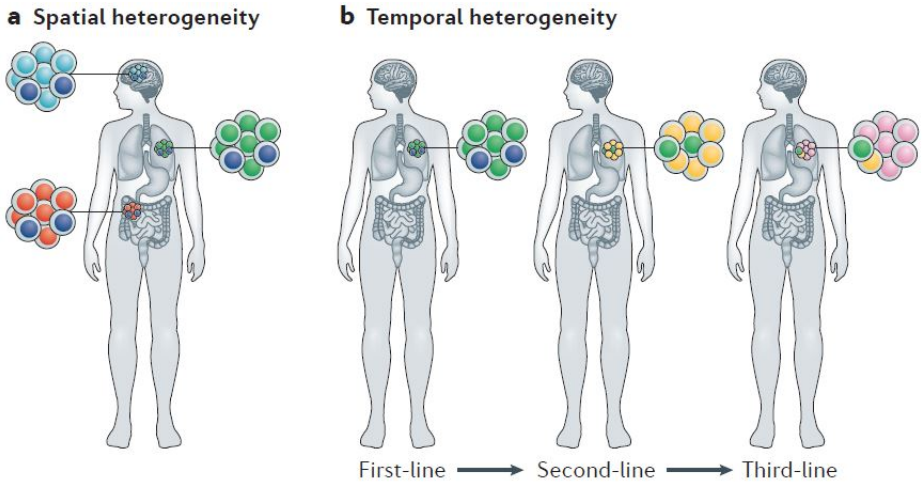


Figure 3: A conceptual framework of heterogeneity in tumor: Spatial heterogeneity (A) and Temporal heterogeneity (B) (Reprinted from [19], Copyright © 2018 by Macmillan Publishers Limited, permission from Copyright Clearance Center's RightLink® service)

acquire de novo resistant alterations that give resistance to therapy (Figure 4B). Given this temporal heterogeneity, interrogating a single genetic snapshot might not be efficient throughout the course of therapy.

Serial sampling of tumors at multiple timepoints is now the only approach to accurately track the dynamics of clonal evolution during the clinical course of treatment. Administration of targeted drugs can be adapted accordingly to the emergence, loss, and reappearance of expanding clones. For example, longitudinal sampling of a patient with adenocarcinoma harboring L858R EGFR and TP53 mutation has shown a dynamic change in clonal structure in response to administration of EGFR tyrosine kinase inhibitor (TKI) erlotinib [20]. The tumor had a substantial response during the first 8 months. A lung core biopsy reveals adenocarcinoma with the same L858R and p53 mutations, as well as an additional $EGFR^{T790M}$ TKI resistant mutation. T790M mutation could no longer be detected from the repeat biopsy after a 10-month interval of TKI withdrawal. The patient afterwards responded to erlotinib again to a therapeutic option that does not target T790M. This study demonstrates the clinical utility of repeat sampling for keeping track of clonal evolution and adjusting therapeutic administration. The development of sensitive technologies to support the early detection of a residual resistant clone is also necessary for the future era of precision medicine.

1.2.2 Resisting cell death through the telomere elongation

In tumor development, cancer cells require supportive mechanisms for unlimited replicative potential. Maintenance of telomere is one of the crucial processes that protect the ends of chromosomes from end-to-end fusions that leads to unstable dicentric chromosomes and finally cell mortality [18]. The maintenance process requires activation of the telomerase protein complex that plays important role in synthesizing telomeric DNA by the function of TERT reverse transcriptase and TERC RNA template. Genetic aberrations of TERT, including amplifications, rearrangements, or mutations in the promoter region are commonly found in human cancers [21]. Another mechanism known as the alternative lengthening of telomeres (ALT) pathway also supporting the telomere elongation by synthesizing telomeric DNA with different DNA recombination. The underlining mechanism of ALT remains unclear. Detection of ALT could indicate the inhibition of ALT as an anticancer treatment that causes cellular senescence [22].
Human telomeric DNA is typically 10–15 kb long and consists of non-coding repetitive sequences of

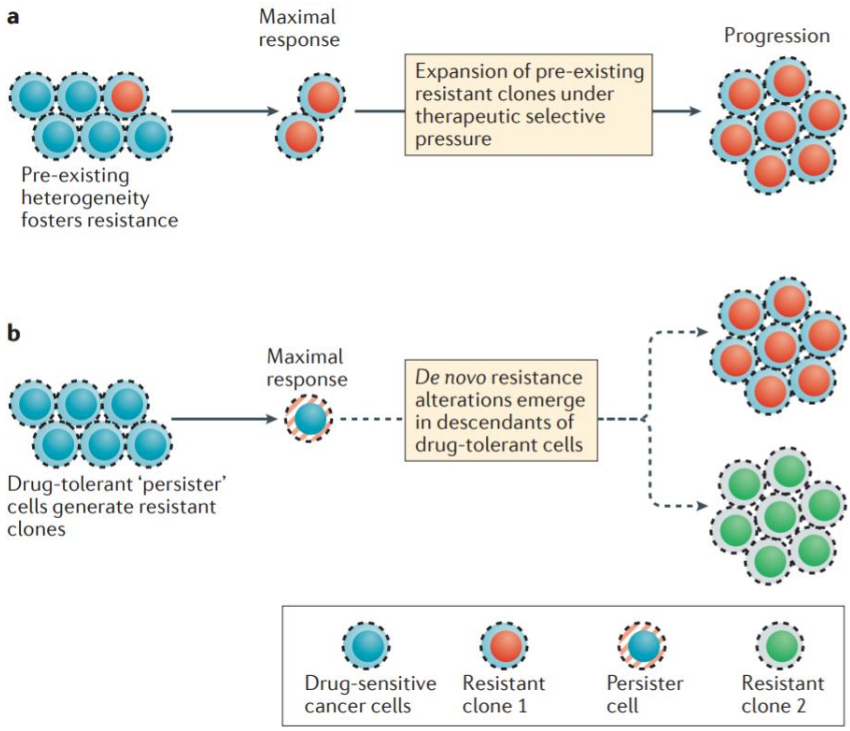


Figure 4: Two temporal evolutionary pathways that drive treatment resistance: A resistant clone (red) is pre-existing in the clone and expands beyond therapeutic selective pressure (A) or acquire de novo resistance alterations after surviving the pressure (B). (Reprinted from [19], Copyright © 2018 by Macmillan Publishers Limited, permission from Copyright Clearance Center's RightLink® service)

TTAGGG (t-type). However, telomeric variant repeats (TVRs) sequences namely TGAGGG (g-type), TCAGGG (c-type), and TTGGGG (j-type) also exist (Figure 5) [21, 23]. Telomeres of cells with ALT have heterogenous lengths and harbor recombination of TVRs [23]. In addition, extra-chromosomal telomeric repeats can exist in forms called C-circles [22, 24] which has been developed as a rapid, robust, and quantitative assay for ALT. Detection and quantification of telomere elongation and TVR has been demonstrated as a part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium [21].

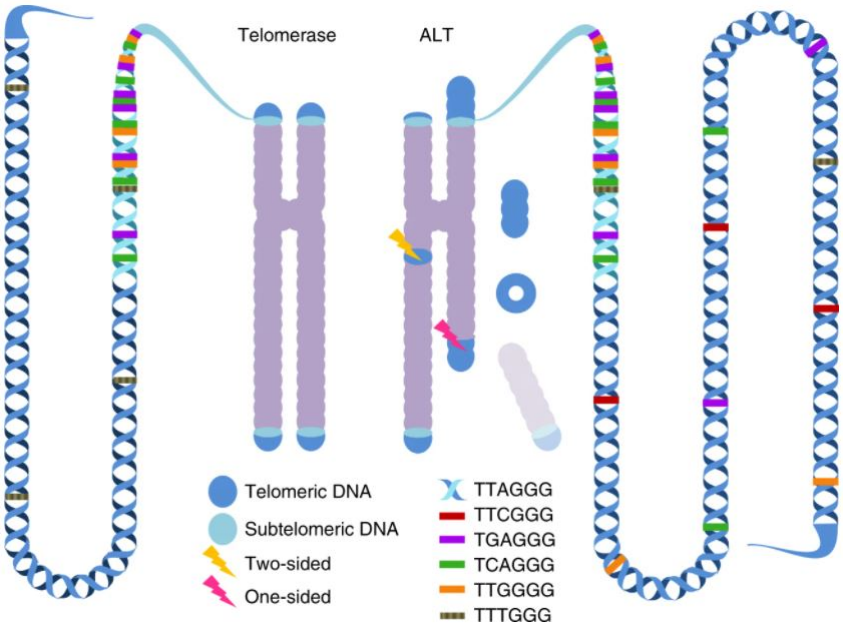


Figure 5: Genomic footprint of telomere elongation by telomerase and ALT (Reprinted from [21], Copyright © 2020 by the authors, CC BY)

1.3 Liquid Biopsy as a Non-invasive Approach to Track Tumor Progression

Due to the aforementioned tumor heterogeneity, a tumor can find an alternative direction to evolve and overcome environmental limitations or resist applied treatment. Taking multiple or serial biopsies seems to be the explicit solution. However, some limitations prohibit the routine tissue biopsy.

1. Multiple tumor biopsies from a patient cannot be always performed as a routine procedure. The patient would feel discomfort and suffer from the surgery. The surgery could be also complicated reaching the tumor site.
2. The procedure might increase the chance of tumor to seed onto other sites.
3. The derived sample from tissue biopsy might not represent the overall clonal structure of the tumor at a particular site. Only a single snap-shot of tumors is taken which ignores the adjacent tumor or at the remote site.

Liquid biopsy has become attractive over the past years as an alternative to derive information from patients non-invasively regarding pathological status. The fundamental objective is to detect a particular biomarker as a sign of the tumor in the body from the liquid samples (e.g. blood, saliva, or urine). Recently, the term "liquid biopsy" is covering the use of various biofluids, analyte materials, and biomarkers (Figure 6). Because it is easy to obtain a liquid biopsy, some of the liquid biopsy-based biomarkers have been used routinely after the completion of treatment as prognosis markers.

The following sections will describe liquid-based biomarkers that have been routinely obtained as tumor markers and emerging biopsy materials.

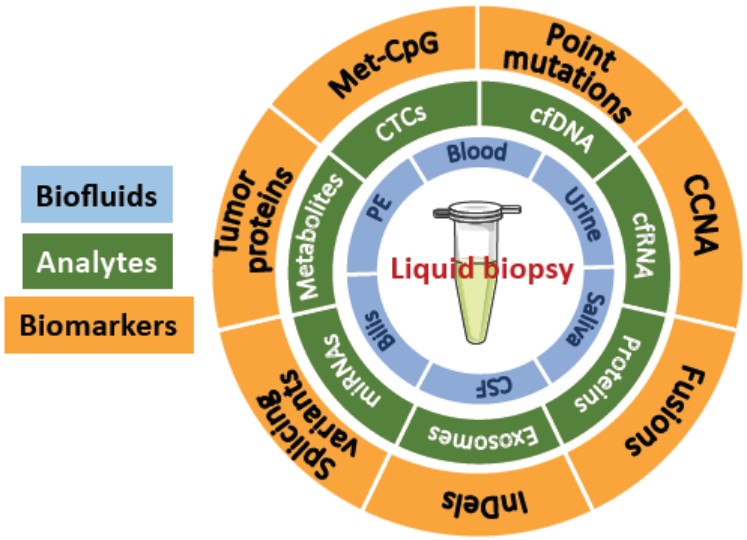


Figure 6: Biomarkers encompassed in liquid biopsy (Reprinted from [25], Copyright © 2020 by the authors, CC BY.)

1.3.1 Serum tumor markers

Serum-derived tumor markers are substances either released by tumor cells or other tissue in response to tumor indicating the existence of tumor in the body. They can be obtained non-invasively from blood, urine, stool, or other bodily fluid. Recent marker candidates, mostly proteins, antibodies, metabolites, and lipids, have shown potential for detecting a tumor in various clinical stages. The majority of these tumor markers were measured periodically after completion of curative treatment of primary tumor as a prognosis marker and to detect recurrence of the disease. For example, postoperative surveillance of asymptomatic women with breast cancer commonly measures the level of CA 15-3, carcinoembryonic antigen (CEA), tissue polypeptide antigen (TPA), tissue polypeptide-specific antigen, and HER2. Other serum-based tumor markers and their utilization have been introduced and reviewed [26].
Enzyme-linked immunosorbent assay (ELISA) is used as the gold standard method to detect serum tumor markers. This assay contains antibodies that bind specifically to targeted tumor antigens on a solid phase. The antibodies were designed to enzymatically react with specific substrates to produce a detectable signal (Figure 7). The general procedure of ELISA involves attaching one specific antigen on a solid well or with antibodies on the well surface. After immobilizing the antigen, the antibodies are added and form immunocomplexes with antigens. The antibody itself can be bound to an enzyme or to another secondary enzyme-conjugated antibody. In the final step, a designed substrate is added to produce detectable signals that can be detected by naked eyes or a spectrophotometer. The intensity of the signal indicates the concentration of tumor marker molecules. The ELISA-based technologies have been adapted extensively to improve its performance, customization and reduced operation cost [27, 28].

Although serum markers have been utilized in many clinical settings, there are limitations on their lack of specificity and sensitivity. Moreover, even though a particular serum marker is detected, it only indicates the existence of disease but lack of diagnostic value nor specify the tissue of origin.

1.3.2 Circulating tumor cells

Circulating tumor cells (CTCs) are a group of tumor cells that were shed from a primary tumor and circulate through blood circulation or the lymphovascular system. The first discovery was in 1869 when the Australian physician Thomas R. Ashworth observed cells with similar features of a tumor in the blood

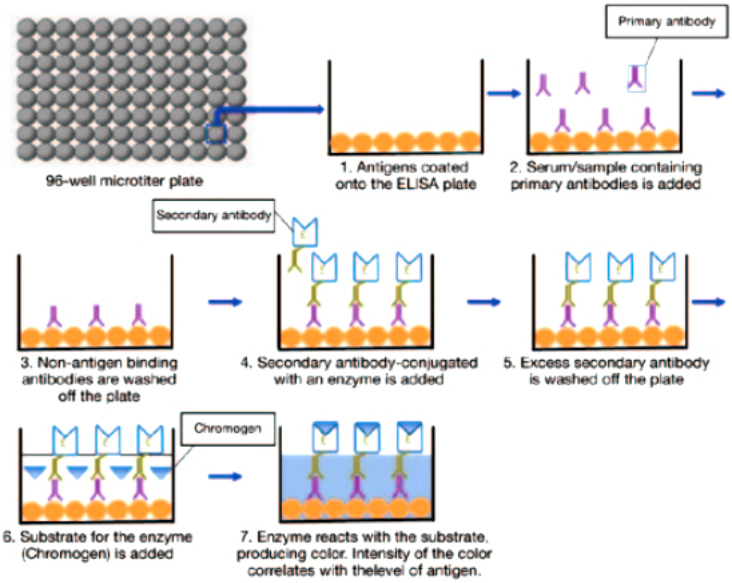


Figure 7: Process of Sandwich ELISA (Adapted from [28], Copyright © 2013 The Society for Investigative Dermatology, Inc., with permission from Elsevier)

of a man with metastatic cancer [29]. However, CTC had never been widely studied until the late 90s when Racila and colleagues first developed a sensitive assay combining immunomagnetic enrichment with flow cytometric methodology for CTC detection (Figure 8) [30]. Importantly, Racila and colleagues found that CTC were also present at the early stage , and they described the correlation between changes in the level of CTCs with both treatment and clinical status. The enrichment method distinguishes epithelial cells from mesenchymal blood cells by the expression of epithelial cell adhesion molecule (EpCAM) or cytokeratin proteins. Base-on this enrichment methodology, the CellSearch® [31] is the only detection and enumeration system approved by U.S. Food and Drug Administration (FDA) to date for monitoring cancer patients. The clinical utility has been demonstrated in advance and metastatic cancer such as lung cancer [32, 33], prostate cancer [34], ovarian cancer [35], and colorectal cancer [36].

The presence of CTCs in a patient's peripheral blood implies the intravasation of a population of tumor cells and the beginning of the metastatic event (Figure 9). CTC can go through epithelial-to-mesenchymal transmission (EMT) and be shedded into the bloodstream via active secretion from the primary tumor. Through EMT, cancerous epithelial cells lose their cell-to-cell adhesion and develop a mesenchymal-like phenotype. Those CTCs can be in the form of single cells or cell clusters which increase their metastatic potential. When reaching the distant site, CTCs transform back to their epithelial phenotype and grow into secondary metastasis. Despite the tumor's ability to secrete CTCs, only a small group of CTCs survive from trauma, oxidative stress, or evade from the immune system. The success of CTC to reach target distant sites depends on their survival mechanisms and influence factors [37]. Depending on the origin clone, CTCs are usually heterogeneous at the genetic, transcriptomic, proteomic, or metabolomic level making them a potential biomarker for deriving information regarding tumor heterogeneity and allow early detection of tumor metastasis.

The challenge of utilizing CTC is due to the low concentration of CTC. Usually, a sample of blood contains approximately 1 CTC per 1×10^6 blood cells with a half-life of less than 2.5 h [39]. This requires the development of a robust, reproducible, and sensitive assay to extract and maintain CTCs from a limited blood sample. Moreover, the FDA-approved platform, CellSearch®, was designed to separate CTCs

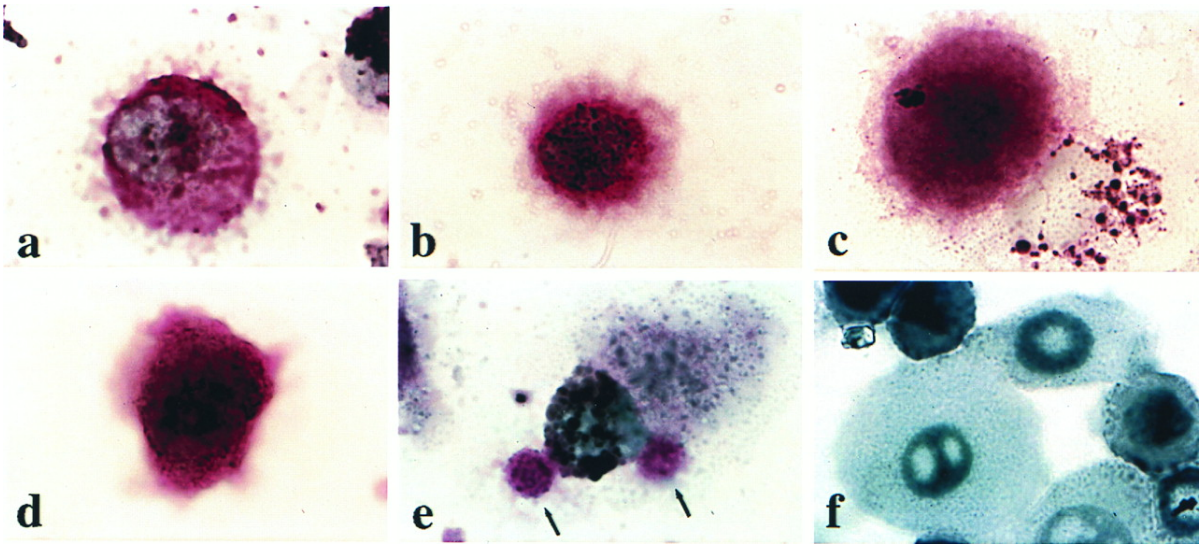


Figure 8: Circulating tumor cell detection with flow cytometric methodology in 1998: Detected circulating tumor cells (A-E) and normal epithelium cells (F) by flow cytometry and combined immunomagnetic enrichment. (A) Circulating tumor cells from a patient with metastatic breast cancer stained with anti-mucin-1. (B) Cells stained with anti-cytokeratin 5,6,8, and 18 from the same patient as *a*. Tumor cells stained with anti-cytokeratins from patient with breast tumor (C) and prostate cancer (D). (E) Two apoptotic tumor cells (arrows) stained with anti-cytokeratin and attached to a macrophage. (F) Normal epithelium obtained from human trypinized foreskined and stained with anti-mucin-1. (Reprinted from [30], Copyright © 1998 The National Academy of Sciences, CC BY-NC-ND)

with the expression of EpCAM from whole-blood cells, whereas CTCs without or low expression of EpCAM would be overlooked. Therefore, it is necessary to develop a method for enrichment, capture, and enumeration of CTCs incorporating other molecular or biophysical properties. Recently, many separation and enumeration methods have been develop and commercially available such as microfluidic chips [40, 41], size-based separation [42, 43], direct-imaging [44–46], and dielectrophoresis[47, 48]. The advantages and disadvantages have been reviewed in detail [38, 49].

1.3.3 Exosome

Exosomes are one of the extracellular vesicles that play important role in the cell-to-cell signal transduction of most eukaryotic cells. The size of exosome ranges from 40 to 160 nm in diameter (average 100 nm) [50]. An exosome is surrounded by a lipid bilayer membrane where inside contains biomolecules, including proteins, DNA, mRNA, non-coding RNA, and metabolites originated from the source cell (Figure 10A).

The basic protein component of the exosome includes a protein family of tetraspanins including CD9, CD63, CD81, CD82, CD106, Tspan8, and ICAM. Other non-specific protein families includes major histocompatibility complex (MHC), heat shock proteins (HSP), membrane fusion and transport proteins (annexins, Rab-GTPase), and cytoskeleton (actin,myosin, and tubulin) [50, 51]. Depending on the cellular origin and physiopathologic state, their actual composition is highly heterogeneous. According to ExoCarta [52], an exosome database (www.exocarta.org;accessed on 21 June 2021), exosome contains almost 10,000 proteins, 3,500 mRNAs, 3,000 miRNA and 1,000 lipids. These components in the exosome can be used as a prognosis marker for cancer progression.

Exosomes are also present in body fluid such as urine, serum, plasma, lymph, or cerebrospinal fluid from both cancer patients and healthy individuals. This makes it another potential non-invasive prognosis biomarker. Many exosomal circulating miRNAs have been related to tumor proliferation, transformation,

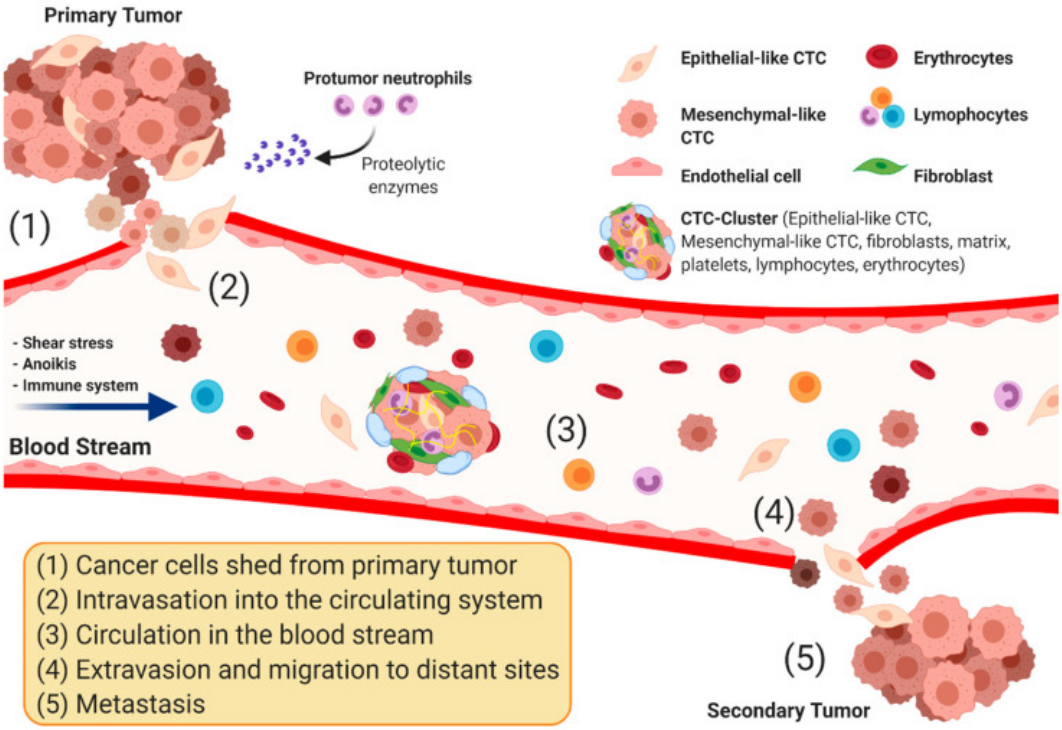


Figure 9: CTC dissemination from the primary tumor to distant sites via blood circulation (Reprinted from [38], Copyright © 2020 by the authors, CC BY)

angiogenesis, and resistance to therapy. Exosomes are important molecules that allow communication between growing tumor cells and surrounding cells in a tumor microenvironment (TME) (Figure 10B). TME is a mixture composition of extracellular matrix, blood vessels, tumor stem cells, tumor fibroblasts, stromal cells, signaling molecules, infiltrating inflammatory cells, and immune cells (T and B lymphocytes, dendritic cells, macrophages, and natural killer cells). The ability to protect those cellular contents from the phagocytic system make exosomes a good messager for cellular communication within TME. Detecting biomarkers from exosomes could be used for cancer early detection, early diagnosis, prognosis prediction, and therapeutic efficacy evaluation [50]. Moreover, engineered exosomes carrying tumor-suppressing proteins could provide new strategies for precise drug delivery in the era of precision medicine.

1.3.4 Cell-free DNA

Cell-free DNA (cfDNA) are extracellular double-stranded DNA fragments released by cells in the body into body fluid such as blood plasma, serum, cerebrospinal fluid, urine, and saliva [53]. The most commonly studied body fluids are blood plasma and urine whereas other liquids have been analyzed for specific type of tumor or disease. In general, the cfDNA fragments are relatively short (~167 bp) but larger fragment (>1 kb) could also be found [54]. The mechanism of cfDNA secretion is still unclear. Cell apoptotic process, in particular endonuclease activity, could be the source of short cfDNA. The length of plasma cfDNA fragments measured by sequencing technology shows a peak at 166-167 bp, which corresponds to the length of DNA wrapped around a nucleosome plus H1 histone linker protein. Nucleases cleaving process on the DNA strand at exposed sites with each turn of the DNA double helix leaves a 10bp ladder pattern on the fragment size trace of cfDNA. The longer fragment may be released by circulating tumor cells or exosome via necrosis (Figure 11). Recent studies has demonstrated that cfDNA carries dynamic information of cancer-specific genetic and epigenetic alterations [55]. The estimated half-life of cfDNA in blood circulation varies from a couple of minutes to 1-2 hours [56]. The short half-life of cfDNA facilitate the real-time analysis for evaluating treatment response and assessing

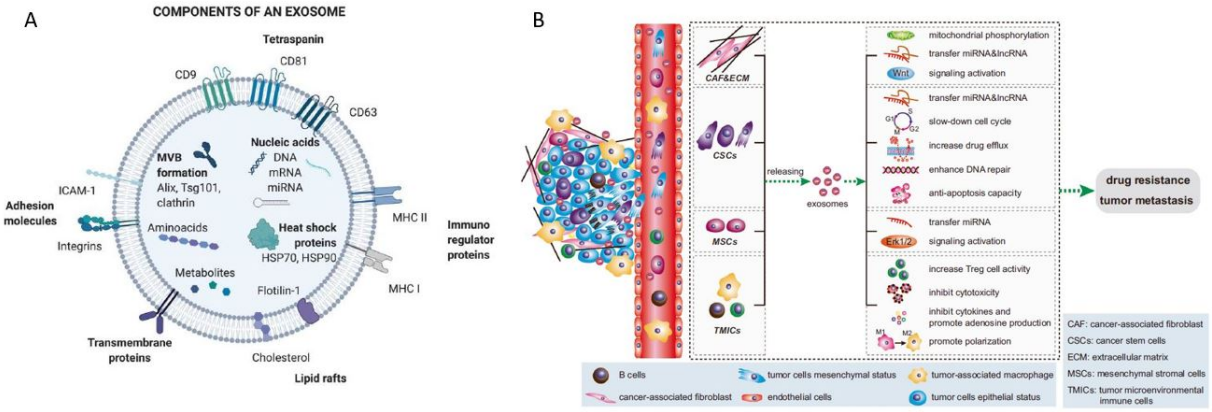


Figure 10: Components of an Exosome (A) (Adapted from [51], Copyright © 2021 by the authors, CC BY) and signal transduction pathway of TME via exosome (B) (Adapted from [50], Copyright © 2020 by the authors, CC BY)

status of tumor tissue.

cfDNA has been widely explored their clinical utilization as a prognostic or predictive marker, and ability to detect cancer [55]. Blood plasma of advance cancer patient contains much higher cfDNA concentration than healthy individuals [57, 58]. In cfDNA derived from a patient, DNA fragments originated from tumor tissue, termed circulating tumor DNA (ctDNA), can be detected via tracking tumor mutation. It is usually specific to tumor and could be used as a marker of tumor. The concentration of ctDNA was found elevated among patients with advanced or metastatic cancer [59]. It usually correlates with tumor stage [59], and response of tumor to the given therapy [55, 60]. With recent advance of high-throughput sequencing technique, cfDNA become an attractive candidate for a routine surveillance in cancer management. However, cfDNA has to be evaluated for its reliability and prognostic significance. Standardization of assay and finding validation has to be done in large-scale clinical trials.

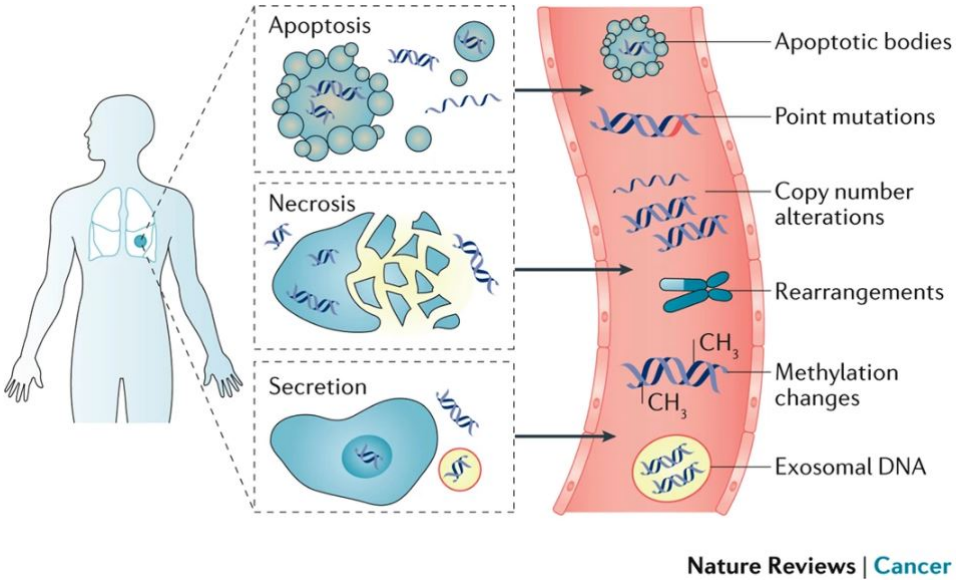


Figure 11: Source and genetic alterations in plasma cell-free DNA (Reprinted from [55], Copyright © 2017 by Macmillan Publishers Limited, permission from Copyright Clearance Center's RightLink® service)

1.4 Circulating Cell-free DNA

The basic information about cfDNA has been described in the previous section. This section contains more specific information about cfDNA including the history, possible source of cfDNA, biological properties, methodology for ctDNA detection, and application.

1.4.1 History of cell-free DNA

The history of cfDNA dates back to 1948 when Mendel and M´etais reported the discovery of nucleic acid in blood plasma [61]. They reported that extracellular DNA and RNA can be detected in the blood of humans without intention to be recently known as "liquid biopsy". This discovery had not gained much attention until 30 years later. The level of cfDNA was significantly increased in plasma of patients with systemic lupus erythematosus [62], and cancer [63]. They found that the concentration of serum cfDNA was higher in half of the cancer patients comparing to healthy individuals [63]. The concentration dropped when the patient positively response to radiation therapy and vice versa. An important discovery by Stroun and Anker in 1989 has demonstrated that cfDNA from the blood of patients contains DNA originated from tumor cells [64]. In the early 1990s, two independent studies were able to detect oncogene (KRAS and NRAS) point mutations in the plasma of patients with pancreatic cancer [65] and acute myelogenous leukemia [66]. Microsatellite instability and loss of heterozygosity (LOH) were found in the serum of patients with small-cell lung cancer [67] and head and neck cancer [68] in 1996. This discovery leads to the following development that supports advancements in liquid biopsy for non-invasive cancer detection (Figure 12). In 2016, FDA approved Cobas® EGFR Mutation Test for patients with non-small cell lung cancer [69]. High-throughput sequencing technology has become the main platform of DNA sequencing. Recently, cfDNA has been widely explored and clinically evaluated to support detection of both genetic and epigenetic alteration [69–71].

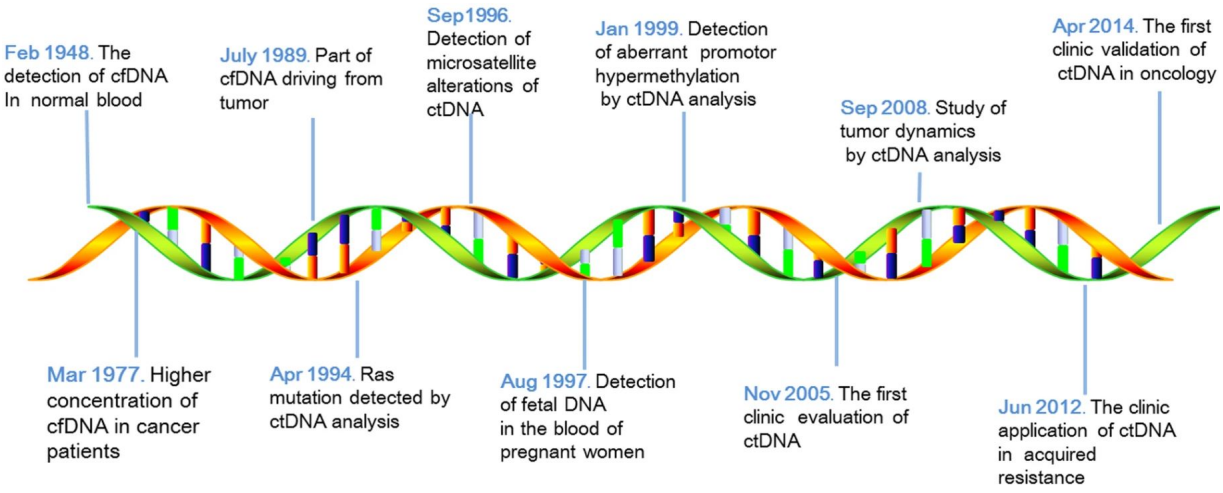


Figure 12: Timeline of cfDNA major research progression (Reprinted from [71], Copyright © 2019 by the authors, CC0 1.0)

1.4.2 Liquid sample of cell-free DNA

Cell-free DNA has been widely explored especially the potential source of cfDNA to be extracted from a patient. CfDNA extracted from different sources harbor unique contributions of cells of origin and provide a specific characteristic of DNA fragment (Figure 13). It has to be considered when planning

the implementation of cfDNA to surveillance on the tumor of interest and the selection of DNA isolation and quantification methodology.

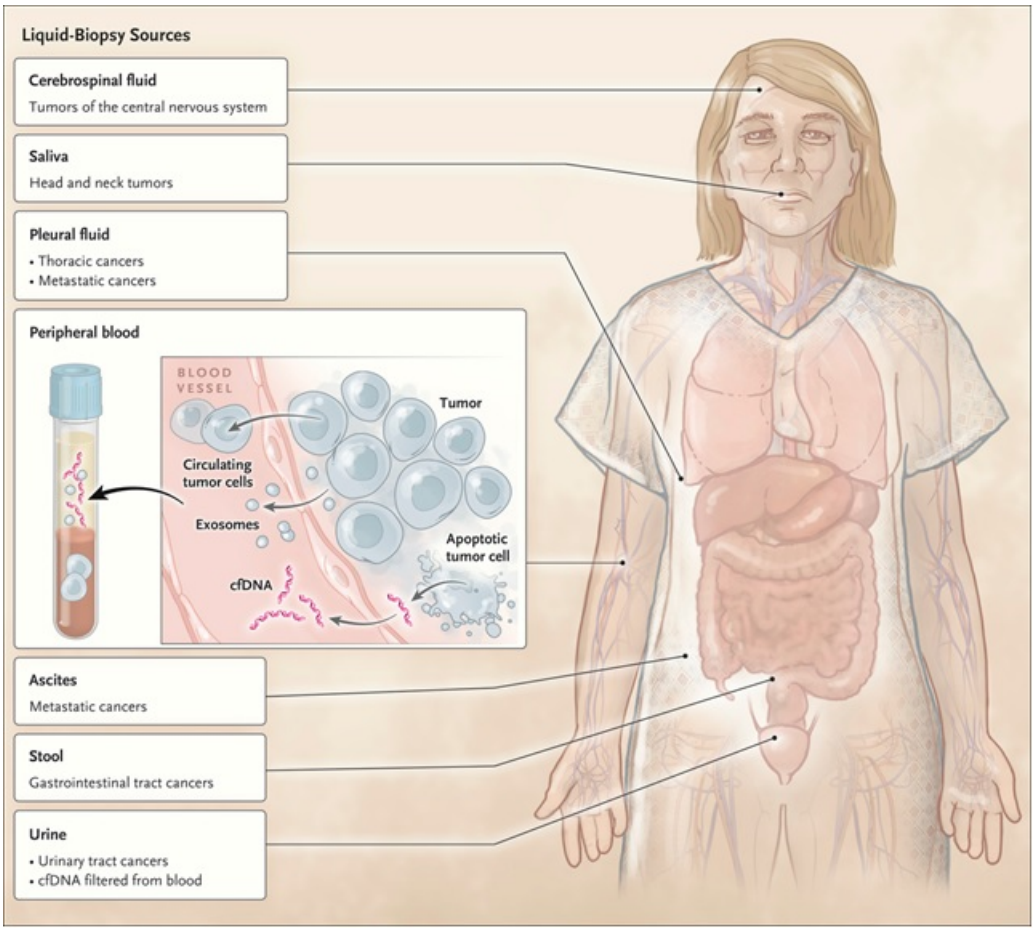


Figure 13: Overview of liquid sample of cell-free DNA (Reproduced with permission from [72], Copyright © 2018 by Massachusetts Medical Society)

Blood plasma (plasma cfDNA)/serum

Plasma cfDNA has been used as biomarkers in several medical areas such as non-invasive prenatal testing [73], inspecting of graft rejections after organ transplantations [74], and oncology. It has been widely explored for a decade. Studies during the past decade of plasma cfDNA has shown some basic properties and suggested their origin. Cells in the hematopoietic system are the major source of plasma cfDNA [54]. The fragment length distribution of plasma cfDNA shows a modal length of 167 bp with a 10 bp peak ladder suggesting apoptosis cells as its origin [54]. Necrotic cells, active secretion, and circulating tumor cells (CTC) also contribute high-molecular-weight DNA to the pool of plasma cfDNA [55]. Since the cfDNA fragment shows the pattern of DNA-binding onto nucleosome, many studies investigate patterns of plasma DNA fragmentation especially the preferred ending of fragment [75] and nucleosome positioning mapping [76]. Despite recent progression, the insight about the origin and the underlying mechanism still has to be further elucidated.

The mechanisms of cfDNA accumulation remain unclear. Concentration of plasma cfDNA varies between 0–1000 ng/ml in patients with cancer [58, 77] whereas approximately 200 ng/ml in healthy control [77]. A significant variation in the level of ctDNA has been observed among plasma cfDNA derived from patients with different tumor types [59]. CtDNA detection rate in patients with a primary tumor located

in the brain, renal, and thyroid was lower than those patients with advanced neuroblastoma, prostate, ovarian, colorectal, breast, and some other tumors [59]. This might be explained by the location of the primary tumor where particular mechanisms such as the blood-brain barrier or capsules block the release of ctDNA into blood circulation. Moreover, excessive physical activity, stroke, and infection also result in elevated concentrations of plasma cfDNA [78, 79]. Possibly concentration of cfDNA alone might not be an appropriate marker for cancer management. The success of utilizing cfDNA in clinical management could be improved by a better understanding of the basic biology of cfDNA and the underlying mechanisms of ctDNA.

Cerebrospinal fluid

Cerebrospinal fluid (CSF) is a clear body, colorless fluid that fills and baths the brain and spinal cord. It provides necessary nutrients and removes waste to maintain the central nervous system (CNS). CSF can be obtained through a minimally invasive procedure of the lumbar puncture which possesses some clinical risk and potential discomfort of the patient [80, 81]. The diagnostic lumbar puncture is performed routinely to evaluate CSF cytology for patients with CNS infectious disease, autoimmune encephalitis, and some tumors such as medulloblastoma. In a cancer patient, CSF cytology is used for diagnosis, tumor staging, and an indicator of response to therapy (Figure 14).

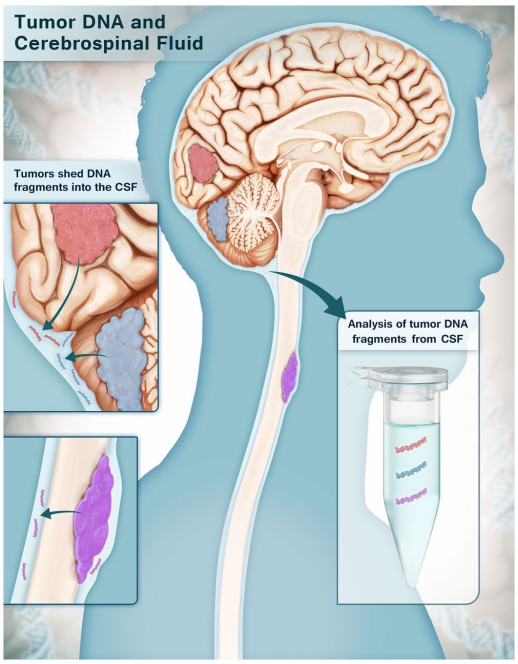


Figure 14: The shedding of DNA from central nervous system malignancies into cerebrospinal fluid (Reproduced with permission from [81], Copyright © 2015 by National Academy of Sciences)

As mentioned previously, blood plasma contains cfDNA derived from various tissue-of-origin especially hematologic cells. CNS, however, has a highly selective semipermeable border, termed the blood-brain barrier, that tightly regulates the transportation of molecules including cfDNA from peripheral blood into the extracellular fluid of the CNS and vice versa. A limited amount of cfDNA from CNS origin is released into the blood plasma. Therefore, blood plasma is not the best liquid solution for detecting cranial malignancies. Compared to blood, cfDNA in CSF has a lower background of normal DNA and contains a much higher proportion of tumor-derived cfDNA [82]. Li Y.S. and colleagues demonstrated that CSF liquid biopsy harbor EGFR mutation in patients with leptomeningeal metastases of the non-small-cell lung [82].

Other body fluids

Urine has been recognized as an important ultra-noninvasive sample source over tissue and blood to detect tumor markers from a patient with bladder cancer and prostate cancer [83, 84]. Extracellular DNA has long been found in urine [85]. There are two categories of urinary cell-free DNA (ucfDNA) depending on its origin: urinary tract cell DNA and transrenal DNA. Urinary tract cells DNA contains both high-molecular-weight ucfDNA, usually longer than 1 kbp released from necrotic cells along the urinary tract, and low-molecular-weight ucfDNA, 150-250 bp fragment originated from apoptotic cells and represent the majority of ucfDNA [86, 87]. Transrenal DNA refers to cell-free DNA in blood plasma that passes through the glomerular basement membrane in the kidney. The transrenal DNA is a low-molecular-weight fragment of size 150 - 160 bp, given that glomerular pores filtering out the large molecule with a diameter $>$ 11.5 nm. including nucleosomes, exosomes, apoptotic bodies, and large protein complexes. Since urinary tract cells have direct contact and their DNA is the majority of ucfDNA, it has great potential as a desirable source of diagnostic biomarkers for bladder cancer, prostate cancer, and renal cancer [88].

Pleural fluid is a common liquid material used in diagnosing cancers of the respiratory system. Many studies have demonstrated the feasibility of pleural effusion fluid in detecting EGFR mutation in patients with non-small cell lung cancer [89, 90]. It showed a potential of being a useful predictor of the gefitinib and erlotinib response. A study reported the high sensitivity (88%) and specificity (100%) of using pleural fluid cfDNA [91].

Ascites were reported to have abundant cell-free DNA and contained mutations in TP53, KRAS in patients with digest system cancer and gynecologic cancer [92]. Another preliminary study detected the presence of copy-number alterations in cancer-associated genes, especially in EGFR, in 6 metastatic cancer patients [93]. High molecular weight cfDNA was commonly found in ascites and indicate extracellular vesicles as the possible source [94].

Other body fluids such as sputum and saliva (for head and neck cancer, and oral cavity cancer), and stools (for colorectal cancer) are also a promising sources of cfDNA [81, 95].

1.4.3 Methodology/Technology for detecting circulating tumor DNA

At the early time of studies on cfDNA, polymerase chain reaction (PCR) was the main technology used for quantification of cfDNA and detection of alteration. Recently, next-generation sequencing has become cost-effective and demonstrated much utilization in the studies on cfDNA. One should consider the clinical situation and goal of ctDNA analysis in order to select which method would be suitable (Table 1). Briefly, the comprehensive approach does not rely on prior knowledge of hotspot mutation or genomic landscape of target tumor entity, while the targeted method can provide more sensitivity toward low-concentration of ctDNA.

Gene-panel deep sequencing

Although targeting a few genomic loci, gene-panel sequencing provides high specificity with a limit of detection at an allele frequency of 0.1. There are two approaches for sequencing a set of target genes: amplicon and hybridization-based sequencing.

The amplicon sequencing method is the most commonly used to detect point mutations in a set of target regions. This method uses PCR to amplify the targeted regions, called amplicon, and create multiplex of amplicon from different samples. If the target region is small (typically $<$ 50 genes), amplicon sequencing is more cost-effective, requires less material (10 - 100 ng) and lesser time than the hybridization-based method. However, the PCR bias of this method can lead to sequencing errors.

The hybridization-based method uses long, biotinylated oligonucleotide baits to capture the targeted

region. The hybridization-based methods are favorable when targeting larger regions (typically > 50 genes). In general, this approach provides better sensitivity (down to 1%) than amplicon sequencing (down to 5%) and enable detection all variant types including single nucleotide variants (SNVs), insertions/deletions (INDELs), and complex genomic alteration [96]. However, the hybridization method requires more input material (1-250 ng.) and a longer time to do purification steps.

Both sequencing methods have been frequently used in cfDNA studies. However, the additional advantage of the hybridization strategy is that it can combine with molecular barcodes which allow the reduction of sequencing error during the PCR process. Moreover, sequencing reads on off-target regions can be used for the detection of copy-number variations (CNVs). These advantages make the hybridization-based method a potential candidate for cfDNA investigation [97].

Whole-exome sequencing

Whole-exome sequencing (WES) provides a broader investigation of coding and non-coding regions of genes. It also allows the identification of genomic signatures such as tumor mutational burden (TMB) and microsatellite instability (MSI). Several studies performed WES on plasma cfDNA in detecting mutations and copy number alterations [98–100]. They demonstrated the longitudinal WES could be used to track tumor mutations during treatment or follow-up [98, 99]. Changes in the level of clonal and subclonal mutations could inform clinical about emerging resistant clones. However, the use of WES is limited by its sensitivity (limit of detection (LOD) >5%) and requires a relatively high amount of input material (>50 ng. required by Illumina Nextera Rapid Capture [101]). Many studies applied WES after a certain level of ctDNA is reached to effectively derive comprehensive mutation information and mutational signatures [98, 99, 102].

Test	Description	Detection Limit	Variant Detected	Advantages	Disadvantages	Cost
Allele-specific PCR	Amplification and quantification of pre-selected variants	0.1–1%	Well-defined SNVs and Indels	• Lower cost	• Small number of variants tested per sample • Lower sensitivity	$
Digital PCR	Amplication of pre-selected variants after partitioning into multiple reactions to increase sensitivity	0.01–0.1%	Well-defined SNVs and Indels	• High sensitivity • Lower cost	• Small number of variants tested per sample	$
Amplicon-based NGS	Deep sequencing of PCR amplicons	0.01–2%	SNVs and Indels	• High sensitivity • Less expensive than other NGS-based methods	• Fewer variants tested per sample than other NGS-based methods	$
Capture-based NGS	Deep sequencing of hybrid captured DNA molecules	0.00025–0.01%	SNVs, Indels, SCNAs, and recurrent SVs	• Highest sensitivity • Broadly applicable	• Less comprehensive than whole exome and genome NGS	$$ - $$$
Whole Exome NGS	Deep sequencing the exome	5–10%	SNVs, Indels, SCNAs, and SVs	• Entire exome analyzed • Broadly applicable	• Expensive • Low sensitivity	$$$$
Whole Genome NGS	Deep sequencing of the genome	1–10%	SNVs, Indels, SCNAs, and SVs	• Entire genome analyzed • Broadly applicable	• Expensive • Low sensitivity	$$ - $$$$

Table 1: cfDNA PCR and sequencing methodologies in comparison: somatic copy-number aberration (SCNA); structural variant (SV) (Adapted from [96] ,)

Low-coverage whole-genome sequencing

Instead of getting sequencing coverage at 10-30X, low-coverage whole-genome sequencing (lcWGS) offers an affordable approach to derive genome sequence at shallow coverage ~0.5-2X. It can be performed instantly using a few input DNA materials (>1 ng.). lcWGS can discover genetic alterations without prior knowledge of the genetic makeup of the tumor and is not limited to a specific set of regions. This ability come in useful because most of the late-stage tumor evolve rapidly as a result of progression and the selective pressure of treatment. Moreover, the majority of solid tumors and 50% of blood-related cancer harbor aneuploidy and aberrated copy-number profile. Bioinformatics workflows can use lcWGS data to investigate genome-wide copy-number profiles, estimate the tumor fraction and extract characteristics of cfDNA fragments. Recently, lcWGS has been performed in many studies and shows a great presentation of genome-wide copy-number profiles from plasma DNA samples. Moreover, longitudinal lcWGS has been recognized as a cost-effective tool in tracking tumor relapse during follow-up and revealing the copy-number profile of the therapeutical-resistant clone. However, the sensitivity of this method is limited to reliable detection of ctDNA at 5-10% [96].

Error rate reduction

The limitation of the next-generation sequencing (NGS) method is due to the high error rate of both the PCR and sequencing process. Theoretically, a true mutation is called only when the frequency of the mutation is higher than a read error rate. The limit of detection of 0.01% can be achieved with 100,000 region supporting reads given the error rate is below 0.01% and 5,000 genomic equivalence. The early NGS and PCR-based genotyping technique cannot reliably detect alleles less than 5% [96]. Several techniques were introduced aiming to reduce the error rate. One potential technique is so-called molecular barcoding strategies [103]. The molecular barcode has been known as unique molecular identifiers (UMI), or unique identifiers (UID). They are designed as a random sequence of 6-8 nucleotides to be assigned to each DNA molecule during PCR. At the end of the process, the bioinformatics approach could reidentify the sequence of template molecules based on consensus reads having identical UMI and mapping genomic location (Figure 15) [104–106]. Implementing UMI with deep panel-sequencing can reduce PCR biases and sequencing errors, improve accuracy in the detection of low-allele frequency mutation in cfDNA [97, 107].

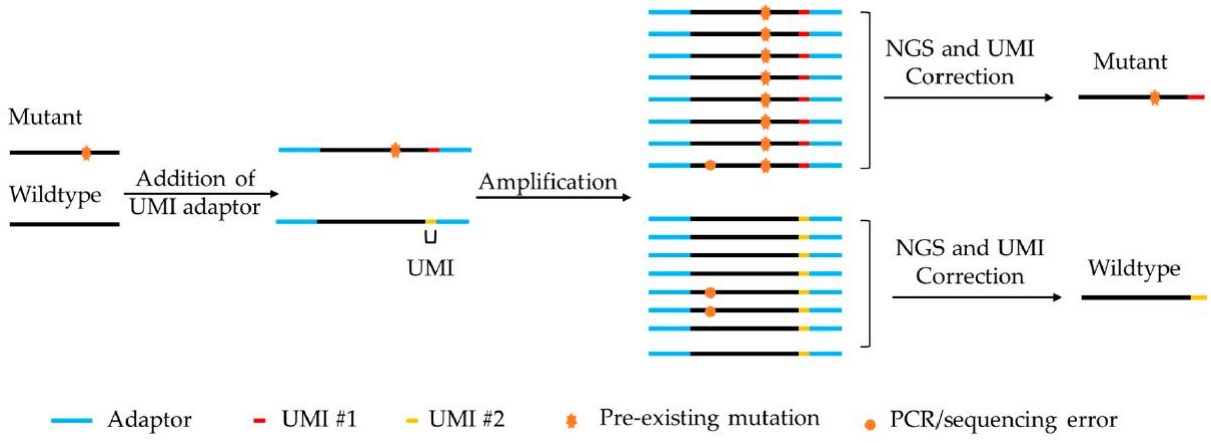


Figure 15: Simplified schematic of somatic mutations calling with application of unique molecular identifiers (UMI) (Reprinted from [106], Copyright © 2019 by the authors, CC BY)

1.5 Characteristical Length of CfDNA Inferring Tumor-origin Plasma CfDNA

Cell-free DNA in blood plasma consists of a pool of DNA fragments released into the blood circulation from various cell types in the body. Differentiating tumor-derived DNA (or ctDNA) from non-tumor DNA required insight on which characteristic of cfDNA could be used as a marker. Many studies have discovered genetic and nongenetic signatures of cfDNA that could infer the origin of cfDNA, for example, methylation, nucleosomal footprint, end-motif sequence, and length of the fragment. This dissertation will investigate the characteristic of cfDNA focusing on the length of the cfDNA fragment. This section describes the underlining mechanism relating to fragmentation of cfDNA and what is the difference between ctDNA and non-malignant cfDNA.

1.5.1 The source of cfDNA determines characteristical length of plasma cfDNA

It has been long discovered that cfDNA fragments are generated by a non-random process. Blood plasma contains a mixture of cfDNA fragments of different sizes where the majority of fragments are short (<200bp). In plasma of healthy individuals, the fragment length distribution of cfDNA shows a dominant peak is ~167 bp. which corresponding to the length of a DNA fragment wrapping around a molecule of mononucleosome (143 bp.) plus an H1 linker protein (~10.4 bp.) (Figure 16a). Within the 100-160 bp range, a characteristic 10-bp periodic peak is observed which is possibly the result of cleavage on the grooves of DNA that is exposed to nuclease. This common finding suggests that plasma cfDNA was secreted via cell apoptosis into blood circulation as a DNA bound to the histone protein. It is often known as "circulating nucleosomes". Recent studies reveal that the fragmentation process involves several endonuclease activities. Inside apoptotic cells, chromatins are digested by DFFB (DNA fragmentation factor sub-unit β) and DNASE1L3 (deoxyribonuclease 1-like3) as a part of cell death program (Figure 16b). Cleaved DNA-nucleosome complex is secreted together with DNASE1L3 and DNASE1 (deoxyribonuclease 1) into extracellular fluid where additional fragmentation is performed. Therefore, the chromatin structure of the original cell would influence the length of the cfDNA fragment. Open chromatin regions would be secreted as highly fragmented cfDNA whereas cfDNA from closed chromatin regions are mostly intact (Figure 16c).

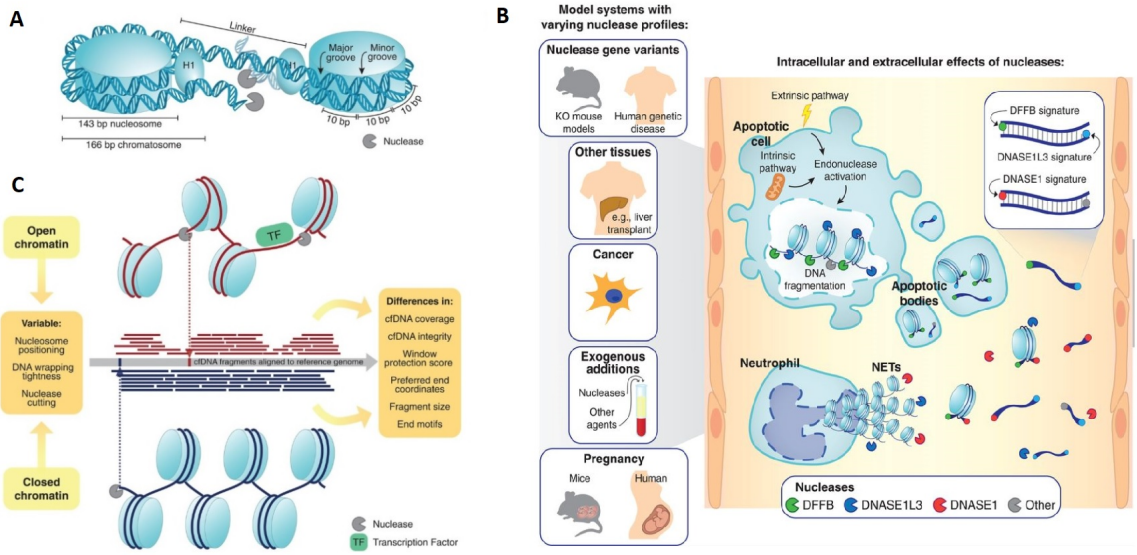


Figure 16: Source and chromatin structure incluence length of cfDNA fragment (Adapted from [72], Copyright © 2021 by the authors with permission from AAAS)

1.5.2 Tumor-derived cfDNA is shorter than non-malignant-origin cfDNA

Since the advance of next-generation sequencing technology, the length of individual cfDNA molecules can be accurately measured in many areas of research. In the plasma of pregnant women, cfDNA derived from the fetus (originated from the placenta) has has been shown to be shorter than cfDNA from the mother. Quantification of short-fragment cfDNA in pregnant women could benefit in quantification of fetal DNA and detect chromosomal aneuploidies of the fetus. A similar phenomenon is observed in patients who receive organ transplantation. Graft-derived cfDNA are shorter than recipient-derived cfDNA and enrichment of short cfDNA indicate the graft-rejection [108, 109]. In patients diagnosed with cancer, enrichment of short cfDNA has been observed in many tumor entities, and correlate with pathological status.

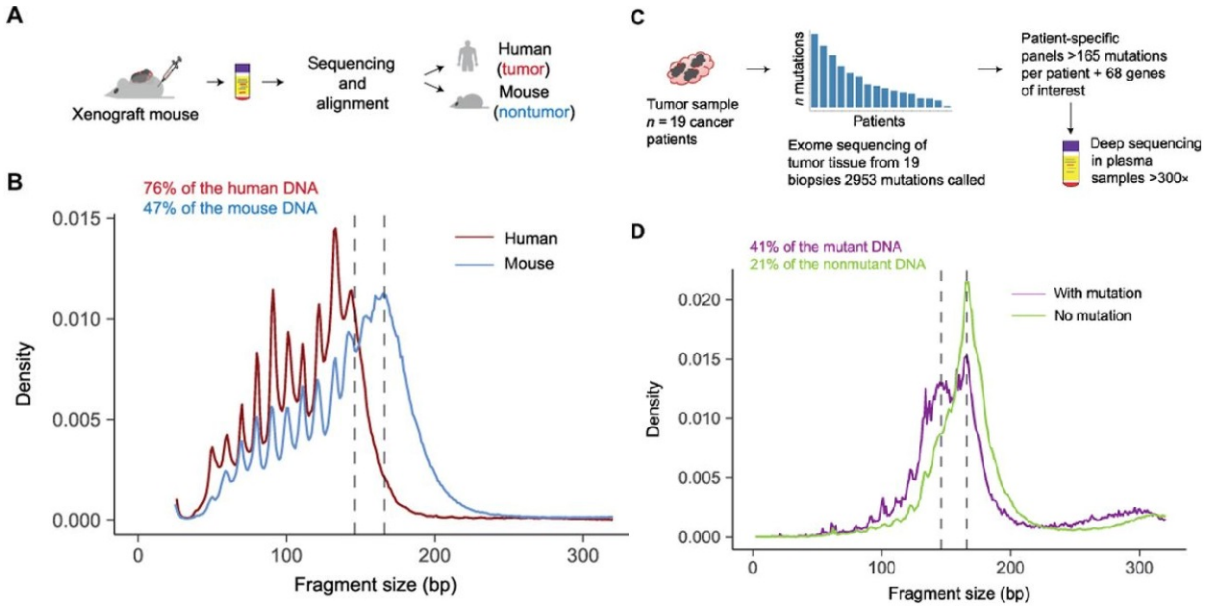


Figure 17: The size profile of mutant ctDNA with animal models and personalized capture sequencing (Reprinted from [110], Copyright © 2018 by the authors with permission from AAAS)

In 2018, Florent Mouliere and colleagues published a comprehensive study demonstrating that tumor-derived cfDNA is shorter than cfDNA from non-malignant cells. An experiment of a xenografted human ovarian cancer was performed in a mouse model in which cfDNA was extracted (Figure 17A). The extracted cfDNA were sequenced and their origin whether were identified via sequence alignment (align onto the human reference genome or mouse reference genome). The length of tumor-derived cfDNA (human cfDNA) was enriched in the range between 90 and 150 bp, while non-tumor cfDNA (mouse cfDNA) is dominated by fragments longer than 150 bp and peaked at 166 bp (Figure 17B). Similar findings were also found in other xenografted human cancers [111–113]. Second, tumor mutations identified by whole-exome sequencing of tumor DNA were used as a patient-specific panel for deep sequencing (>300 depth of coverage) of matched cfDNA samples (Figure 17C). The size profiles of detected ctDNA in 19 patients with cancer were analyzed. cfDNA fragments that harbor tumor alleles were enriched in fragments ~20 and 40 bp shorter than the length of DNA-mononucleosome and dinucleosome complex (Figure 17D). This study finds that circulating tumor DNA consists of highly fragmented DNA between the length of 90 and 150 bp, and 250 to 320 bp. They also survey fragment length of 344 plasma samples derived at late-stage in a pan-cancer study and 65 healthy controls (Figure 18A). It shows a significant difference in the proportion of short-fragment cfDNA between samples with high ctDNA and samples

from healthy individuals (Figure 18B). cfDNA samples from late-stage melanoma, breast, ovarian, lung, colorectal, and cholangiocarcinoma show enrichment of short-fragment cfDNA when comparing to other tumor entities and healthy individuals (Figure 18C).

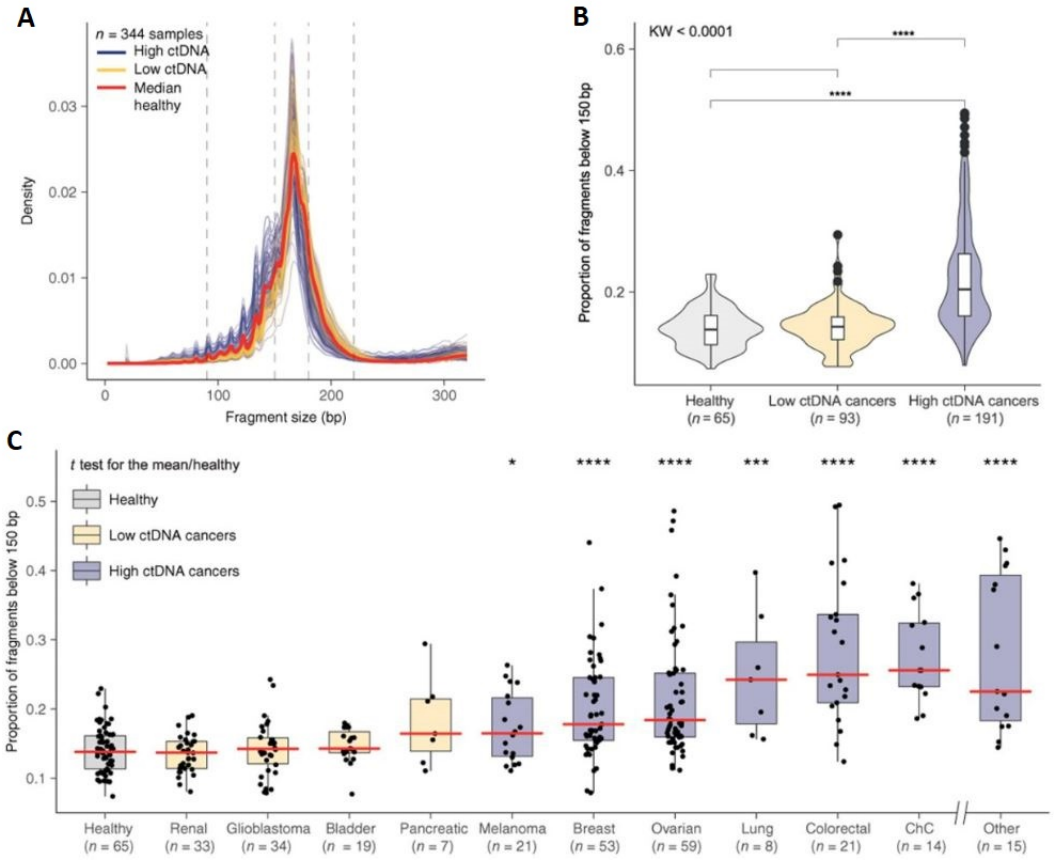


Figure 18: A survey of plasma DNA fragmentation on a pan-cancer scale (Reprinted from [110], Copyright © 2018 by the authors with permission from AAAS)

1.5.3 Size-selection enhances detection of circulating tumor DNA

The finding that circulating tumor DNA is shorter than non-tumor cfDNA has been discussed during the past decade and comprehensively demonstrated by Florent Mouliere and colleagues [110]. This study is also the first study that presents the utility of size-selection strategy, both in vitro and in silico, (Figure 19A) and quantitatively assesses its impact on detecting tumor alteration in plasma cfDNA. In vitro size-selection used a bench-top microfluidic device to select fragments with a particular size. In silico size-selection, fragment length is inferred from the mapping distance between the beginning and the end of a mapped paired-read. Both methods can filter cfDNA with the length between 90 to 150 bp (Figure 19B).

The effect of size-selection in detecting somatic copy number alterations (SCNAs) has been determined in plasma cfDNA samples derived from a group of patients with high-grade serous ovarian cancer. They identified cfDNA at pretreatment with a high concentration of ctDNA where many SCNAs were detected (Figure 19C). Without size-selection, a few SCNAs were detected in the posttreatment sample derived

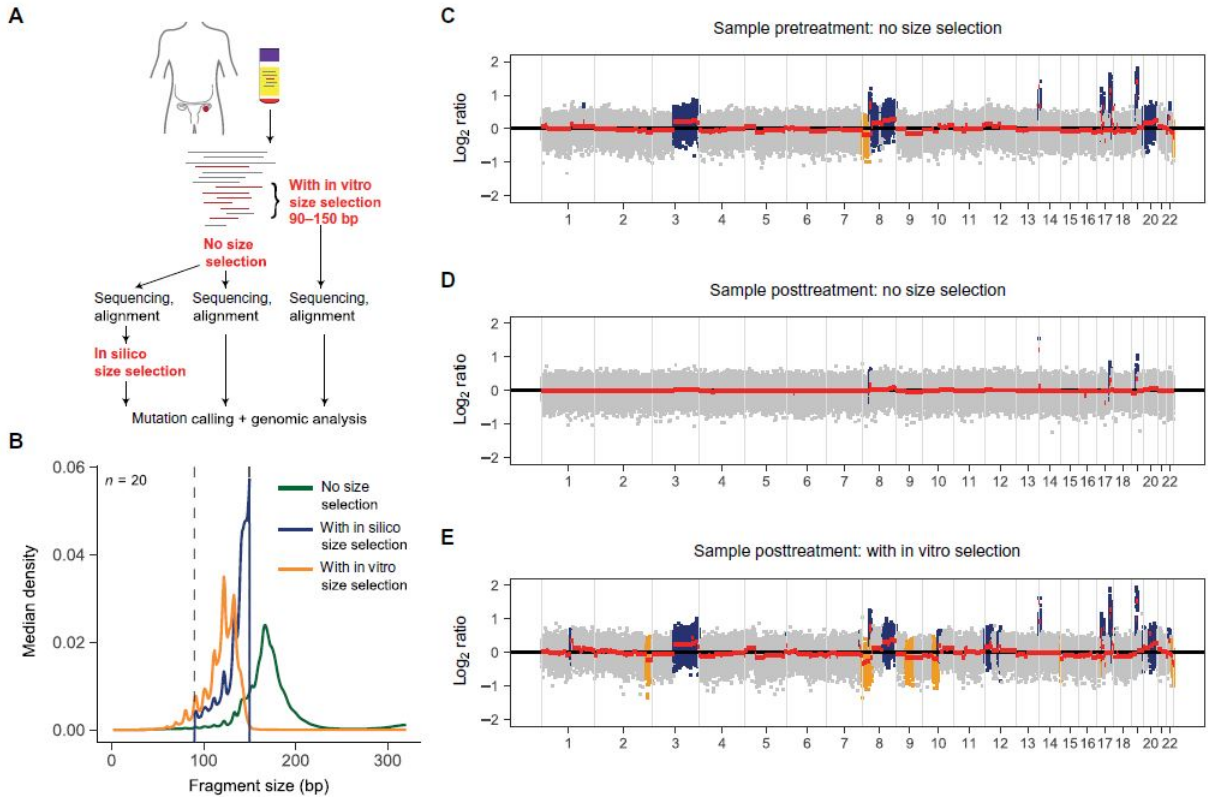


Figure 19: Enhancing the tumor fraction from plasma sequencing with size selection (Reprinted from [110], Copyright © 2018 by the authors with permission from AAAS)

3 weeks after the beginning of chemotherapy (Figure 19D). It is possibly due to the low concentration of ctDNA. When applying in vitro size-selection on the posttreatment sample, amplitudes of detected SCNAs were increased approximately 6.4x comparing to without size-selection (Figure 19E). Moreover, it shows SCNAs not only those observed in pretreatment but also additional SCNAs that were not detected in the pretreatment sample. Not only SCNAs, both in vitro and in silico size-selection strategies also improve SNV/INDELs detection using WES. Integrating cfDNA fragment size analysis and SCNAs together increases the performance of the classification model discriminating between cfDNA samples from patients and those from healthy individuals. Their experiment demonstrates exploring the biological properties of cfDNA, fragment length in this study can overcome the current limitation of sensitivity and support downstream clinical and research applications.

1.6 Aims of Thesis

Liquid biopsy offers non-invasive approach to get genetic material from a patient while also getting pooled genetic profile from heterogeneous origin including entire tumor mass. In collaboration with the Early Cancer Diagnostics and Reverse Translation unit, KiTZ Hopp Children's Cancer Center, we collected cfDNA from a group of pediatric cancer patients. We aim to use the advantages of cfDNA in the clinical management of pediatric cancer using multi-omic data. However, the utilization of cfDNA in pediatric cancer have not been investigated comprehensively with multiple next-generation sequencing technique. This thesis aims to **investigate the utilization of cfDNA in detecting genetic alterations based-on three next-generation sequencing approachs namely low-coverage whole-genome sequencing (lcWGS), whole-exome sequencing (WES) and deep gene panel-sequencing (Panel-seq)**. A set of druggable genes in pediatric cancers would be the alteration to focus on. To support this investigation, two analyses were performed

1. Evaluate the performance of cfDNA in detecting copy-number variations (CNVs), somatic point mutations (SNVs and INDELs) using lcWGS, WES and Panel-seq base-on information from tumor sequencing data
2. Detect genetic aberration from cfDNA that could potentially indicate the use of targeted therapy

It has been shown in many adult cancer studies that the tumor-derived cfDNA is shorter than cfDNA shed from non-malignant cells. The increasing proportion of short-fragment cfDNA is correlating with pathological stage of tumor. Moreover, the size-selection for short-fragmented cfDNA enchances the detection of tumor copy-number aberrations. It opened an opportunity to use this characteristics as a quantitative measurement of tumor from cfDNA. Recently, none of bioinformatics tool can comprehensively extract fragment-length profile from the next-generation sequencing data and provide genome-wide pattern of fragment length of cfDNA. In this study, we explored **the fragment-length characteristic of cfDNA in pediatric cancers** and aims to **increase the success of detection of tumor-derived cfDNA**. The accomplished these aims, we have to

1. Demonstate the fragment-length chracteristic of tumor-derived cfDNA in pediatric cancer patients
2. Develop a bioinformatics tool that extract fragment-length profile of the sample and analyse genome-wide pattern of fragment length of cfDNA
3. Evaluate the fragment-length characteristic of cfDNA as a marker of tumor aberration in cfDNA assay

2 METHODS

In this dissertation, we developed several bioinformatics workflows for analyzing specific next-generation sequencing data including low-coverage whole-genome sequencing (lcWGS), whole-exome sequencing (WES), and gene-panel sequencing (Panel-seq) of cfDNA samples (Figure 20). This chapter describes technical details involving the detection of copy-number variants (CNVs), druggable alterations, and alterations in telomeric regions. The fragment length analysis with the new bioinformatics method is described further in Chapter 3.

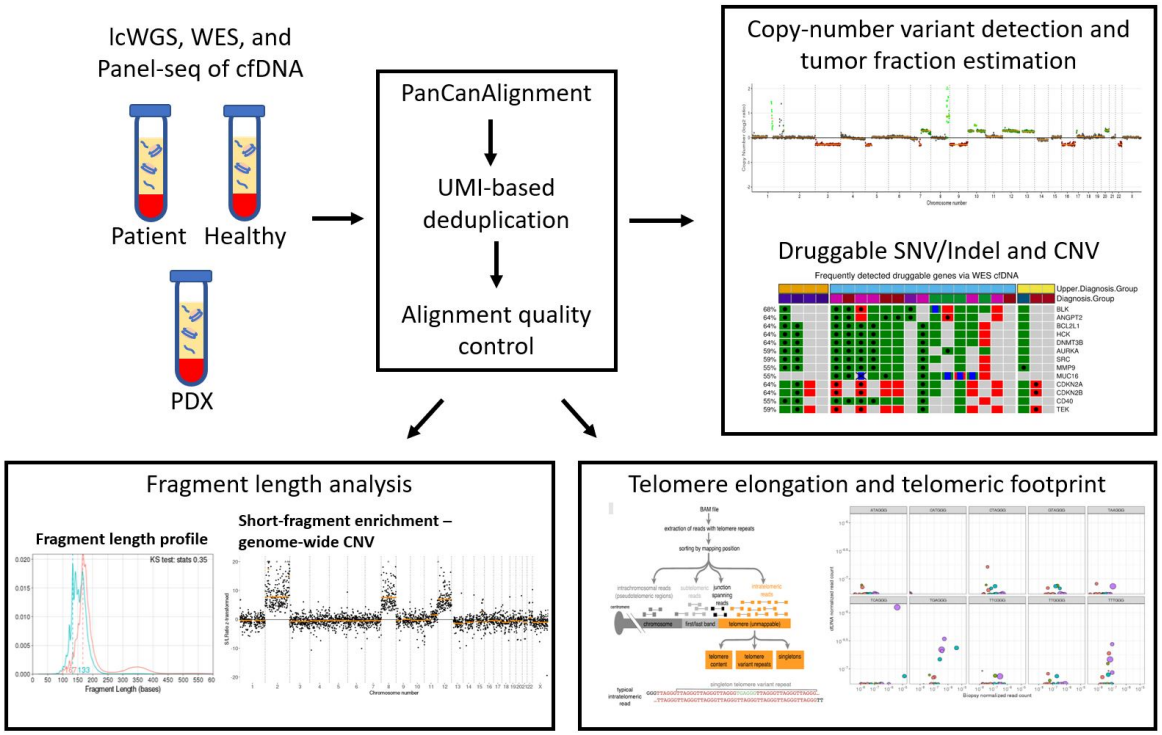


Figure 20: The overall analysis workflow to analyse next-generation sequencing data of cfDNA

2.1 Library Preparation and Next Generation Sequencing (NGS)

2.1.1 Tumor and blood control samples - whole-exome sequencing

In the collaboration with the Department of Pediatric Neurooncology at the German Cancer Research Center (DKFZ), pan-pediatric cancer samples have been collected from children, adolescents, and young adults. The library preparation, and the sequencing process of individual-matched tumor and blood samples have been previously described in the INFORM pilot study [114]. Primary tumors and matched controls from each patient were submitted to DKFZ Genomics and Proteomics Core Facility. Either SureSelect Human All Exon V5 or SureSelectXT HS Human All Exon V7 capture kit were used to capturing the coding regions of the genome without untranslated regions. The whole-exome sequencing was operated by Illumina HiSeq sequencing machines with paired-end sequencing strategy.

2.1.2 Cell-free DNA sequencing

The processes of the sample extraction and library preparation have been performed by the Early Cancer Diagnostics and Reverse Translation unit, KiTZ Hopp Children's Cancer Center. The cell-free DNA samples from each patient were extracted and submitted to the DKFZ Genomics and Proteomics Core Facility. The exons without untranslated regions were captured by either SureSelect Human All Exon V5 or SureSelectXT HS Human All Exon V7 capture kit. Sequencing was performed by Illumina

HiSeq sequencing machines with a paired-end sequencing strategy. For low-coverage whole-genome sequencing, the library preparation was carried out with either the Accel-NGS 2S Plus DNA library kit, which allows unique molecular barcoding, or PicoPLEX DNA-Seq. Gene-panel sequencing utilized the customized gene-panel developed by the Department of Neuropathology, Heidelberg University Hospital [115]. The library preparation was carried out with Accel-NGS 2S Plus DNA library kit with unique molecular barcoding process.

2.2 Sequencing Data Pre-processing : ODCF Sequence Alignment and Somatic Variant Calling Workflow

Sequencing data of tumor, control, and cfDNA were transferred to DKFZ Omics IT and Data Management Core Facility (ODCF). In-house bioinformatics workflows for sequence alignment and somatic variant calling were performed. Briefly, this workflow performed sequence alignment onto the GRCh37 (hg19) human reference genome plus PhiX sequence by using BWA-MEM [116]. Duplicated marking, sorting and indexing processes were performed by using Sambamba [117] and samtools [118] respectively. Quality matrices of the alignment (e.g. coverage, percentage of mapped reads, percentage of duplicates) were extracted by in-house scripts. This workflow is publicly available at [`https://github.com/DKFZ-ODCF/AlignmentAndQCWorkflows`].

Somatic SNV and INDEL calling was performed by ODCF with the in-house SNVCallingWorkflow and the IndelCallingWorkflow from individual-matched tumor-control or cfDNA-control BAM files as previously described [15]. In brief, somatic SNVs were detected by using Samtools mpileup and bcftools. Somatic INDELs were detected by using Platypus [119]. All detected variants were annotated by using ANNOVAR [120] and GENCODE database version 19 [121]. Only somatic high-confidence coding or splice site variants were used for downstream analysis. The somatic SNV and INDEL calling from matched cfDNA-control were performed with option -t 500 -c 0 -x 1 -l 1 -e 0 and set the score of 7 as the threshold of high-confidence variant to allow detection of low allele frequency mutations. Finally, One Touch Pipeline (OTP) [122] provides a web-based portal showing the overview of available sequencing data, quality matrices, and the result of variant calling.

2.3 Copy-number Variant Calling for Tumor Sequencing Data

Copy-number variants were inferred from whole-exome sequencing of individual-matched tumor-control samples by using CNVkit [123]. CNVkit used both on-target reads and off-targets reads to determine copy-number aberrations across the genome. It also corrects variability of the sequencing read depth regarding GC content, library size, and spacing of target regions.

The segmentation and CNV calling processes were already described in detail [124]. Briefly, genomic positions with alternative allele frequencies between 0.3 and 0.7 are considered heterozygous SNPs. Segmentation was performed on the alternative allele frequency information. Only segments that contain at least 20 heterozygous SNPs were later used in the estimation of tumor ploidy and tumor cell content. The segments were classified into balanced, ambiguous, and imbalanced segments using the distribution of the alternative allele frequency. The ambiguous segments were excluded from the analysis. For imbalanced segments, the average B-allele frequency (BAF) of all SNPs was calculated per segment. The average read count of the B-allele of a segment was calculated as the read count multiply by the BAF of the segment. Estimation of tumor cell content (TCC) and tumor ploidy method was adapted from ACEseq [125]. The range of TCC between 0.15 and 1.0, and tumor ploidy between 1 and 6.5 were included in the model fitting procedure. The distance per TCC/ploidy solution was calculated as the local minimum in the weighted mean distance.

2.4 Developing a Bioinformatics Workflow for CfDNA Sequencing Analysis

The following section describes the bioinformatics workflow for cfDNA sequencing analysis in detail.

The overview of the workflow is shown in Figure 21.

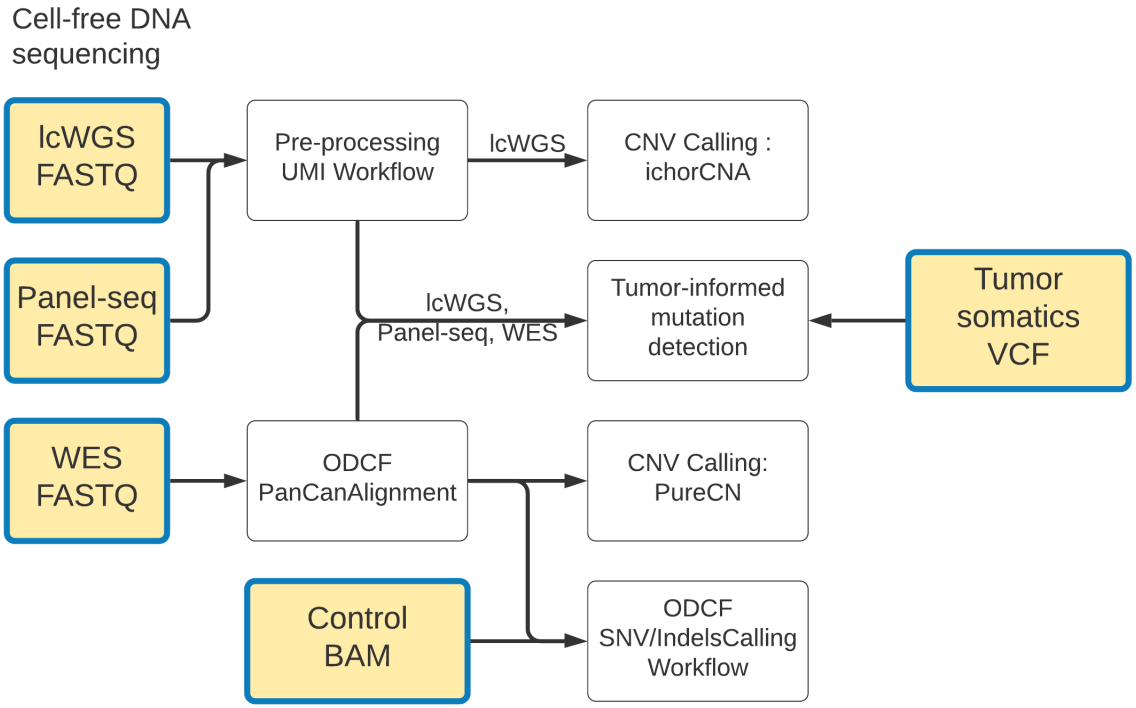


Figure 21: Overview of bioinformatics analysis workflow

2.4.1 Unique molecular index integration workflow for lcWGS and Panel-seq

Unique molecular index (UMI) barcoding is a sequencing strategy that can suppress sequencing artifacts that occur during PCR by calling consensus sequences from reads originating from the same DNA molecule. Moreover, molecular barcodes allow differentiation between reads of molecular origin from PCR products. It increases the overall sequencing coverage comparing to the regular markduplication process when one deeply sequences highly fragmented cfDNA. Fgbio toolkit, developed by Fulcrum genomics, provides the UMI processing workflow [126]. This workflow required sequencing FASTQ files of the paired-end reads (R1 and R2), a FASTQ file of sample-matched UMI (I1), and a BAM file as inputs of the workflow. The workflow is implemented as follows (Figure 22).

1. fgbio-FastqToBam matches UMI sequences (I1) with sequencing reads (R1 and R2 files) using the read name. A sorted unmapped BAM file is created. The UMI is added per alignment record into the RX tag.

2. Picard-MergeBamAlignment merges information of the unmapped BAM with the alignment information from the mapped BAM file.

3. fgbio-GroupReadsByUmi groups sequencing reads that originate from the same original molecule by sub-grouping those reads by the UMI sequence and the mapping positions. The output of sub-grouping is assigned to molecular index (MI) tag per alignment record.

4. fgbio-CallMolecularConsensusReads calls consensus reads from those reads with the same MI tag. Reads must have a minimum mapping quality of 20.

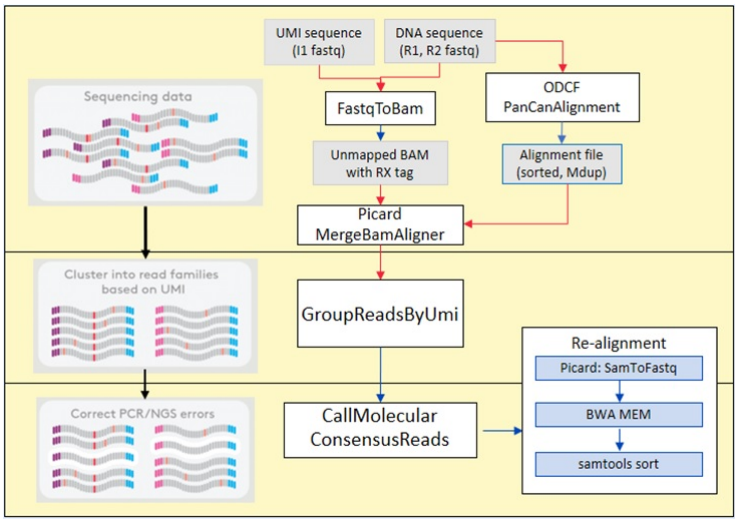


Figure 22: Pre-processing workflow : UMI-based deduplication and errors correction.

5. Re-alignment is performed by Picard-SamToFastq extracting the consensus reads as FASTQ format and BWA-MEM aligning reads onto the reference genome.

6. Samtools creates the sorted alignment file and index file (.bai).

7. For panel-sequencing data, on-target reads are extracted by using bedtools-intersect function from a given target-region bed file.

2.4.2 Extracting sequencing coverage matrices

The sequencing coverage of low-coverage whole-genome sequencing was extracted by Picard-CollectWgsMatrics [127]. For this assessment, paired-end reads (flag value 3) with minimum mapping quality of 20 and excluded mark-duplicated reads (flag value 1024) were used.

The on-target coverage of whole-exome sequencing was extracted from QC matrices table provided by ODCF AlignmentAndQCWorkflows.

For panel-seq data, the median on-target depth of coverage was calculated by using samtools-depth and an in-house bash script. Only reads with minimum mapping quality of 20 were considered.

2.4.3 Assessing the effect of DNA oxidation artifact

For both WES and Panel-seq of cfDNA, DNA oxidation artifacts C > A/G > T [128] has been considered as one of the quality control measurements. Picard-CollectOxoMatric was used to collect these alterations and calculate the Phred-scaled probability of the oxidation artifact. The lower Phred-score implies higher 8-oxoguanine artifact rate. For each sample, the average Phred-score of substitution C(C>A or G>T) were calculated as a quality measurement of whole-exome sequencing and panel-sequencing of cfDNA samples.

2.4.4 Copy-number variant calling for low-coverage whole-genome sequencing

ichorCNA (v0.3.2) [102] was used for segmentation, tumor fraction estimation and CNV calling. To reduce noise and correct the systematic biases introduced by the sequencing platform, sample preparation protocol, cfDNA-specific fragmentation structure, creating a Panel-of-Normal (PoN) from a group of selected cfDNA samples is necessary. The PoN was created from patient-derived cfDNA that does not have large copy-number alterations. The NIPTeR package [129] was implemented to filter-out cfDNA

with the large copy-number alteration. Since cfDNA samples in this study were prepared by using two preparation kits, Accel-NGS and Picoplex, two separated PoNs were created as the following instruction.

1. BAM files with coverage between 0.24 - 2.35 for Accel-NGS samples and more than 0.1 for Picoplex samples were initially selected.
2. Each of the selected BAM files was loaded into R environment using the NIPTeR package as a NIPTSample object. The GC bias correction was performed using LOESS method via NIPTeR:gc_correct function.
3. A NIPTControlGroup object was created from a list of NIPTSample objects in the previous step.
4. The function NIPTeR::diagnose_control_group was used iteratively to compute z-scores per chromosome of every sample in the NIPTControlGroup object. In each iteration, the function reported samples with the aberrant chromosomal event. The reported samples were removed from the control group and then the process continued until no aberrant sample was reported.
5. The final samples in the NIPTControlGroup were used in the creation of PoN by ichorCNA.

Once a PoN was created, it was used in the copy-number detection by ichorCNA. Since the majority of cfDNA sample contains low concentration of tumor-derived cfDNA, ichorCNA parameters were modified to improve CNV detection having low ctDNA samples. The parameters were changes as followed. -ploidy "c(2,3)" -normal "c(0.8,0.9,0.95,0.99,0.995)" -maxCN 4 -includeHOMD FALSE -estimateScPrevalence FALSE -scStates "c()" -chrTrain "c(1:22)". These parameters setting allows fitting ranges of non-tumoral contamination: 80%, 90%, 95%, 99% and 99.5% ; cell ploidy of 2 and 3; segment copy-number from 1 to 4 copies. Subclonal fraction estimation was ignored. The most likelihood tumor fraction was interpreted as the final estimated tumor fraction.

2.4.5 Copy-number variant calling for whole-exome sequencing

Unlike CNV calling of the matched tumor-control data, the result of CNV calling of cfDNA-control sample produces rather a high level of noise and unstable segmentation. It is possibly due to the differences in sequencing protocol, DNA capture-kit, coverage, and genomic structure of the source. PureCN [130] was selected as software for CNV calling on the whole-exome sequencing data of cfDNA. Similar to ichorCNA, PureCN allows the creation of PoN selected from process-matched samples. To be selected as a PoN, the sample must have the median on-target depth of coverage between 142 and 269 By the result of tumor-informed SNV/indel variant detection process (Section 2.4.7), the samples also must support less than 3 somatic variants in the matched tumor.

Once a group of samples were selected, PureCN requires them for the creation of NormalDB. The instruction of this process can be found in PureCN vignettes document. Briefly, the coverage of each sample was extracted and normalized for the GC-bias. A normal panel VCF containing mutations commonly found in the selected samples was created by the following instruction.

1. For each of selected BAM files, germline and somatics variant were detected by using GATK Mutect2 [131] in "tumor-only" mode with parameters -max-mnp-distance 0 -min-base-quality-score 20 -annotation BaseQuality -read-filter MappingQualityReadFilter -read-filter OverclippedReadFilter -minimum-mapping-quality 30 -read-filter FragmentLengthReadFilter -min-fragment-length 30. This process produced a VCF file per individual sample.

2. Only common variants found in at least 3 samples were selected. This can be done by using VCF files in the previous step as the input of GATK:CombineVariants where parameter minimumN is set to 3.

Lastly, PureCN runs CNV calling in the setting that allows the detection of samples with lower tumor purity. The software parameters were set as followed: -minpurity 0.05 -minaf 0.01 -error 0.0005 -maxploidy 3 -maxcopynumber 8 -padding 25 -model betabin -funsegmentation PSCBS -postoptimize. This parameters fix the PureCN solution space down to tumor purity of 5% as recommned by the softwere developer. The model search for solution with the tumor ploidy up to 3 ploidy and 8 number of copy. The segmentation were performed by PSCBS.

2.4.6 Sequencing quality control of cfDNA sequencing data

Before further analysis, a cfDNA sequencing data must pass the following quality threshold.

For low-coverage whole-genome sequencing, a sample must have genomic coverage above 0.1 reported by Picard-CollectWgsMatrics. GC-Map correction MAD, reported by ichorCNA, must be less than 0.15 to reduce high variance in the data.

For WES, a sample must reach 60 on-target coverage, reported by the ODCF workflow, to achieve the detection of the tumor variant allele frequency at 2%. No coverage threshold was applied for Panel-seq samples. WES and Panel-seq samples with the average Phred-score of substitution C(C>A or G>T) below 30 (Section 2.4.3) were excluded from downstream analysis.

2.4.7 Tumor-informed SNV/indel variant detection in cfDNA sequencing data

In addition to the somatic variant calling, a set of in-house scripts were developed for interrogating a cfDNA sample if the tumor-derived cfDNA exists. Tumor high-confidence somatic variants, in the tumor VCF file, were used as ground truth and look them up from the read pileup information of individual-matched cfDNA. Each variant was sorted into three categories. If a tumor variant is present in cfDNA, the variant will be reported as "var_present", otherwise it will be reported as "not_present". The tumor variant will be initially reported as "pos_not_covered" when no read was aligned onto the position of the variant. Only the read pileup that has the read minimum mapping quality of 1 and the base quality of 20 were considered. The variant positions with less than 5 supporting reads were marked as "pos_not_covered" and were discarded from the analysis. A tumor variant needs at least one read in the cfDNA sample that supports the tumor allele to be reported as "var_present".

To support the evalutation of CPA Score in detecting high ctDNA, we categorise cfDNA WES into two classes: high ctDNA and low ctDNA. Threshold were estimated by the power of detection detecting tumor purity > 2.5 %, average coverage 210, and tumor ploidy 2 using calculatePowerDetectSomatic of PureCN package. With this parameter, samples that detect at least 17% of tumor point mutation and 3 tumor point mutations were categorised as high ctDNA otherwise as low ctDNA.

2.5 Xenograft-derived Sequencing Data Analysis

The sequencing data from the patient-derived xenograft experiment were also processed by the ODCF sequence alignment workflow and the UMI sequencing workflow. All reads were mapped onto the reference FASTA file containing both the human reference genome (GRCh37) and the mouse reference genome (GRCm38). The separation between human-derived cfDNA and mouse-derived cfDNA was done by using samtools. Human-derived cfDNA was further analyzed by ichorCNA for CNV calling and tumor

fraction estimation. The fragment length profiles of the human-derived and mouse-derived cfDNA were analyzed by using cfdnakit (Chapter 3).

2.6 Telomere Content Estimation and Quantification of Telomeric Variant Repeat

The telomere content of both tumor and cfDNA was estimated by using TelomereHunter [132]. Briefly, TelomereHunter extracted reads containing at least six non-consecutive repeat sequences (TTAGGG, TCAGGG, TGAGGG, and TTGGGG) from a BAM file. The extracted reads were sorted into four categories depending on their mapping position on the genome. Only unmapped reads or reads with a mapping quality lower than 8 were considered intratelomere reads. The telomere content was calculated as the number of intratelomere reads per million reads having a GC content of 48-52%. Telomere variant repeats (TVR) were detected in the intratelomeric reads by searching for the hexamer NNNGGG where 'N' can stand for A, C, G, or T. The TVR that has a neighboring t-type context, (TTAGGG)3-NNNGGG-(TTAGGG)3, were called "singletons". The absolute counts of each TVR singleton were normalized by the total number of reads in the sample and used for further analysis.

For tumor samples, the matched tumor-control WGS were used to calculate the log2 ratio of the estimated telomere content and TVR singleton count. For cfDNA samples, only the lcWGS (BAM files from the ODCF workflow) of cfDNA was used as the input of TelomereHunter. The input BAM file of cfDNA were obtained from the standard ODCF sequence alignment not the result of UMI workflow.

3 DEVELOPMENT OF BIOINFORMATICS METHODOLOGY (cfdnakit Package)

3.1 Background

Cell-free DNA (cfDNA) has become an attractive source of DNA that shows potential benefits to the management of cancer patients. Detection of tumor-derived cfDNA or circulating tumor DNA (ctDNA) has been extensively demonstrated in many cancers and different clinical settings. Nevertheless, the low concentration of ctDNA has been a major challenge to the success especially for those patients with early-stage or localized tumors. Research on the biological characteristics of cfDNA have provided new insights regarding its cellular origin and mechanism behind the secretion [54, 72, 133]. These discoveries post new opportunities also in terms of data analysis to increase the success of ctDNA detection. We are interested in the characteristical length of cfDNA showing that the ctDNA is relatively shorter than non-ctDNA fragments [110, 112]. The enrichment of short-fragmented cfDNA correlates with the pathological stage of the tumor and mimics the genomic copy-number alteration of the tumor population [110]. Analyzing the fragment length of cfDNA could provide complementary evidence of ctDNA in the pool of cfDNA fragments [134, 135].

Due to the lack of specific bioinformatics tools, a software package "cfdnakit" has been developed. This package provides functions to explore the length of cfDNA from low-coverage next-generation sequencing data. Comparing the amount of short-fragmented cfDNA (<150 base-pairs) relative to long-fragmented cfDNA between multiple samples is simple by using this package. The amount of short-fragmented cfDNA can be explored throughout genomic loci and infers aberrant copy-number in the tumor genome. cfdnakit also estimates the most likely tumor fraction from the signal of short-fragmented cfDNA and calculates a copy-number tumor burden score. This score could be used to indicate overall genomic instability from the tumor-derived cfDNA. In this dissertation, this package has been used in the exploration of cfDNA samples from a pan-pediatric cancer dataset. The following sections are dedicated to methodological details of this package.

3.2 Fragment-length Distribution

The fragment length of a cfDNA can be inferred from the mapping distance between the outer end of the two paired-end reads. This information can be extracted from a BAM file in the TLEN field. cfdnakit uses Rsamtools package [136] to read a given BAM file and extracts the TLEN information. Using the Rsamtools function to read the BAM flag information, cfdnakit keeps only mapped paired-end reads with minimum mapping quality of 20 and excludes reads with markduplicated flag or being the secondary alignment. cfdnakit also excludes those reads that mapped onto blacklisted regions (described in Section 3.3). After that, sequencing reads are then separated into equal-size (100, 500, or 1000 kilobase pairs) non-overlapping genomic windows (bins). Finally, the input sequencing sample is formated as a SampleBam object in the R environment.

cfdnakit provides a function to visualize the fragment-length distribution of a SampleBam object. Given a list of SampleBam objects, this function allows comparisons between multiple cfDNA samples (Figure 23). The fragment-length distribution should present a pattern of association between cfDNA and nucleosomes. In general, plasma cfDNA would show modal length at 167 bases (the size of a DNA wrapping around a unit of nucleosome plus an H1 linker protein) and 10-bp periodically peak in the distribution of fragment lengths below 150 bases [54, 110]. CfDNA from other sources (e.g. CSF or urine) is more fragmented into less than 147 bases suggesting a different mechanism behind their secretion [137, 138]. Enrichment of short-fragmented cfDNA (<150 bp) is often observed from tumor-derived cfDNA (ctDNA) and has been recognized as a potential tumor marker [110]. The other pattern of distribution has not been reported yet. The fragments such as PCR primers are usually short (<50

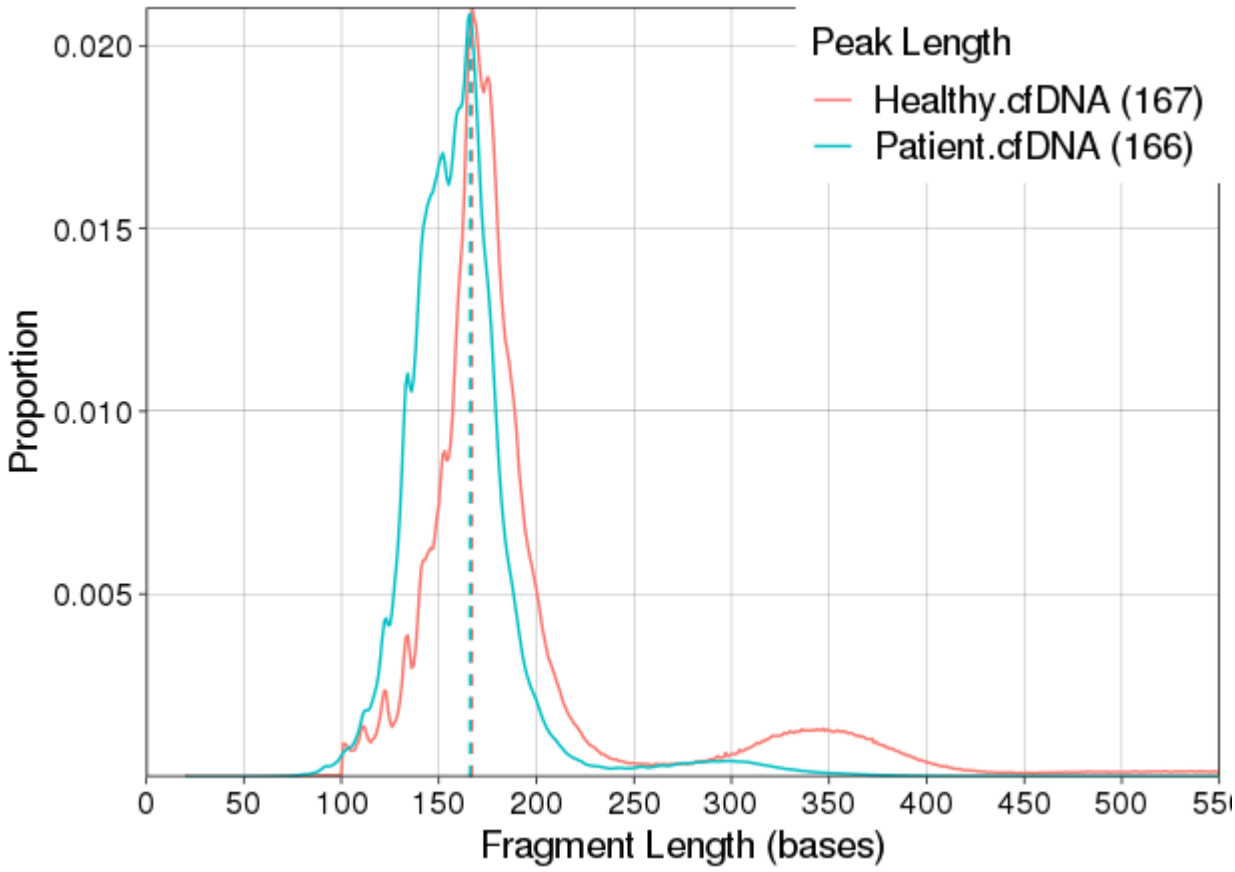


Figure 23: A fragment-length distribution plot showing in comparison between two cfDNAs derived from a cancer patient (cyan line) and a healthy donor (red line)

bases) and indicates the sample quality issue if the plot shows a high distribution within this range.

3.3 ENCODE Excluded Regions

It is recommended when analyzing genomic data to exclude sequencing reads locate within the ENCODE blacklist loci to assure the quality of the result [139]. When using the GRCh37 as the reference in cfdnakit, a set of genomic regions including the ENCODE blacklist and centromere loci, provided by UCSC Genome Browser [140], were used. Users can introduce customized blacklist regions by creating a bed file or a tab-separated file where the first three columns are chromosome, start, and end position respectively. The future cfdnakit would be able to support blacklists of other reference genomes such as GRCh38 or GRCm38.

3.4 Calculation of Short-fragmented Ratio

The number of short and long fragments of every bin is counted. The count value is called fragment-count. By default, a short-fragment is defined as a fragment with a size between 100 to 150 base pairs whereas the size of long-fragment is 151 to 250 base pairs. The short and long fragment-count are then further corrected for GC and mappability bias (Section 3.5) using the information provided by the QDNAseq package [141]. The corrected fragment-counts of short and long fragments are used to calculate the short/long-fragment ratios (S.L.Ratio) of the sample ($S.L.Ratio_{sample}$) and ratios per bin ($S.L.Ratio_{w}$) as follows:

$$S.L.Ratio_{sample} = \frac{N_{F.short}}{N_{F.long}}$$

$$S.L.Ratio_{w} = \frac{N_{F.short_w}}{N_{F.long_w}}$$

where $N_{F.short}$ is number of short fragments; $N_{F.long}$ is number of long fragments;

$w = \{1, 2, 3, ..., n\}$; where n is number of bins;

$N_{F.short_w}$is number of short fragments in bin w; $N_{F.long_w}$is number of long fragments in bin w.

$S.L.Ratio_{sample}$ can be used as a general comparative quantification of ctDNA between plasma cfDNA samples. This ratio increases when a sample contains the higher contribution of ctDNA. The $S.L.Ratio_{w}$represents the short-fragment cfDNA in a genomic bin. The aberration of ratios over a continuous locus correlates with the copy-number status in the matched tumor genome. The ratio increases when the tumor acquires more segment copies and slightly decrease in the copy-loss segment.

The results of the calculation are then returned from the function as a SampleFragment object. The object contains S.L.Ratio per bin (in table per_bin_profile) and S.L.Ratio of the sample (in table sample_profile). cfdnakit provides a plot function to visualize the S.L.Ratio throughout the genomic regions (Figure 24). The noisy plot might be the result of too low sequencing coverage or too low DNA material.

3.5 GC and Mappability Bias Correction

A LOESS regression model is created from the relation between the fragment count and the percent of GC per bin. The raw count per bin is deduced with the read count predicted by the model. Then, the values are added with the median of raw counts to bring back the range of values similar to the raw count. After correction for GC bias, the GC-corrected read counts are then corrected for mappability bias using a similar process. The mappability bias indicates the mapping capability of a genomic region

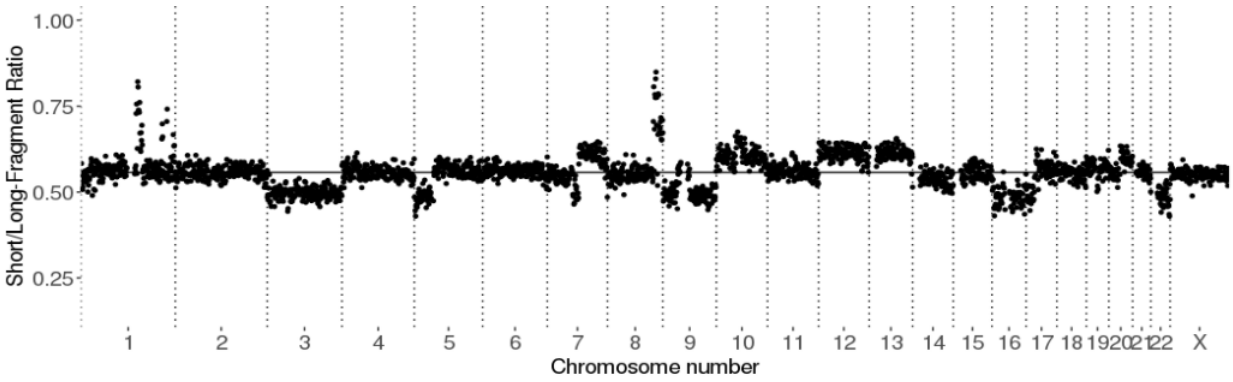


Figure 24: A plot of genomic short/long-fragment ratios (S.L.Ratio) representing enrichment of short-fragmented cfDNA in different genomic loci

to be mapped uniquely by sequencing reads. cfdnakit also produces a plot describing the read count bias within a sample and a plot showing the read count and S.L.Ratio after GC and mappability correction.

3.6 Creation of Panel-of-Normal Dataset

To estimate the rate of both technical and biological artifacts, creation of a Panel-of-Normal is usually recommended by most bioinformatics workflow. A Panel-of-Normal (PoN) of cfDNA analysis should be made from healthy samples or a group of selected patient-derived cfDNA. There is no definitive rule on how to select or how many samples should be included in a PoN. Creating a PoN will in general be better than analysis without a PoN. Nevertheless, the most important approach is including normal samples that are generated by similar techniques (such as DNA preparation methods, sequencing platform, and biological sources) as many as possible.

cfdnakit requires a PoN dataset for further analysis. Every selected sample must be initially processed by cfdnakit to extract S.L.Ratio per bin and saved the result as a separated RData file. Once every sample is processed, a text file containing paths to those RData files is created. cfdnakit will read this text file and create a matrix of S.L.Ratio. The matrix must be saved into an RData file to be used repetitively in downstream analysis.

3.7 Transforming Short-fragmented Ratio with PoN

The bias-corrected S.L.Ratio indicates the quantity of short-fragmented cfDNA and can be compared within a sample. However, to relatively compare between samples, standardization is required. cfdnakit transforms the S.L.Ratio by subtracting the median and dividing by median absolute deviation (MAD) of S.L.Ratio as follows:

$$S.L.norm_w = \frac{S.L.Ratio_w - median(\{S.L.Ratio_1,...,S.L.Ratio_n\} - \{S.L.Ratio_w\})}{mad(\{S.L.Ratio_1,...,S.L.Ratio_n\} - \{S.L.Ratio_w\})}$$

$$w = \{1, 2, 3, ..., n\}; \text{ where } n \text{ is number of bins}$$

The MAD is a term representing the median of the absolute deviation from the median. As an alternative to the standard deviation, MAD a robust measure of variability of the data. We calculated MAD from a sample as follows:

$$mad\left(S.L.Ratio_{1..n}\right) = median\left(|S.L.Ratio_i - median\left(S.L.Ratio_{1..n}\right)|\right)$$

where n is number of bins

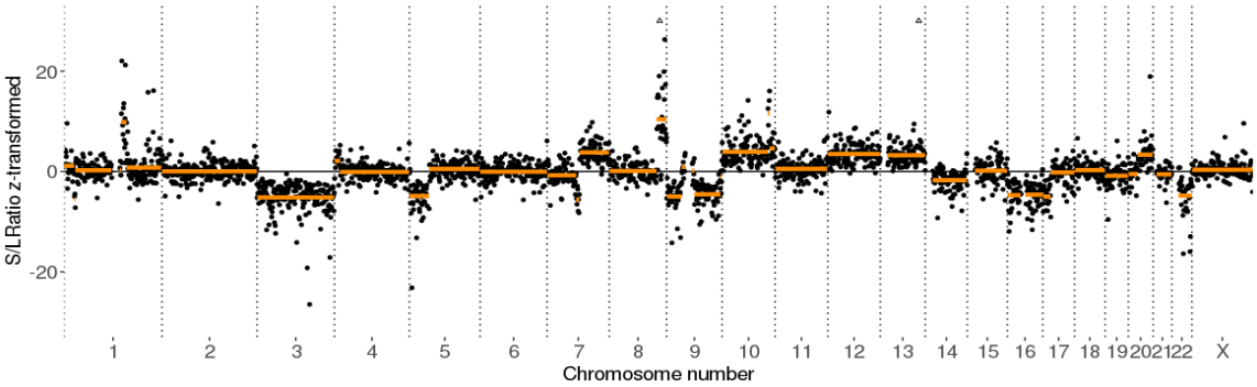


Figure 25: A plot of genomic segmentation with circular binary segmentation (CBS)

The standardized S.L.Ratio (S.L.norm) of a bin is then transformed in z-score using the PoN dataset. The z-score is calculated from the S.L.Ratio by subtracting the median and dividing by the mad of S.L.Ratio of PoN samples at the same locus.

$$zscore_w = \frac{S.L.norm_w - median(\{S.L.norm_{w,1},...,S.L.norm_{w,p}\})}{mad(\{S.L.norm_{w,1},...,S.L.norm_{w,p}\})}$$

$w = \{1, 2, 3, ..., n\}$; where n is number of bins

and $p = \{1, 2, 3, ..., m\}$; where m is number of samples in PoN

3.8 Circular Binary Segmentation

Circular Binary Segmentation (CBS) is a partition method commonly used in partitioning a genome into segments of total copy-number (TCN)[142]. Implementation of CBS in R packages (DNAcopy[143] and PSCBS[144]) is widely used in many copy-number analysis tools, for example, ACEseq [125] and cnvkit. cfdnakit utilizes the CBS algorithm and additional functions provided by the PSCBS package.

Once the S.LRatio is calculated per genomic windows and transformed into a z-score, the CBS is performed. Outlier signals that are significantly different from the neighboring loci are identified by PSCBS function dropSegmentationOutliers with default parameters. The biological gaps such as centromere where two adjacent loci should be treated as non-neighboring loci are identified. cfdnakit defines a region as a gap if the distance between two loci is larger than 10 Mb with no observed signal between them. The actual segmentation is then performed using the function segmentByCBS. The function produces a segmentation result using the median as a representative value (Figure 25). To avoid oversegmentation, cfdnakit also applies hierarchical clustering to prune the segmentation result by setting the tree height threshold to 0.5.

3.9 Copy-number Variant Calling and Tumor Fraction Estimation

The median S.L.Ratio of segments can be used as the signal for the estimation of tumor content and ploidy of the tumor cell population. cfdnakit calculates the expected signal for tumor fraction (tf) between 0.0 to 0.8 (with increments of 0.01), tumor ploidy ($ploidy$) between 1.5 to 4 (with increments of 0.05), and integer copy numbers (TCN) between 1 and 5 as followed by the package default:

$$Expected.S.L.Ratio = median.segment \cdot \left(\frac{tf \cdot TCN + 2 \cdot (1 - tf)}{tf \cdot ploidy + 2 \cdot (1 - tf)}\right)$$

where $median.segment$ is the median segment S.L.Ratio of all segments ;

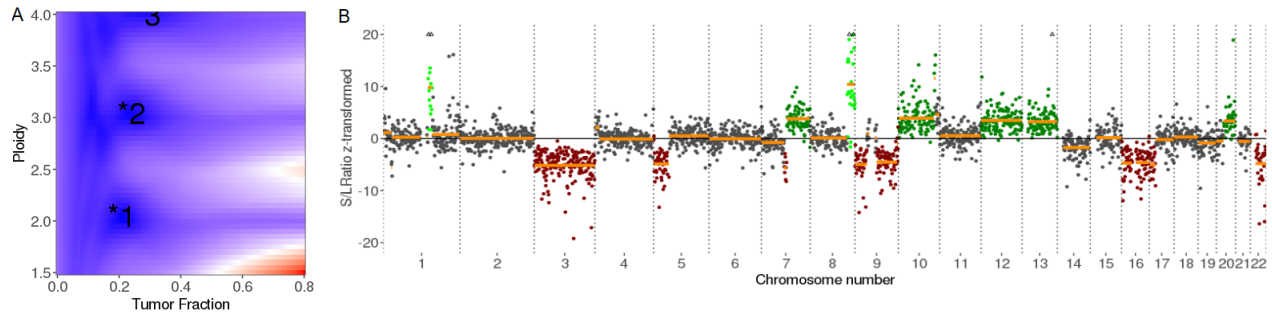


Figure 26: Copy-number variant calling solution space and coverage plot: A) Heatmap plot showing all solution distances. The color gradient ranges from the lowest distance (blue) to the highest distance (red). The lowest distances per rounded ploidy (2, 3, and 4) are marked with asterisks where the ranking number are nearby. B) The genome-wide copy-number plot of the best solution (lowest distance). The color represents the associated copy-number alteration: deletion (red), neutral (grey), gain (green), amplification (light green).

$$tf = \{0.0, 0.01, ..., 0.8\} \; ; \; ploidy = \{1.5, 1.55, ..., 4\} \text{ and } TCN = \{1, 2, .., 5\}.$$

Cfdnakit calculates the distance between the observed signal and the expected signal as the absolute difference of the expected signal and the signal of the segment.

$$distance_{segment} = |Expected.S.L.Ratio - S.L.Ratio_{segment}|$$

where $S.L.Ratio_{segment}$is the median S.L.Ratio of segment (the signal of segment)

and $distance_{segment}$is the distance of a segment to the expected S.L.Ratio

For a distinct set of parameters ($ploidy$, and tf), cfdnakit selects a TCN that provides the minimum distance to the expected signal. cfdnakit calculates the distance per distinct set of parameters (solution) as the mean distance weighted by the segment length as follows:

$$distance\,(ploidy, tf) = \frac{\sum_{i=1}^{N_{segment}} (distance_{segment_i} * length_{segment_i})}{\sum_{i=1}^{N_{segment}} length_{segment_i}}$$

where $distance_{segment_i}$is the distance of segment$_i$;

$length_{segment_i}$is the number of bins in segment$_i$

and $N_{segment}$is the total number of segment

Cfdnakit reports the distances of all solutions and visualizes them with a heatmap plot (Figure 26A). The color and color intensity represents the distance of a solution. The asterisks (*) indicate solutions with the minimum distance per integer ploidy (ploidy 2, 3, and 4 by the package default). Finally, cfdnakit provides CNV profiles that represent the best solution per round ploidy (Figure 26B). Users can select which solution to be reported and visualized.

3.10 Copy-number Abnormality Score

As the result of copy-number solution fitting, the tumor fraction (tf) indicates the estimated quantity of ctDNA from the amplitude of signals. cfdnakit also implements the copy number profile abnormality (CPA) score [145] to quantify the tumor burden from the segmentation result. In cfdnakit, this score is defined similarly as follows:

$$CPA = \left(\sum_{i=1}^{N_{segment}} \left(|Z_{segment_i} \times l_{segment_i}| \right) / N_{segment} \right) \cdot S.L.Ratio_{sample}$$

where $Z_{segment_i}$ is the z-score of segment$_i$;

$l_{segment_i}$ is the number of bins in segment$_i$

and $S.L.Ratio_{sample}$ short/long-fragment ratio of the sample (Section 3.4)

This score is robust to coverage bias and noisy fragmented signals. The full formula and its advantages were emphasized in the original publication [145]. Briefly, the Gaussian noise does not affect the score because the z-scores of segments, instead of the z-score of bins, are considered. Second, the average segment length is used as a penalty for sample quality. The signal of a bad quality sample does not strongly affect the score whereas a true highly unstable genome would overcome this penalty.

3.11 Package Repository

The cfdnakit package is currently accessible via the GitHub repository (https://github.com/Pitithat-pu/cfdnakit) as an open-source software under the GNU General Public License v3.0. The package information and analysis instructions are available on the wiki page of the repository.

4 RESULTS

4.1 The Pediatric Cohort Dataset

The Early Cancer Diagnostics and Reverse Translation unit, KiTZ Hopp Children's Cancer Center collected serum/plasma from patients with brain tumors (n= 62), sarcoma (n=55), and other pediatric cancers (n=14) and additional healthy individuals (n=10). Cell-free DNA was extracted and sequenced with three different strategies, namely: low-coverage whole-genome sequencing (lcWGS), whole-exome sequencing (WES), and gene-panel deep sequencing (Panel-seq). The collection of cfDNA sequencing data includes 137 samples with lcWGS, 71 with WES, and 77 with Panel-seq. The individual-matched tumor genomic data are available through the study "Individualized Therapy for Relapsed Malignancies in Childhood" (INFORM) project. The tumor genomic data include 131 matching tumor WES-, 131 lcWGS-dataset, and 129 methylation arrays. Figure 27 and Supplement Table S1 show the overall collection of cfDNA and solid tumor samples.

More than half of cfDNA samples (53.5%; n=84) were sequenced by more than one sequencing method (Supplement Figure S1). The majority of cfDNA samples from brain tumors were sequenced by all three strategies (n=24) (Supplement Figure S1A). Sarcoma cfDNAs were more exclusively sequenced through lcWGS (n=40) and contains few overlaps of all three strategies (n=8) (Supplement Figure S1B). CfDNA from other pediatric cancers are mostly overlapped by three strategies (n=12) (Supplement Figure S1C). Nevertheless, this dataset allows the comparison between different sequencing strategies in detecting different types of genetic alterations including copy-number variant (CNV) and point mutation (SNVs and INDELs).

4.2 Data Preprocessing

This section describes the result of the preprocessing including the result of applying sequencing quality control and unique molecular index integration workflow.

4.2.1 Quality control filters samples with sequencing artifact and insufficient coverage

Prior to further analysis, several quality measurements have been performed as previously described (Method Section 2.4.6). Figure 28 shows the overall filtering process and the number of samples passing the quality threshold. We excluded 4 lcWGS samples from downstream analysis because they have less than 0.1x genomic coverage or ichorCNA MAD less than 0.15. Out of 71 WES sets, 4 samples have insufficient genomic coverage (less than 60x median on-target depth of coverage) or excessive levels of oxidative artifacts. In addition, we checked the genotyping similarity between the cfDNA WES and the matched germline WES. We excluded one WES sample that had a correlation coefficient of 0.54 to the matched germline WES. We discarded 3 of 77 Panel-seq samples that failed oxidative quality control. Finally, 133 lcWGS, 66 WES, and 74 Panel-seq cfDNA samples were subjected to downstream analysis.

4.2.2 Unique molecular indexing improves the sequencing coverage

A regular bioinformatics workflow applied a duplicate alignment marking (e.g. samtools-markdup) on next-generation sequencing (NGS). This process locates and tags duplicate reads, originating from a single DNA fragment, in an alignment file (BAM file). The aim is to remove duplicates that arise from PCR which are likely to contain sequencing artifacts and coverage bias such as GC-extreme regions [146]. CfDNA sequencing may require PCR because of inadequate amounts of starting DNA material, and losses during size selection. Nevertheless, cfDNA is known to be highly fragmented as the result of endonuclease reaction before and after secretion into the circulation (Section 1.5.1). Several fragmented molecules would have been mistaken for being PCR duplicates and excluded from downstream analysis.

Unique Molecular Indexing (UMI) is one of the sequencing strategies that attach additional sequences to each input molecule of DNA. With this barcode index, PCR duplicates can be accurately identified and be distinguished from real DNA duplicate fragments. Hence, UMI could enhance the performance of deep coverage sequencing in detecting point mutations from cfDNA. We applied the UMI integration

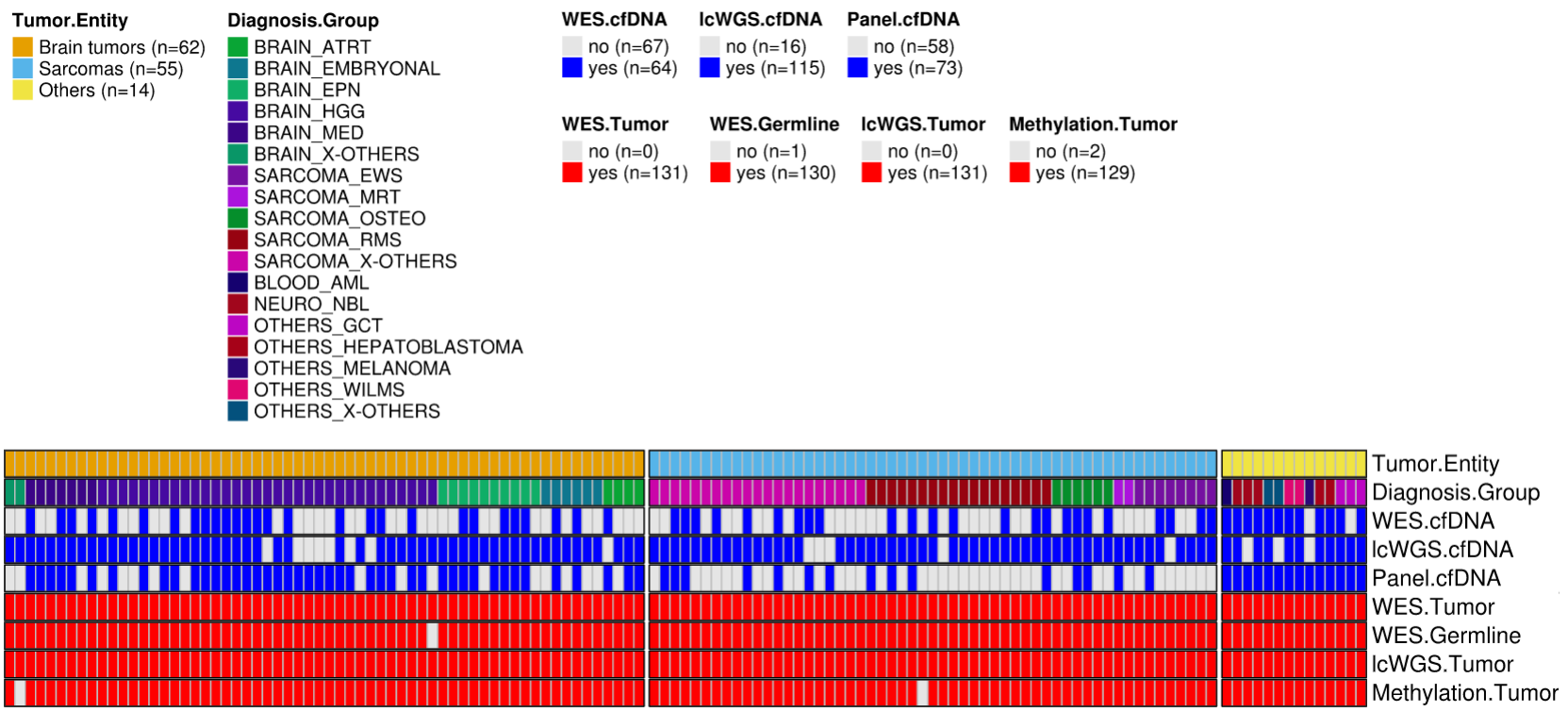


Figure 27: The overview of pediatric cohort (INFORM) - liquid biopsy cfDNA in this thesis

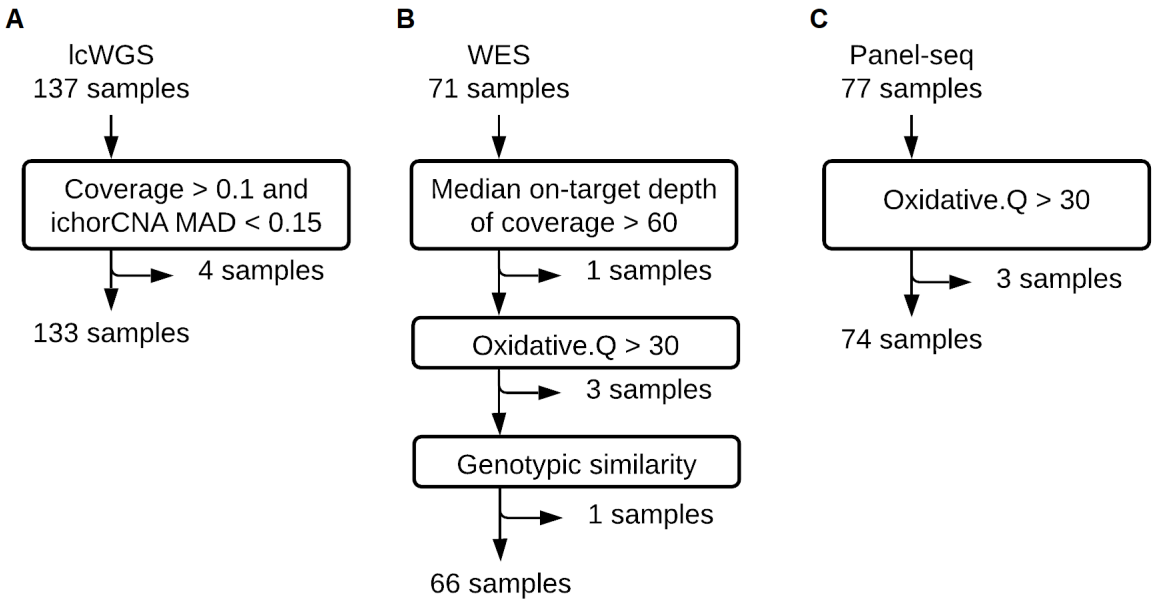


Figure 28: The number of samples that passed or failed the sequencing quality measurement of lcWGS (A), WES (B), and Panel-seq (C).

workflow (Method section 2.4.1) to both lcWGS (Accel NGS library only) and Panel-seq. We extracted median on-target depth of coverage from Panel-seq data and genomic coverage from lcWGS data (Method Section 2.4.2).

We compared the coverage of result BAM files before and after the implementation of the UMI integration workflow (Figure 29). The median on-target depth of coverage increased approximately threefold from 328.5 to 820.5 in Panel-seq (Figure 29A). On the other hand, the median genomic coverage of lcWGS increases approximately 7% from 1.27 to 1.38 after integrating UMI deduplication (Figure 29B). The increase appears to be influenced by the degree of duplication of the DNA template. The fold change in coverage correlates positively with the mark-duplication rate for both Panel-seq (Figure 29C) and lcWGS (Figure 29D). The Panel-seq library was constructed using a larger amount of input DNA and produced more throughput than the lcWGS. Moreover, cfDNAs are already highly-fragmented DNA when isolated from the blood sample. The more duplicated templates present in a sample, the more reads can be obtained using the UMI barcoding strategy and the greater the chance of detecting point mutations with low allele frequency.

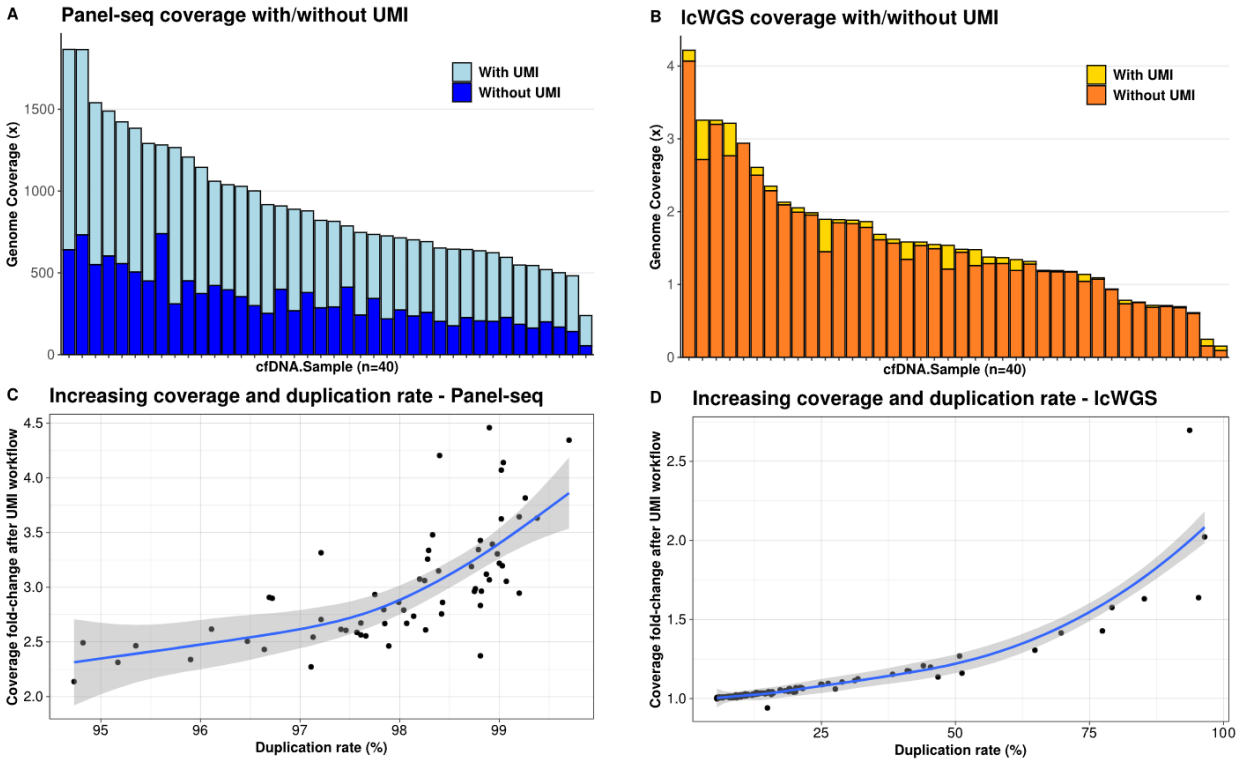


Figure 29: Increasing coverage of deep and shallow cfDNA sequencing with UMI: (A) On-target coverage of cfDNA Panel-seq data was compared between the regular bioinformatic Markduplication procedure and the UMI-based deduplication workflow. The UMI-based deduplication workflow (light blue) improves on-target coverage by about 3-fold compared to the regular Markduplication (blue). (B) UMI-based deduplication (light yellow) improves coverage over mark-duplication (yellow) by only 7% in lcWGS samples. The increasing coverage in Panel-seq (C) and lcWGS (D) correlates with the degree of duplicated reads removed by the Markduplication procedure.

4.3 Result of CfDNA Sequencing Data from Bioinformatics Workflows

4.3.1 Low-coverage whole-genome sequencing is a comprehensive strategy to detect large copy-number alteration

Since CNV is the most common alteration in pediatric cancers, obtaining this information non-invasively via liquid biopsy could aid in the clinical management of childhood cancers. Large CNVs of tumors were detected using the matched tumor/germline WES. For each tumor sample, the CNV calling workflow (Method Section 2.3) reported the normalized log2 ratio per bin of on-target and off-target regions, genomic segments with the associated integer copy number (Figure 30A). To facilitate comparison between tumor and cfDNA CNVs, the reported CNV event was adjusted according to the reported tumor ploidy. For example, a segment with an absolute copy-number of 3 is designated as neutral if the tumor has ploidy 3.

IchorCNA detected CNVs based on sequencing coverage segmentation per 1-megabase of genomic non-overlapping windows. Based on the segmentation result, the software detected copy number aberrations by fitting a model with a range of parameters for tumor-fractions (TF) and tumor ploidy. Multiple solutions of CNV profiles were reported but only the profile with the highest likelihood score was considered (Figure 30B). When a tumor with large CNVs secretes enough DNA into the blood circulation, cfDNA would likely be able to capture those CNVs and reported high TF. However, more than of cfDNA in this cohort did not recapitulate the alteration that existed in their matched tumor profile because they had a very low tumor-fraction.

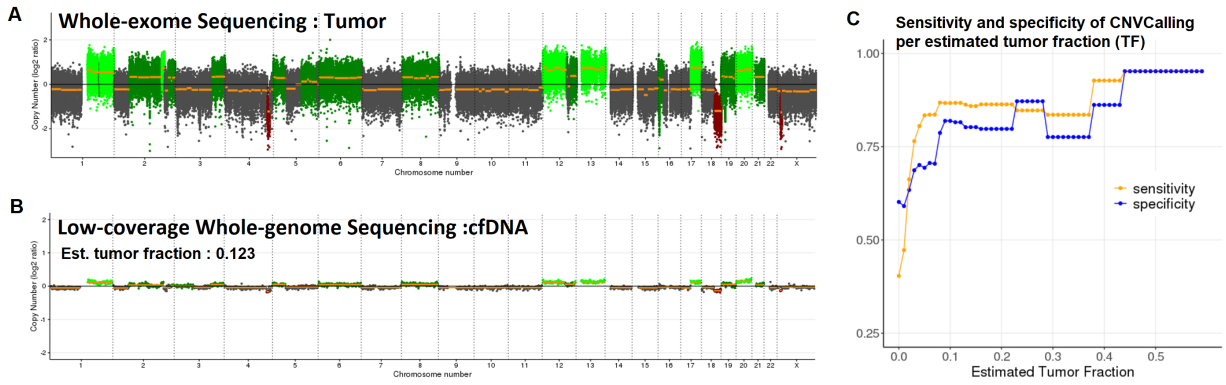


Figure 30: Detection of genome-wide copy-number alterations in cfDNA with low-coverage whole-genome sequencing (lcWGS): tumor copy number alterations were first determined from tumor tissue using WES (A). Cell-free DNA was obtained from liquid biopsy samples. The lcWGS (~2X) were generated and provided a comprehensive genomic copy number abberations where tumor-fraction estimation can be performed by ichorCNA (B). Colors in a genomic profile represent CNV events (grey:neutral, red:deletion, green:gain (3N), light green:amplification (> 3N)). Using the tumor profile as the ground truth, the sensitivity and specificity of lcWGS reaching specific estimated tumor-fraction were determined (C).

Using the matching tumor CNV profile as a ground truth, we evaluated the performance of the lcWGS strategy in detecting copy number aberrations in liquid biopsy cfDNA (Figure 30C). The sensitivity and specificity of lcWGS are relatively stable at 80% to 90% and when a sample reaches 5% or more TF. When a sample reaches the estimated tumor-fraction of 3%, lcWGS detects CNVs with a sensitivity of 76.5% and a specificity of 68.9% in this cohort. This indicates that cfDNA has the ability to detect CNVs and focal amplifications/deletions when the tumor fraction reaches a certain threshold. For further analysis and evaluation, we classified a sample with TF greater than 3% as "high ctDNA" samples, and otherwise as "low ctDNA" samples. We later determined the success of detection based on this sample classification, regardless of the lack of clinical status at the time the liquid biopsy was taken.

4.3.2 Whole-exome sequencing complements low-coverage whole-genome sequencing by detecting point mutations

The utility of the cfDNA WES strategy was demonstrated here by performing the CNV calling workflow with PureCN (Method Section 2.4.5) and tumor-informed mutation detection (Method Section 2.4.7). PureCN calculated the log2 copy number ratio of the normalized read-count of the sample and the group of process-matched cfDNA samples for both on- and off-target regions. Segmentation was performed using PSCBS, which is included in the package. As a result, genome-wide copy number events are reported per segment. CNVs with a tumor-fraction of 3% or more are likely to be identified.

We compared the number of tumor alterations, including point mutations and CNVs, between lcWGS and WES of 6 high ctDNA samples (Figure 31A). Due to higher coverage, WES provides the ability to detect tumor-derived cfDNA having deleterious somatic SNVs and INDELS. In this cohort, WES detected 89.6% (190/212) of SNVs and INDELs in sarcomas and 81.6% (58/71) in other pediatric cancers. On the other hand, lcWGS detected 14% (30/212) of SNVs and INDELs in sarcomas and 4.2% (3/71) in other pediatric cancers.

For CNVs, the detection results of lcWGS and WES are very similar among high ctDNA samples (Figure 31B). LcWGS of high ctDNA samples detected 90% (161/179) of CNVs in sarcomas and 30.4% (7/23) in other pediatric cancers. The matching WES sample detected 88.2% (158/179) of CNVs in sarcomas and 39.1% (9/23) in other pediatric cancers. In general, WES can detect CNVs similarly to lcWGS and also provides sequencing coverage that allows detection of tumor point mutations. However,

it is important to remember that the limitation of CNV calling with WES is initially set to a minimum of 5% tumor content. CNVs in samples with lower tumor purity may not be detected.

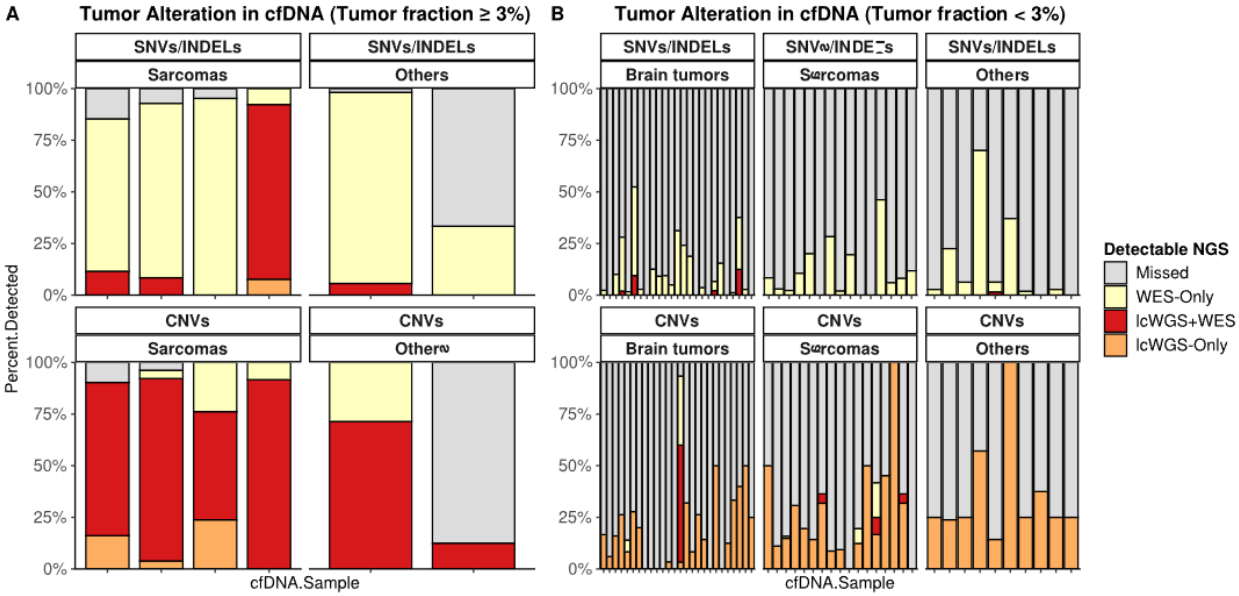


Figure 31: Comparison between lcWGS and WES fo cfDNA in detecting CNVs and point mutations: The deeper coverage of WES allows detection of point mutations (SNVs and INDELs). Approximately 80% of point mutations were detected by WES in cfDNA with high tumor-fraction (A). Both WES and lcWGS detected a comparable amount of CNVs (B). In samples with low tumor-fraction, a lower number of point mutations (C) and CNVs (D) were detected by both WES and lcWGS.

The sensitivity of 50 cfDNA samples with low ctDNA content decreases in both WES and lcWGS. WES detected 11.1% (132/1181) of point mutaitons in brain tumors, 9.0% (94/1042) in sarcomas, and 8.9% (67/752) in other childhood cancers (Figure 31C). Only 5 mutations in brain tumors and 1 mutation in a germ cell tumor were detected with lcWGS. As for CNVs, 16.2% (78/480) were detected with lcWGS in brain tumors, 22.1% (113/511) in sarcomas, and 30.3% (30/99) in other pediatric cancers (Figure 31D). Meanwhile, only 6% (29/480) of CNVs were detected in brain tumors, 2.7% (14/511) in sarcomas, and none of the other cancers were detected in cfDNA with WES.

4.3.3 Whole-exome sequencing allows detection of druggable mutations

Because the coverage of WES enables detection of tumor CNVs, SNVs and INDELs, we wanted to investigate the application of WES in tumor mutation detection, particularly for druggable genes from cfDNA samples. We extracted mutations in 367 genes that could be candidates for targeted therapy in pediatric cancer patients (Supplement Table S2). The WES data included individual-matched tumor, germline, and cfDNA from 27 brain tumors, 26 sarcomas, and 13 other pediatric cancers. We performed the tumor-informed process (Method Section 2.4.7) and somatic mutation calling (Method Section 2.2) in cfDNA WES. The number of tumor mutations, druggable mutations, and the detection rate were counted and calculated per cfDNA sample. The variant allele frequency (VAF) of detected tumor variants was calculated as the frequency of variant-supporting reads found in all supporting reads. Table 2 shows the descriptive statistics of VAF per cancer type. We collected the point mutation status of druggable genes from tumor and cfDNA WES. The point mutation status was visualized using the Oncoplot function of the ComplexHeatmap R package [147] (Figure 32). We found that approximately half of the tumor genomes contained druggable mutations, including 16 brain tumors, 16 sarcomas, and 6 other pediatric

cancers.

Disease Types	Max.VAF	Median.VAF	Mean.VAF	Min.VAF
Brain tumor	30.56	0.81	2.45	0.09
Sarcomas	92.11	10.58	14.69	0.35
Other Cancer	21.21	2.06	4.19	0.11

Table 2: Variant allele frequency (%) of tumor mutations detected in WES of cfDNA; Variant Allele Frequency (VAF)

Among 18 cfDNA samples from brain tumors (Figure 32A), the average tumor mutation detection rate is 11.3%. We detected at least one druggable mutation in 10 cfDNA samples. PIK3CA is the most frequently detected gene in both tumor and cfDNA, while PLK4 is most frequently detected in cfDNA. We found 2 cfDNA samples containing multiple mutations that are not present in the primary tumor genome. We checked their genotypic fingerprint with the matching tumor genome and confirmed that it was not an individual-mismatch variant calling error. It is possible that the cfDNA containing these mutations was secreted by the refractory tumor.

Increasing detection rates were observed in 17 cfDNA samples derived from sarcoma patients (Figure 32B). The average tumor mutation detection rate was 32%. We detected druggable mutations in 9 cfDNA samples. 6 samples contain mutations present only in the primary tumor; 3 samples have additional druggable mutations. Interestingly, we found two cfDNA samples (2LB-037-P01.01 and 2LB-019-P01.01) derived from desmoplastic small round cell tumors that contain multiple extra druggable mutations.

In 11 cfDNA samples obtained from patients with other pediatric cancers, we found a lower number of drug-effective mutations (Figure 32C). The mutation detection rate is 27% on average. Druggable mutations are presented in 5 samples. One sample, obtained from a patient with neuroblastoma, contains all 3 tumor druggable mutations including mutations in the CTNNB1, ALK, and ATM genes. We followed a set of 5 serial liquid biopsies (2LB-049-P01 to 2LB-049-P05) from a patient with hepatoblastoma. The genome of the tumor contains deleterious mutations in 4 druggable genes, including CTNNB1, FBXW7, PTCH1, and FLT1. The mutation in CTNNB1 was detected only in the first biopsy (2LB-049-P01), while the mutation remained undetected in other liquid biopsy samples.

We obtained a sample from a patient with bilateral Wilms tumor (2LB-053-P01). The tumor genome does not have a druggable mutation. However, a deleterious mutation in the druaggable gene NOTCH2 was found in the patient's cfDNA. This cfDNA has been shown to contain a variety of aberrations that are not present in the primary tumor (Section 4.6.3). The possible source of the distinct alterrations in the cfDNA could be the tumor in another kidney or at a distant metastatic site in the liver, lymph nodes and abdominal wall.

Copy-number status (amplification, neutral or deletion) was extracted based on genomic position. In sarcomas, brain tumors, and other childhood cancers, CNVs were detected in tumors at an average rate of 33.7%, 29.1%, and 39.1%, respectively. Overall, the rate of CNVs detected is 30.4%. The most frequently detected genes include MMP9 (associated with tumor invasion, metastasis, and modulation of the tumor microenvironment [148]), AURKA (oncogene that promotes tumorigenesis in many cancers including solid tumors and hematologic malignancies [149]), and EIF4E (oncogene involved in multiple hyperactive signaling pathways promoting tumorigenesis [150]).

4.3.4 Panel-sequencing of cfDNA provides more sensitivity in detecting druggable point mutations

We have investigated the utility of Panel-seq in detecting SNVs and INDELs in tumors, particularly

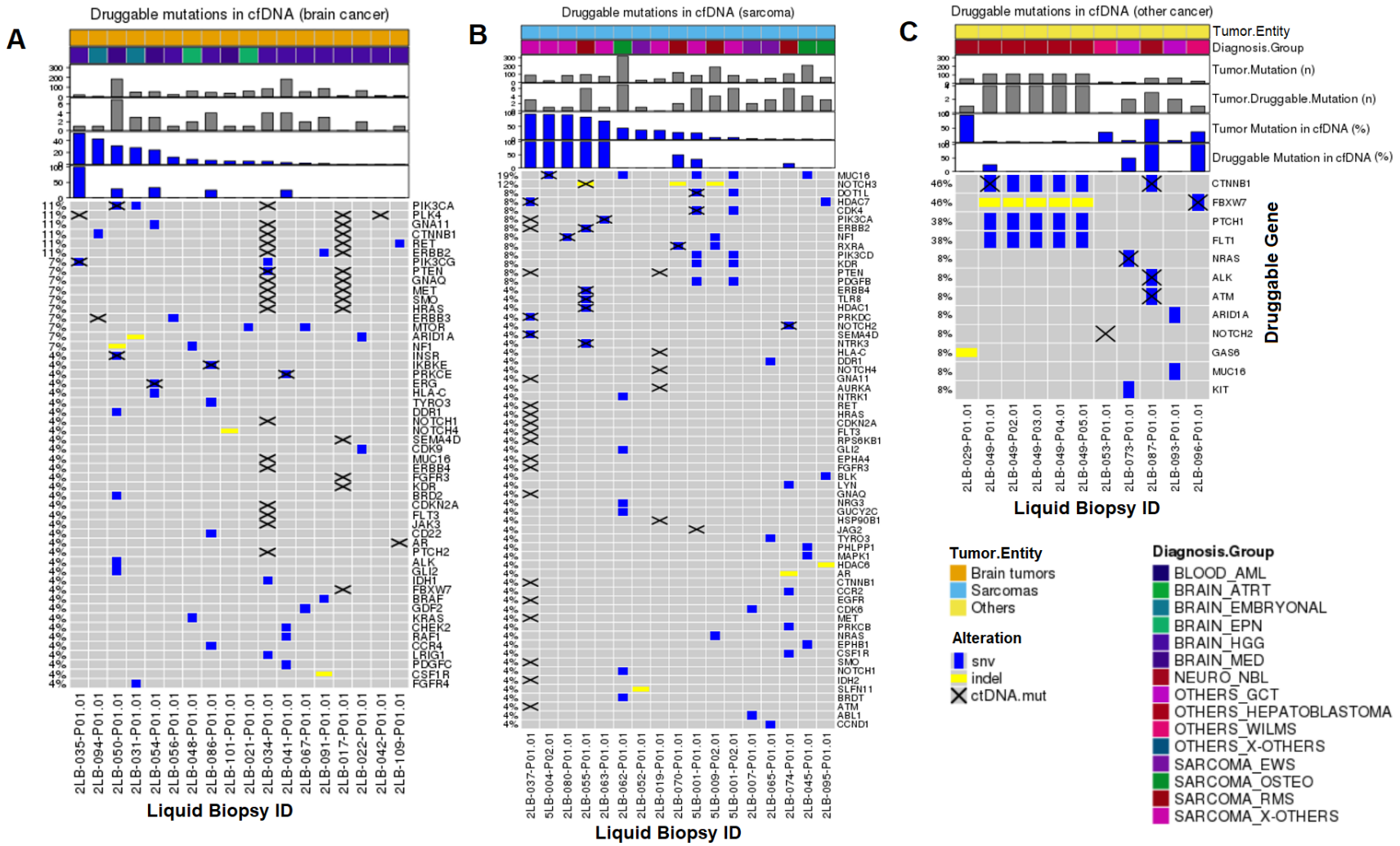


Figure 32: Mutations in druggable genes detected in cfDNA from WES in (A) brain tumors, (B) sarcomas, and (C) other pediatric cancers patients: This oncoplot shows genetic alterations detected in tumors WES including SNVs (blue), INDELs (yellow) and in cfDNA (crossed tiles). The header annotation includes the number of tumor mutations and the number of mutated druggable genes (grey barplots), as well as detection rates of matching cfDNA (blue barplots). Only samples from patients who have druggable mutation are shown.

in druggable genes. Overall, Panel-seq can detect tumor mutations with a low allele frequency in plasma cfDNA. The gene-panel includes 261 genomic loci with a library size of 897,805 bases. We performed only the tumor-informed process (Method 2.4.7) by using somatic functional mutations from tumors WES. Data from WES included individual-matched tumors WES and cfDNA Panel-seq from 44 brain tumors, 15 sarcomas, and 15 other pediatric cancers. The gene-panel captured at least 1 somatic deleterious point mutation in 19 brain tumors, 7 sarcomas, and 7 other pediatric tumors. We found 24 cfDNA samples (9 brain tumors, 11 sarcomas, and 4 other cancers) with at least one tumor mutation.

Table 3 shows the descriptive statistics of VAF in cfDNA. Tumor variants were found with very low frequency in brain tumors. The VAF of brain tumor variants ranged from 1.23% to 0.04% (median = 0.14%). In contrast, tumor variants of sarcomas and other cancers had a higher VAF. VAF of detected variants ranged from 63.03% to 0.7% (median = 10.60%) in sarcomas and from 30.21% to 1.08% (median = 3.1%) in other childhood cancers. This shows that the deep sequencing strategy could detect the tumor variant with an allele frequency as low as 0.1% in a liquid biopsy sample.

Disease Types	Max.VAF	Median.VAF	Mean.VAF	Min.VAF
Brain tumor	1.23	0.14	0.32	0.04
Sarcomas	63.03	10.60	24.58	0.70
Other Cancer	30.21	3.10	7.34	1.08

Table 3: Variant allele frequency (%) of tumor mutations detected in Panel-seq of cfDNA

When considering only druggable genes, only 66 of 367 druggable genes were covered by this gene panel. There were 31 tumor WES (19 brain tumors, 6 sarcomas, and 6 other cancers) that have at least one mutation in druggable genes. The most common druggable genes are CTNNB1, FBXW7, PTCH1, NF1, and MUC16. We detected druggable mutations in the cfDNA of 4 brain tumors, 5 sarcomas, and 4 other childhood cancers (Figure 33). An identical mutation in PIK3CA was found in 3 cfDNA samples from medulloblastoma patients at VAF 0.34% and the other two sarcomas at VAF 10.6% and 4.39%. This demonstates that the Panel-seq can be detected point mutations at very low allele frequency in cfDNA across tumor entities.

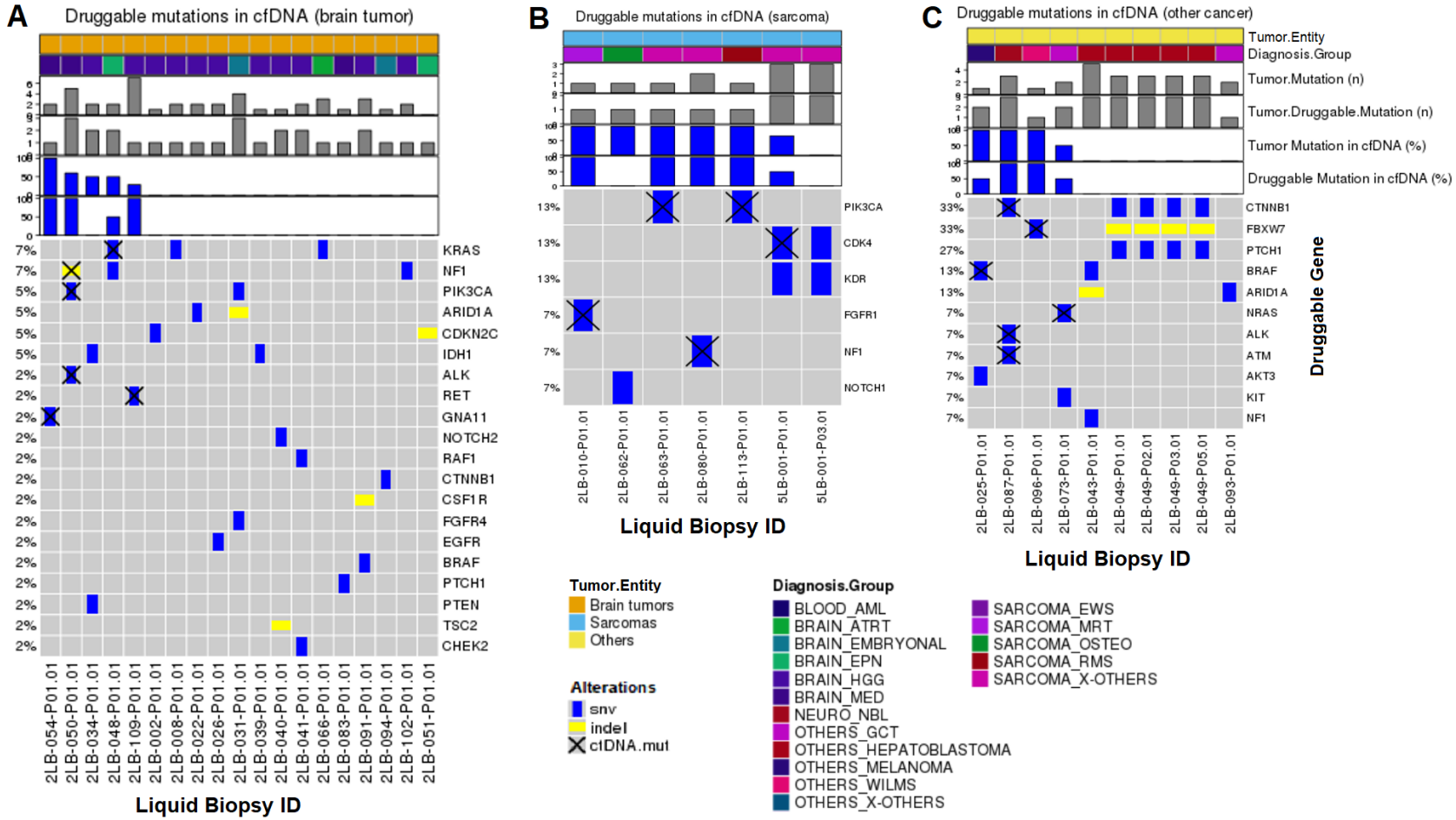


Figure 33: Detected tumor point mutation in druggable genes using Panel-sequencing in (A) brain tumors, (B) sarcomas, and (C) other cancers patients: This oncoplot shows genetic alteration detected from tumor WES including SNVs (blue), INDELs (yellow) and from cfDNA (cross). The header annotation includes the number of tumor mutations and the number of mutated druggable genes (grey barplots), as well as detection rates of matching cfDNA (blue barplots). Only samples from patients who have druggable mutation are shown.

We compared the performance in detecting mutations in druggable genes between WES and Panel-seq (Figure 34). In brain tumors, we detected at least one point mutation in 30% (5/16) with WES and 25% (4/16) with Panel-seq (Figure 34A). In sarcomas, we detected at least one point mutation in 71% (5/7) with WES and 43% (3/7) with Panel-seq (Figure 34B). The majority of mutations were detected by WES and Panel-seq did not exclusively detect the additional druggable mutation. In other childhood cancers, we detected at least one point mutation in 44% (4/9) with WES and 33% (3/9) with Panel-seq (Figure 34C). Similar to sarcoma cfDNA, Panel-seq did not report extra druggable point mutation. Interestingly, WES from a neuroblastoma patient (2LB-087-P01) detected the SNV in CTNNB1 while this mutation was missed by the Panel-seq. Since reported as low TF (TF = 0.8%), the mutation may be missed in the Panel-seq library by chance. Overall, WES can provide broader coverage to detect actionable mutations in non-cranial tumors (sarcomas and other cancers). On the other hand, brain tumors showed the variability of the comparison result. The mutated cfDNA fragments can be missed by chance because of the low concentration of the tumor-derived cfDNA in the blood circulation.

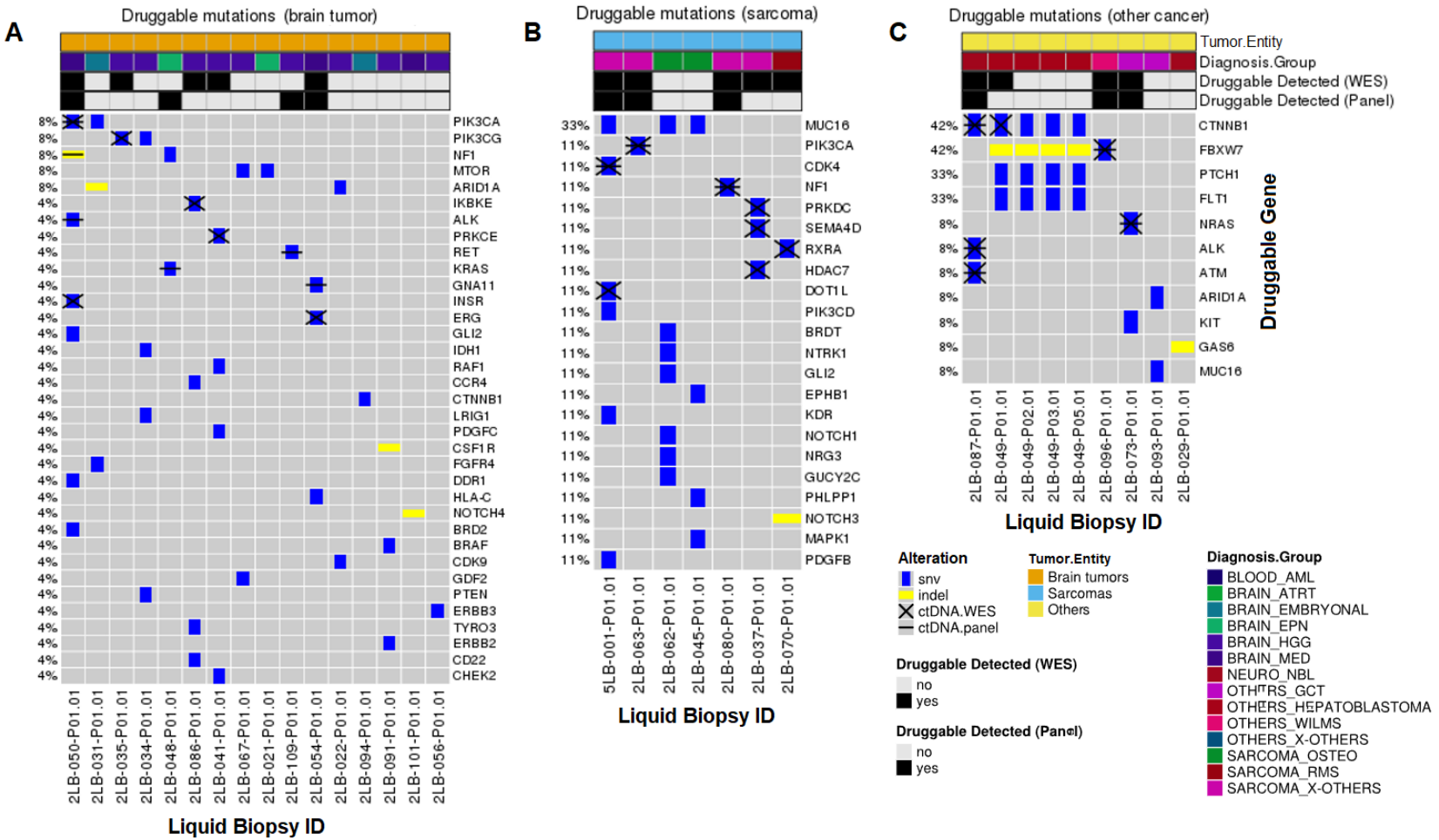


Figure 34: Comparison of druggable mutations detection with WES and Panel of cfDNA in brain tumors (A), sarcomas (B), and other pediatric cancers (C): The oncoprint matrix presents tumor somatic functional SNVs (blue) and INDELs (yellow) in druggable genes. Those mutations were intergated in individual-matched cfDNA if can be detected with WES (cross) or Panel-seq (vertical line).

4.4 CfDNA Fragment Length Analysis with cfdnakit

We developed an R package called cfdnakit. It is specifically designed for analyzing cfDNA sequencing data focusing on extracting the characteristical length of cfDNA, quantify tumor cfDNA contribution, and inferring CNV base-on the short-fragmented cfDNA. The package extracts the fragment length of cfDNA from a sequencing file (BAM file) and creates the fragment-length profile of the sample. For comparison and QC inspection purposes, cfdnakit allows visualization of a fragment-length profile and comparing between multiple cfDNA profiles. This section describes the application of the package. First, we compares the fragment-length profile of tumor-derived cfDNA with non-malignant cfDNA in the PDX experiment. Moreover, cfdnakit is also used to explore the fragment length profiles of cfDNA in the pediatric cancer cohort. Finally, genome-wide fragment-length are explored and used as a signal to infer tumor CNVs by using the proportion of short-fragmented cfDNA.

4.4.1 Circulating tumor Cell-free DNA is shorter than cfDNA from non-malignant cells

We extracted human-derived cfDNA from plasma cfDNA of mice with patient-derived xenograft (PDX) cell-lines (Figure 35A) by separating sequencing reads mapped onto human chromosomes (GRCh37) from those mapped onto mouse chromosome (GRCm38). When a sample was reported having high tumor fraction (Tf $>$ 3%), the genomic copy-number profile of the human-derived cfDNA (Figure 35B) was similar to the genome of the tumor (Figure 35C). This similarity could confirm that human-derived cfDNA was secreted from tumor cells.

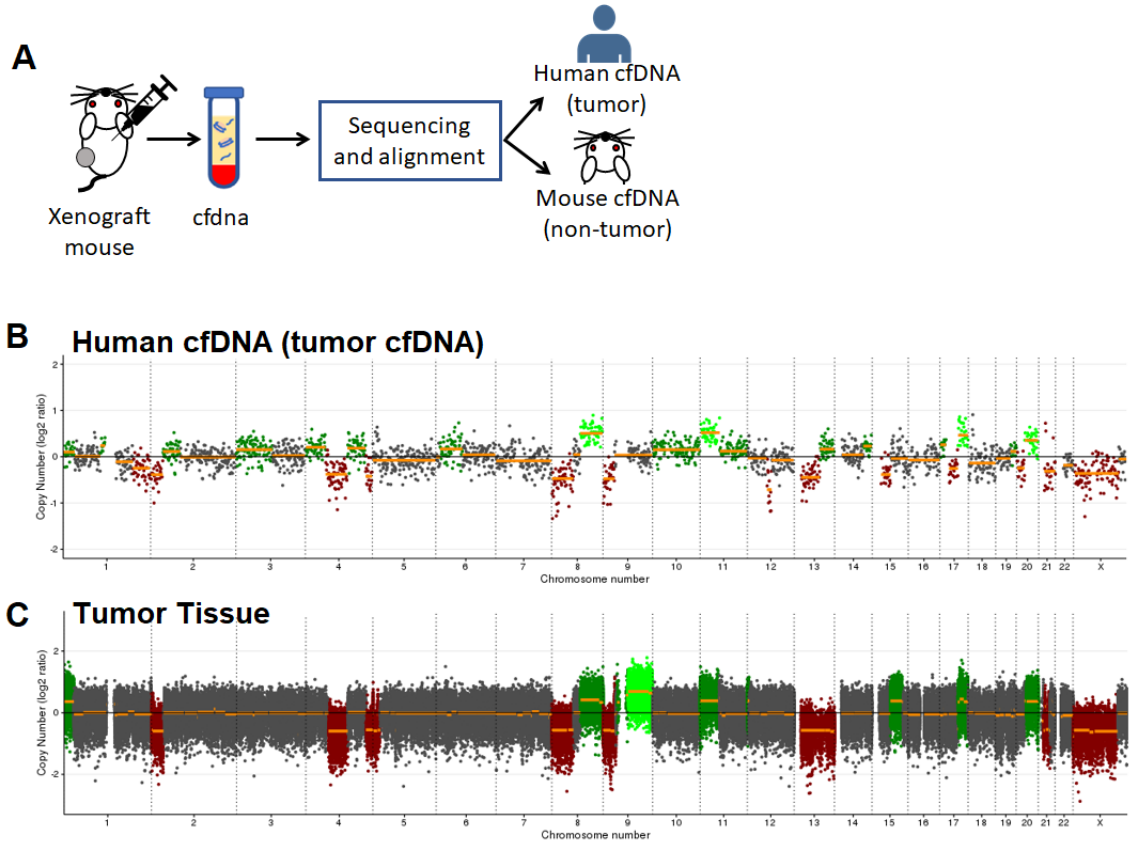


Figure 35: Extraction of tumor-derived cfDNA from a patient-derived xenograft liquid biopsy: (A) A collection of cfDNA from a mouse model with xenografted pediatric cancer cells allows characterization of cfDNA released by cancer cells (reads aligning onto the human genome) from the DNA released by non-malignant cells (reads aligning onto the mouse genome). The result of reads separation is confirmed by the similarity of CNV genomic profile between human-derived cfDNA (B) and tumor DNA (C). Colors in a genomic profile represent CNV events (grey:neutral, red:deletion, green:gain (3N), light green:amplification ($>$ 3N)).

Using cfdnakit, the length of cfDNA fragments was extracted from their alignment information. The fragment-length distribution plot showed that the human-derived cell-free DNA was shorter than mouse-derived cfDNA (Figure 36A). The size of ctDNA was distributed between 80 - 150 base pairs with the peak at $\approx$142 base pairs. Meanwhile, the mouse-derived cfDNA showed the modal length of 167 base pairs with a 10-bp periodical peak among fragments shorter than 150 base pairs. The modal length of the human-derived cfDNA was around 142 bases and is significantly shorter than the modal length of the mouse-derived cell-free (Wilcoxon rank sum test; p=0.024) (Figure 36B). The finding supported the observation in many experiments in adult cancers [110, 151] that tumor-derived cfDNA are relatively shorter than cfDNA from non-malignant cells. The fragment-length characteristic has been used as a quantitative tumor marker in many tumors [110, 151]. The most recent application is to apply the size-selection method to increase the success of tumor mutation detection from liquid biopsies.

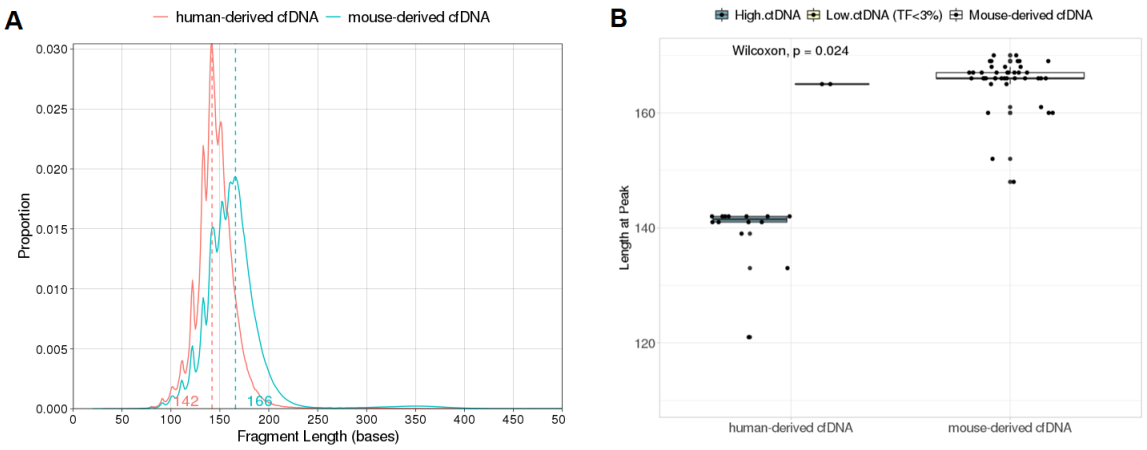


Figure 36: Comparison of fragment-length between tumor-derived cfDNA and non-tumor-derived cfDNA in the PDX experiment: (A) Fragment-length distribution plot of a human-derived cfDNA and a mouse-derived cfDNA. The human-derived cfDNA is shorter than the cfDNA from non-malignant origin. (B) The modal length of cfDNA released by cancer cells (human-derived) and non-maligant cells (mouse-derived).

4.4.2 Short-fragment size-selection in-silico enriched copy-number aberration detection in plasma cfDNA

Since the previous section indicated that the plasma tumor-derived cfDNA is shorter than non-malignant cfDNA, the success of detecting tumor genomic aberration can be increased by selecting only the short-fragmented cfDNA. We performed in-silico size-selection to the lcWGS of cfDNA in the pediatric cohort. The short-fragmented cfDNA was extracted in-silico by selecting cfDNA fragments having a size less than 150 base pairs (Figure 37A).

Using cfdnakit, the lengths of cfDNA fragments were extracted from the alignment information before and after in-silico size selection (Figure 37B). The fragment length distribution showed the clear cut at fragment length 150 observed from samples having in-silico size-selection. The enrichment of short-fragmented cfDNA by the in-silico method enhanced the log2 ratio of genomic regions with copy-number aberrations found in the tumor genome (Figure 37C). The estimated tumor fraction increased from 12% without size selection (Figure 37D) to 36% after size-selection (Figure 37E). By applying in-silico size-selection, we can in general increase the detection rate of CNVs from lcWGS of cfDNA. However, this is possible only when the read-coverage of the sequenced sample is high enough. Otherwise, the result could rather due to noise which will increase the rate of false positives.

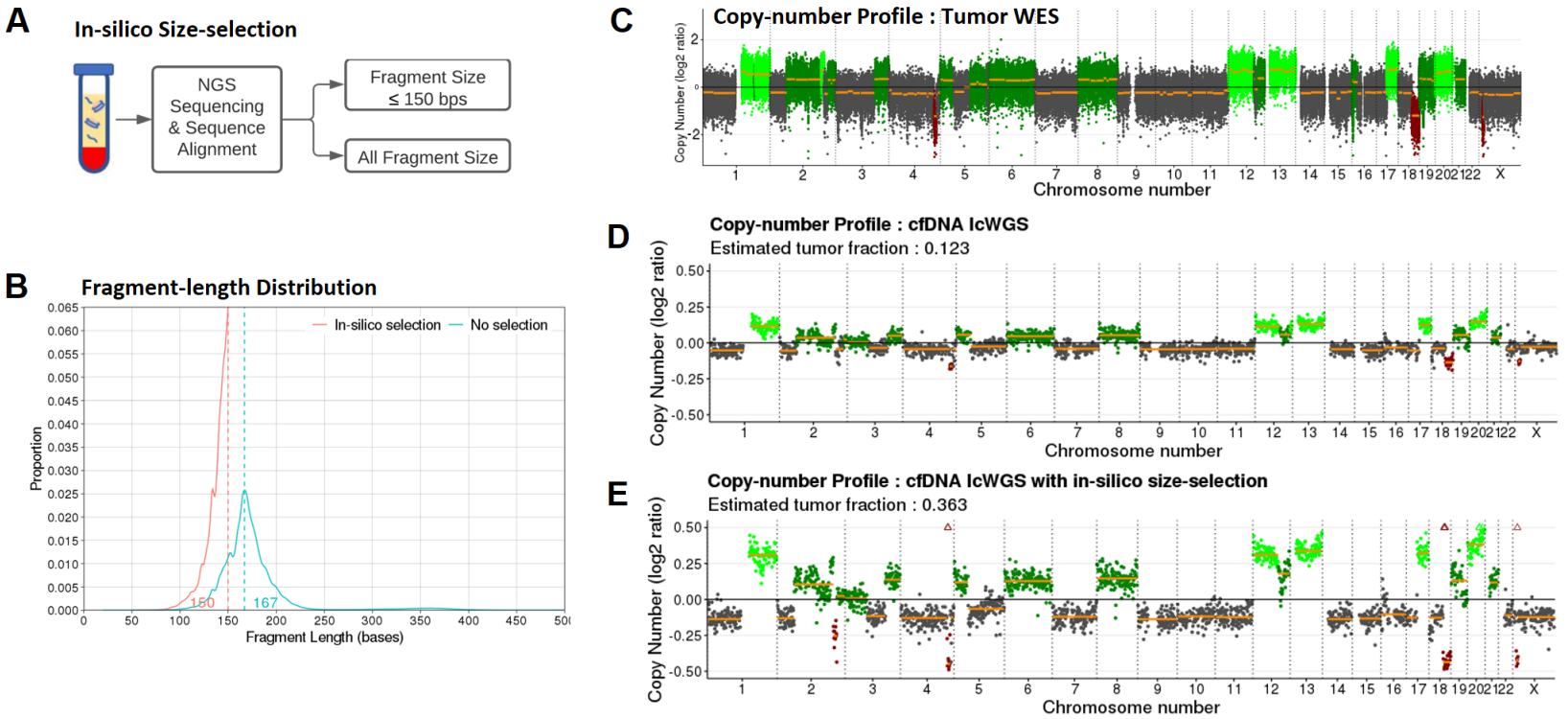


Figure 37: In-silico size-selection enhances the detection of tumor copy-number aberration in cfDNA: (A) Short-fragmented cfDNA ($\leq$150 bases) were separated from a regular cfDNA sequencing. (B) Fragment-length distribution of a cfDNA after the in-silico size-selection (red line) and a regular cfDNA without size-selection (blue). (C) The genomic CNVs profile were determined from the tumor genome using WES. Genomic CNV profile and estimated tumor fraction of cfDNA without size-selection (D) and with size-selection (E). The cfDNA with size-selection for short-fragmented cfDNA enhances the CNV signal and increases the tumor fraction estimates. Colors in a genomic profile represent CNV events (grey:neutral, red:deletion, green:gain (3N), light green:amplification ($>$ 3N)).

4.4.3 Short-fragmented cfDNA correlates with the copy-number aberration

We expected that the enrichment of short-fragmented cell-free DNA per genomic loci correlates with the number of copies in the tumor genome. To demonstrate the relation, we selected a cfDNA sample derived from an embryonal rhabdomyosarcoma patient showing multiple copy-number alterations and high estimated tumor fraction (Figure 38A). Cfdnkit reported the sample having a short-fragment ratio of 1.03 which is approximately 5 times more than the average of healthy individuals. The short-fragment ratio per 1 MB was extracted and visualized by cfdnakit (Figure 38B). It shows that a short-fragment ratio of a genomic segment is increasing in the amplified segment and decreasing when the segment is lost. The copy-number log2 ratio and the short-fragment ratio is highly correlated (Pearson correlation 0.95; 95% CI [0.949,0.956]). Moreover, it shows that the short-fragment ratio is increasing accordingly with the number of copy-number aberrations reported by ichorCNA (Figure 39).

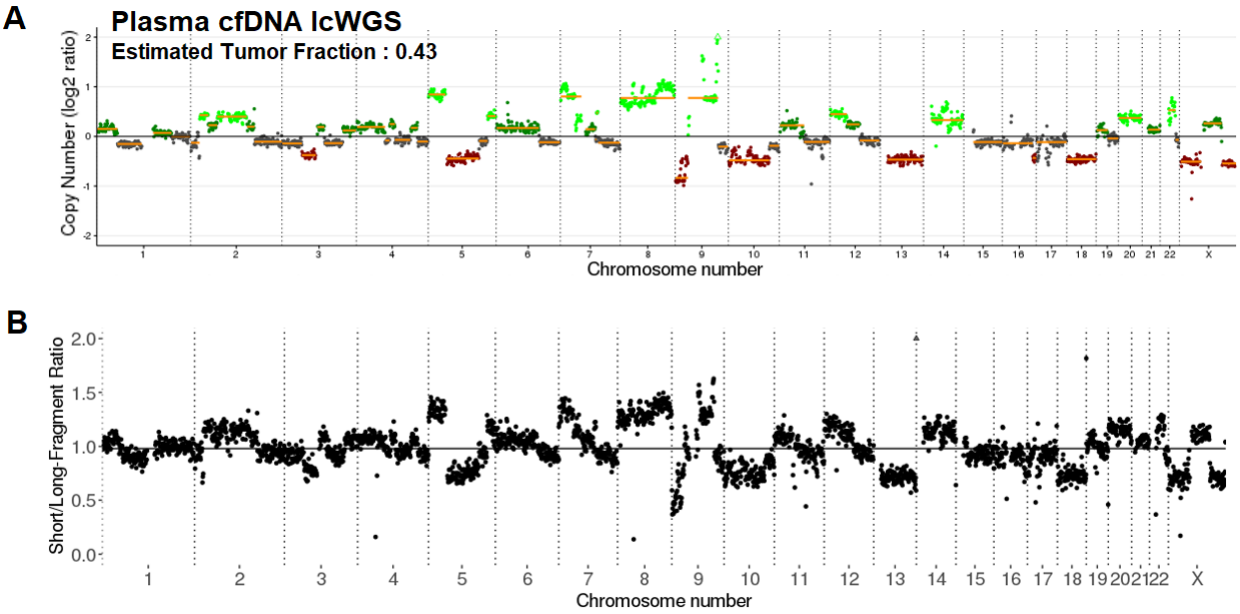


Figure 38: Correlation between short-fragment ratio and copy-number alteration per 1 Mb non-overlapping windows: (A) Genomic CNV profile of a high-TF cfDNA from a patient with embryonal rhabdomyosarcoma. Colors in the genomic profile represent CNV events (grey:neutral, red:deletion, green:gain (3N), light green:amplification (> 3N)). (B) Short-fragment ratio extracted by the cfdnakit package.

4.4.4 CPA score is asssociated with both copy-number aberration and tumor mutational burden

Cfdnakit transformed those short-fragment ratio per 1 MB into normalized score (z-score) using the Panel-of-Normal (PoN) dataset (Method Section 3.7). Similar to the CNV calling workflow of lcWGS, the PoN included a group of selected cfDNA samples without large CNV (Method Section 2.4.4). The z-score could represent the aberration of short-fragment cfDNA at a locus of the sample compared to the PoN dataset. The segmentation using PSCBS packages has been performed through z-scores and created continuous genomic segments; each showing aberration of short-fragment cfDNA. We performed CNV calling and tumor fraction estimation using these segmentation result (Method Section 3.9).

Cfdnakit finally reported the copy-number aberration (CPA) score. This score were calculated as average of segment z-scores multiplied by short-fragment ratio of the sample (Method Section 3.10). The CPA score can be used as a qualitative score to detect cfDNA that contains a certain level of tumor-derived DNA. To demonstrate the relationship, the following experiment and measurement have been

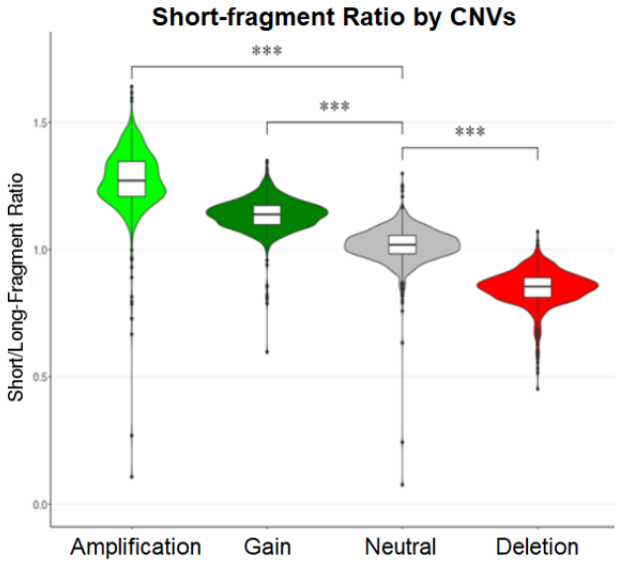


Figure 39: Short-fragment ratio by copy-number aberrations: Short-fragment ratio increases accordingly with absolute CNVs found in cfDNA.

performed in the dataset of 34 WES cfDNA that have matched lcWGS data. This dataset contains cfDNA from 14 brain tumors, 10 sarcomas, and 10 other pediatric cancers. The TF of each lcWGS was reported by ichorCNA. In total, there are 30 low-ctDNA (TF$<$3%) and 4 high-ctDNA. For this dataset, we find a correlation between the CPA score and the estimated tumor-fraction (Pearson correlation 0.89; 95% CI [0.84,0.95]) (Figure 40). The correlation coefficient declines in low-ctDNA samples, (Pearson correlation 0.35; 95% CI [0.00,0.63]).

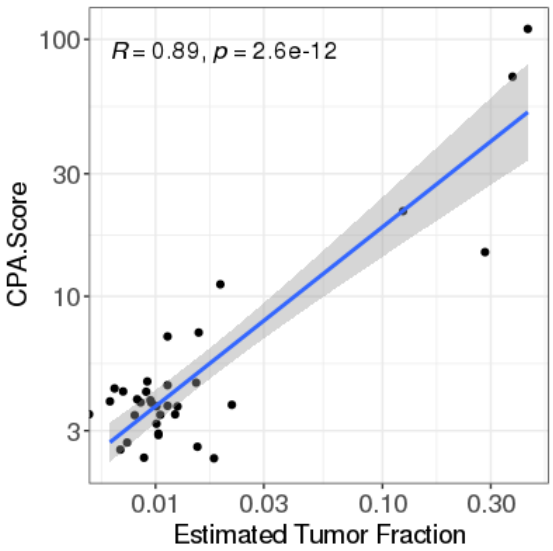


Figure 40: The correlation between CPA score and TF and tumor mutations in 34 lcWGS (Pearson correlation 0.89; 95% CI [0.84,0.95]).

For each WES, the process of mutation calling and the tumor-informed variant detection have been processed. The mutation burden is the total number of somatic functional SNVs supported by at least 5 reads and have $\geq$ 1% VAF. The tumor-informed detection reported the number of detected (variant presented) and undetected tumor somatic functional mutations. The percentage of detection (percent detected) is calculated afterwards. The principal component analysis (PCA) of all cfDNA samples shows

that the CPA score, tumor fraction and mutation burden were contributing to the principal component (PC) 1 while "variant presented" and "percent detected" were more contributing to PC2 (Figure 41A). Those high-ctDNA samples were explained by either having high mutation burden, CPA score and tumor fraction or high variant presented and percent detected. Considering only 30 low-ctDNA, majority of low-ctDNA samples were not assoicated with any variables. The first component of the PCA, although only explaining 46% of variance, was strongly associated with variant presented and percent detect (Figure 41B). The second component explaining 24% of variants had contributions from the variance CPA score and tumor fraction. Only a few number of samples were associated with the CPA score or the tumor fraction.

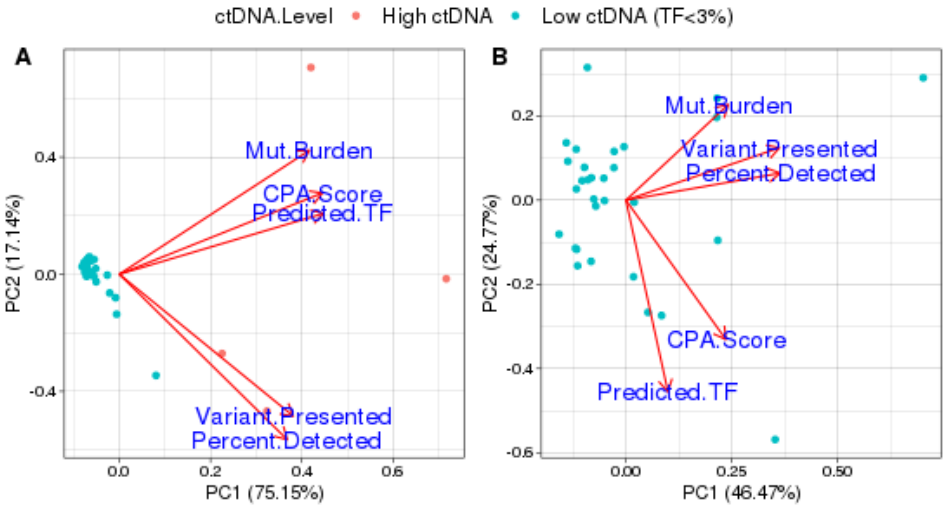


Figure 41: Principal component analysis showing the correlation between estimated TF, CPA Score, and mutational burden. (A) The first two components of all cfDNA samples (n=34). The high-ctDNA samples (red dots) were either associated with high mutation burden, CPA score, and predicted tumor fraction or with high percentage and number of tumor variant detected. Meanwhile, some of low-ctDNA (blue dots) were associated with PC2. (B) First two components of cfDNA with low-ctDNA (TF<3%; n=30). Mut.Burden represents mutatonal burden. Variant Presented is the number of tumor variants detected in cfDNA WES. Percent Detected is the percentage of tumor variants detected. Predicted.TF is the estimated tumor fraction reported by ichorCNA.

The correlation of both CPA score and tumor fraction to all 3 mutational variables, namely Mutation Burden, Variant Presented, and Percent Detected, were calculated (Table 4). The correlation coefficient values was comparable between the tumor fraction and the CPA score when including high-ctDNA samples. However, no correlation was found from the tumor fraction value among low-ctDNA samples. In this group, the CPA score shows a weak but stronger correlation to the Variant Presented and Percent Detected. With this finding, we assumed that the CPA score could also be associated with the number of tumor mutations and the detection rate in a patient's cfDNA.

4.4.5 CPA score performed better in detecting high ctDNA

We compared the performance between the CPA score and the tumor fraction value in detecting cfDNA with a high concentration of tumor-derived cfDNA. We divided WES data into two categories: high ctDNA and low ctDNA by using detection thresholds as described (17% of tumor mutations detected, and 3 tumor mutations) (Method Section 2.4.7). The tumor fraction value is not significantly different between high-ctDNA and low-ctDNA (p-value=0.1; T-test) (Figure 42A). Meanwhile, the CPA score shows a clearer difference between high-ctDNA and low-ctDNA (p-value=0.03; T-test) (Figure 42B).

	Mutation Burden	Variant Presented	Percent Detected
All cfDNA (n=34)			
Tumor Fraction	0.87	0.58	0.66
CPA Score	0.9	0.63	0.54
Low ctDNA (n=30)			
Tumor Fraction	0.05	0.08	0.05
CPA Score	0.1	0.27	0.42

Table 4: Pearson correlation between the CPA Score and the Mutation burden, Variant Presented, and Percent Detection in comparison to the tumor fraction.

The tumor fraction variable cannot differ between high-ctDNA and low-ctDNA when considering the number of tumor variant given the overall mutation burden (Figure 42C). On the other hand, the CPA score can differ at least 3 high-ctDNA from other low-ctDNA (Figure 42D).

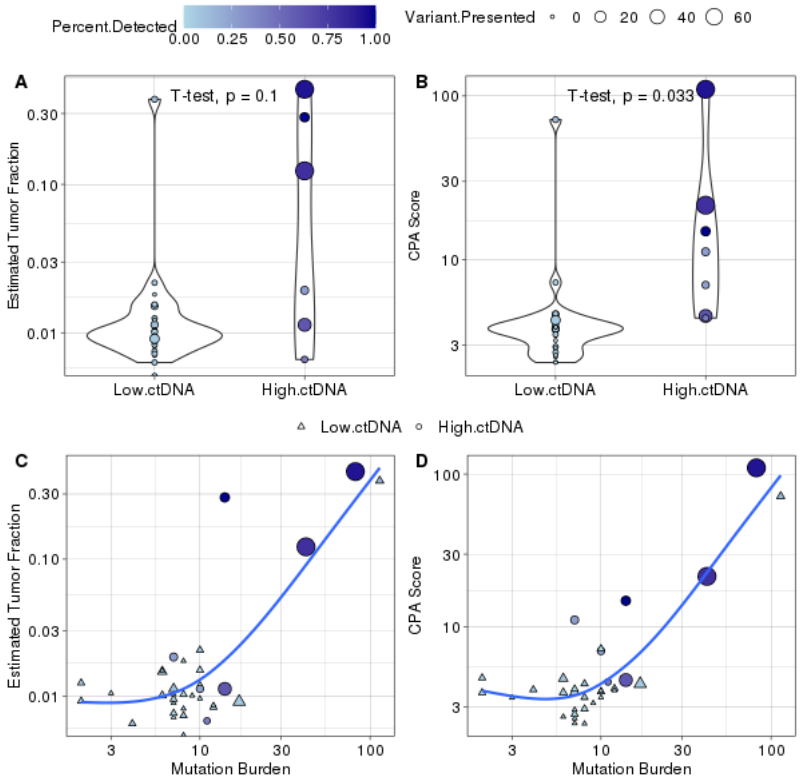


Figure 42: Comparison between the tumor fraction and the CPA score in high-ctDNA and low-ctDNA: (A) Distribution of estimated TF between low ctDNA and high ctDNA samples; (B) Distribution of CPA Score between low ctDNA and high ctDNA samples; (C) A scatter plot showing correlation between mutation burden and estimated TF; (D) A scatter plot showing correlation between mutation burden and CPA Score. Blue line: scatter plot smoothed line using LOESS model.

We calculated the sensitivity and the specificity of the CPA score and the tumor fraction in discriminating high-ctDNA and low-ctDNA samples. We manually change the class of sample 5LB-053 (bilateral Wilm's tumor) from low-ctDNA to high-ctDNA regarding the tumor heterogeneity. Receiver operating characteristic (ROC) curves were used to virtualize and calculate the area under the ROC curve (AUC) (Figure 43). It is demonstated that the CPA score (AUC=0.97) performs better than the estimated

tumor fraction (AUC=0.81) in detecting high-ctDNA samples.

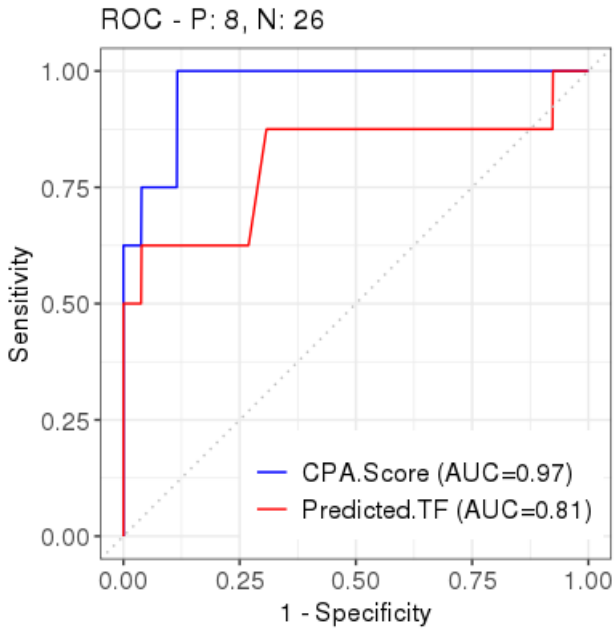


Figure 43: ROC curve showing the performance of CPA score and ichorCNA TF in detecting cfDNA with high tumor mutations.

4.4.6 CPA score of the pediatric cancer cohort

We analyzed lcWGS of cfDNA in the pediatric cancer cohort with cfdnakit. The CPA score has been calculated per sample. We found a high correlation between the CPA score and the tumor fraction (TF) reported by ichorCNA (Pearson correlation : 0.82; 95% CI : [0.76,0.88]) (Supplement Figure S2). The distribution of CPA scores in the cohort has shown a difference between cfDNA from healthy donors and cfDNA from cancer patients (Figure 44A). The CPA score of healthy cfDNAs (median=2.14) is lower than low ctDNA samples (median=4.23) and high ctDNA samples (median=29.3). Compared to the short-fragment ratio (Figure 52B), the CPA score can differentiate cfDNA of healthy donors from cfDNA of patients. In high ctDNA samples. CPA score of sarcoma samples were highest (median = 29.3; n=11) comparing to brain tumors (6.68; n=1) and other cancers (71.6; n=1) (Figure 44B). Those CPA scores of low ctDNA samples were indiferrent between tumor entity (median CPA score 4.38, 3.46 and 4.23). Supplement Table S4 shows CPA score per tumor entity and tumor fraction.

The utility of the CPA score in guiding the detection of tumor point mutations with WES is shown in Figure 45. In brain tumors, high CPA scores (score $>$ 6; false positive rate 0.14) were found in two cfDNA samples (Figure 45A). In particular, a cfDNA from patient with high-grade gliomas detected 11 tumor mutations (52% of all tumor mutations) and 2 druggable mutations in PLK4 and PIK3CG. This sample would not have been detected by using TF as a guiding measurement value. In sarcomas and other pediatric cancers, the CPA score correlates with the mutation burden, and the percentage of tumor mutations detected by WES of cfDNA (Figure 45B and C). The detectability of the CPA score and TF is comparable especially those high TF samples. Using a CPA score of 6 as the threshold, we detect an additional cfDNA from a patient with Wilms tumor (Figure 45C). This cfDNA contains 10 tumor point mutations (37% of all tumor mutations) and a mutation in FBXW7 druggable gene. Overall, the CPA score could increase the sensitivity of cfDNA WES as a guiding measurement to determine the success of detecting tumor alterations and estimation of mutational burden.

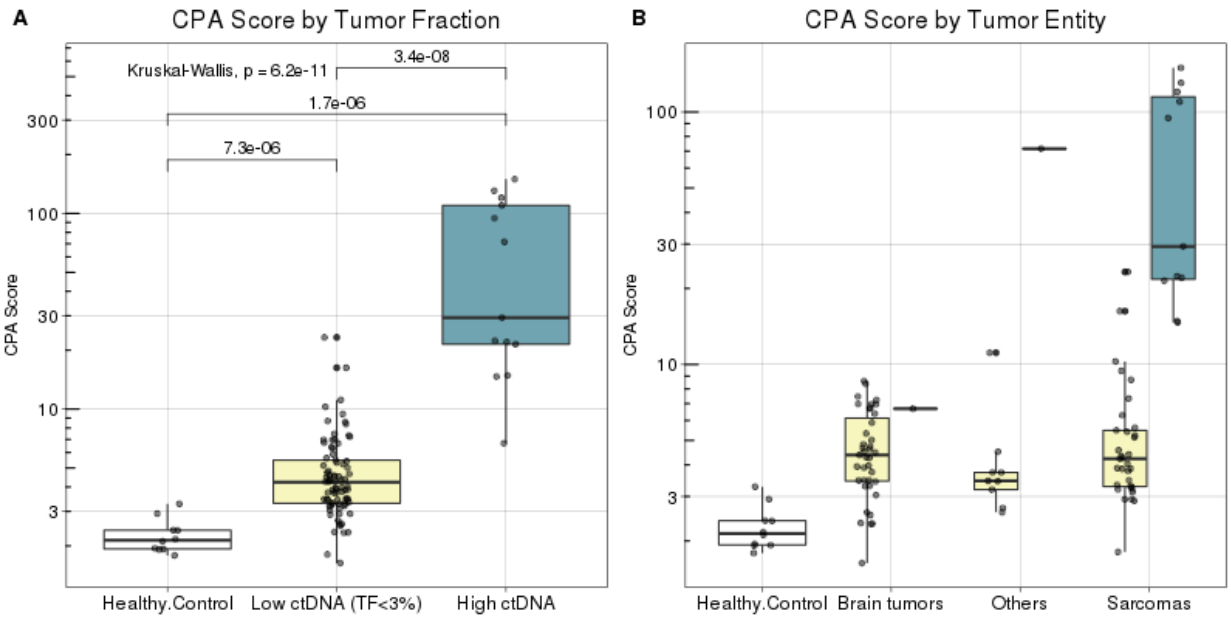


Figure 44: Distribution of CPA Score of plasma cfDNA samples in the pediatric cohort: (A) The distribution of CPA Score in cfDNA samples grouped by sample's esitmated tumor fraction. CPA scores of cfDNA of patients are higher significantly than healthy donors (Wilcoxon rank sum test). (B) Distribution of CPA Score per tumor entities of cfDNA samples in the pediatric cohort. High CPA scores were commonly found among cfDNA from sarcoma patients.

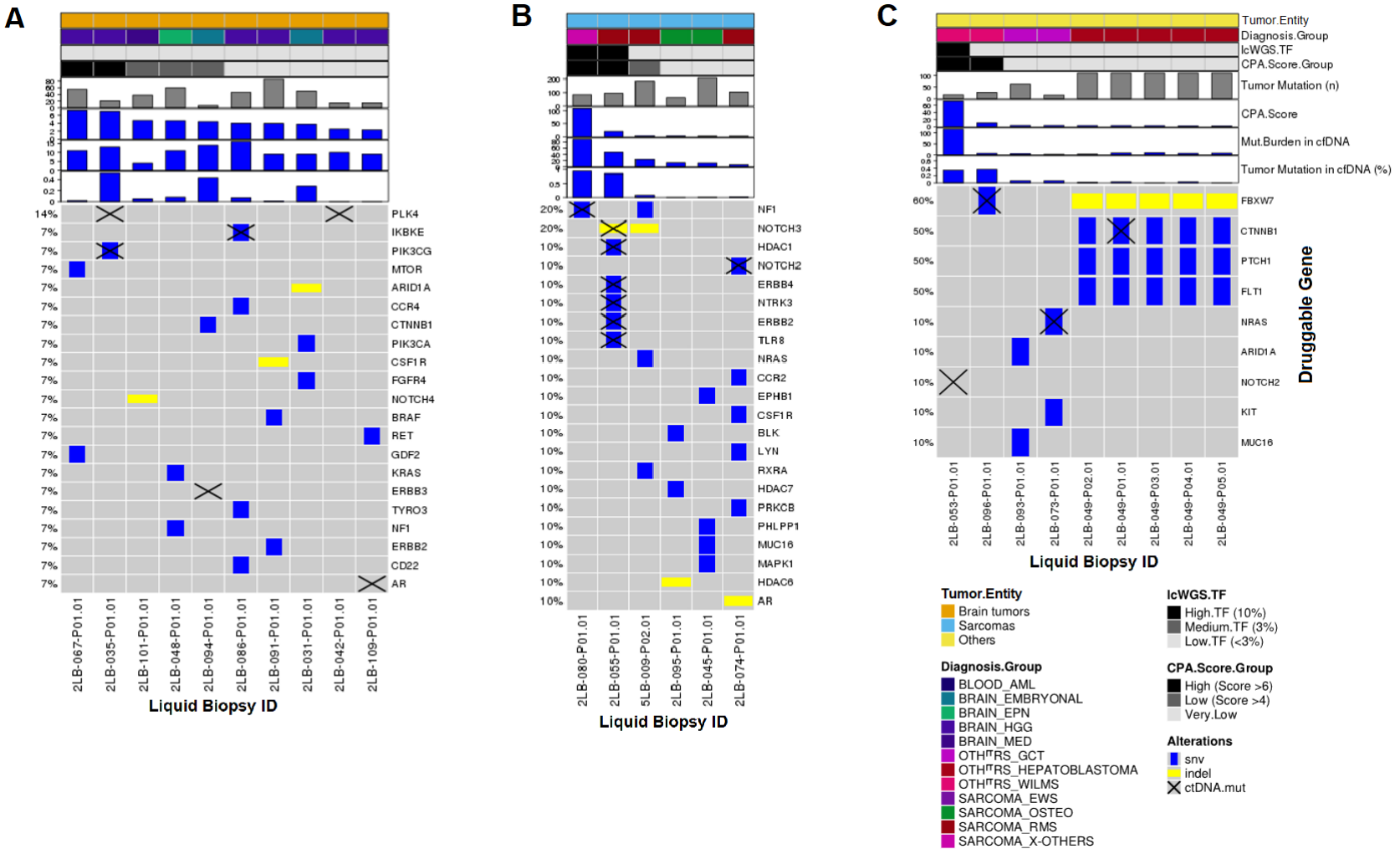


Figure 45: Druggable mutations detected guided by CPA score and estimated tumor fraction: The oncoplot shows the matrix of druggable mutations ordered by CPA score of samples from brain tumor (A), sarcoma (B), and other childhood cancer (C). The genetic alterations were detected from tumor WES including SNVs (blue), INDELs (yellow) and from cfDNA WES (cross).

4.5 A Preliminary Analysis of Detecting Telomeric Alterations with Liquid Biopsy CfDNA

This section presents the result of telomeric alteration analysis from liquid biopsy cfDNA. Here, lcWGS data of tumor and plasma cfDNA was analyzed by using TelomereHunter software (Methods Section 2.6). We compared the estimated telomere content and normalized count of telomeric variant repeats (TVRs) between cfDNA of patients and healthy donors. We demonstrate the possibility of using sequencing data of cfDNA to track telomere shortening and detect integration of TVRs.

4.5.1 Telomere elongation and telomeric variant repeats were found in some brain tumors and sarcomas

First, we explored telomeric aberration of 110 tumor samples in the pediatric cohort, including 51 brain tumors, 48 sarcomas, and 11 other pediatric tumors. Using individual-matched tumor/control lcWGS data, TelomereHunter calculated telomere contents and reported it as the ratio of tumor over control. Sequencing reads were classified as telomeric reads when six non-consecutive repeat types (t-type, c-type, g-type, or j-type) or their reverse complements appear in a 100 bp read. The telomere content was calculated as intratelomeric read counts normalized by the total number of reads having similar GC composition. We found that most tumors had a decreasing telomere content compared to their matched control (Figure 46A).The average telomere content log2 ratios were -0.26 (95% CI [-0.54,0.00]) in brain tumor, -0.21 (95% CI [-0.36,0.10]) in sarcoma and -0.59 (95% CI [-0.98,0.19]) in other cancers. There were 13 brain tumors (25%), 9 sarcomas (16%), and 1 other cancer (9%) with increasing telomere content (log2 ratio $>$ 0.5). Among those diagnostic types with an increased or stable telomere content were high-grade gliomas (HGG), germ cell tumors, and osteosarcomas (average telomere content log2 ratio = 0.01, -0.05, and 0.77, respectively) (Figure 46B).

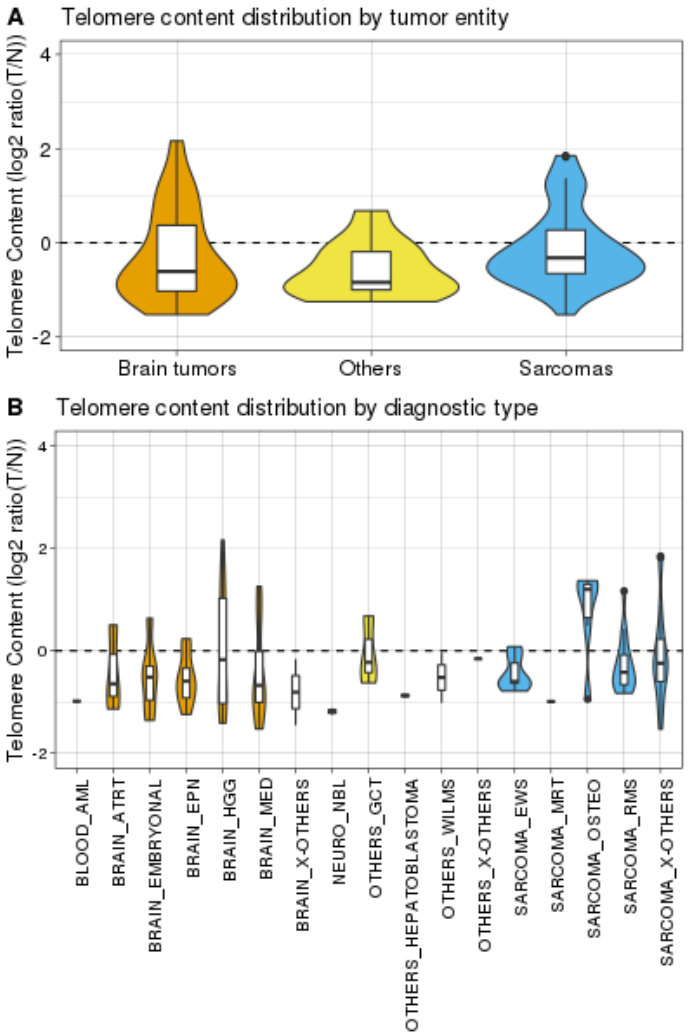


Figure 46: Distribution of telomore content log2 ratio of tumors in the pediatric cohort: (A) Distributon of telomere content of all tumor entities; (B) Distribution of telomere content of all tumor diagnostic types

We identified tumor samples with deleterious somatic point mutations in ATRX, DAXX, H3F3A, TERT, TP53, IDH1, and IDH2 from matched tumor WES (Supplement Figure S3). Those genes are associated with telomere maintenance mechanisms (TMM) or alternative lengthening of telomeres (ALT) in brain tumors [152–154]. In total, there were 44 samples with at least one mutation in those genes. Point mutations in ATRX were found in 4 brain tumors (3 HGGs and 1 diffuse intrinsic pontine glioma) and 1 osteosarcoma. Additional mutated H3F3A was found in 18 brain tumors. No samples had a point mutation in IDH1, IDH2, DAXX, and TERT. Lastly, TP53 is the most frequently mutated gene and was found in 29 samples. Interestingly, all samples with a mutation in both ATRX and TP53 (ATRX/TP53) had increased telomere content.

Since alternative lengthening of telomere (ALT) leads to increased integration of TVRs into telomeres, TelomereHunter extracted and calculated the normalized count of TVRs in both intratelomeric and subtelomeric regions. In this study, we focused on the number of each 5 common TVR singletons (variant hexamers surrounded by at least three t-type repeats) in intratelomeric regions. The normalized count of TVR singletons generally increased with telomere content increase in brain tumors and sarcomas (Figure 47). In brain tumors, TGAGGG and TTCGGG singletons were frequently found in tumors having mutations in both ATRX and TP53 (ATRX/TP53). On the other hand, the integration

of TGAGGG singletons was relatively stable. The ATRX-mutated osteosarcomas did not show any particular enrichment of any TVRs. The insertion of TVRs in pediatric sarcomas could be mandated by mutations in other genes.

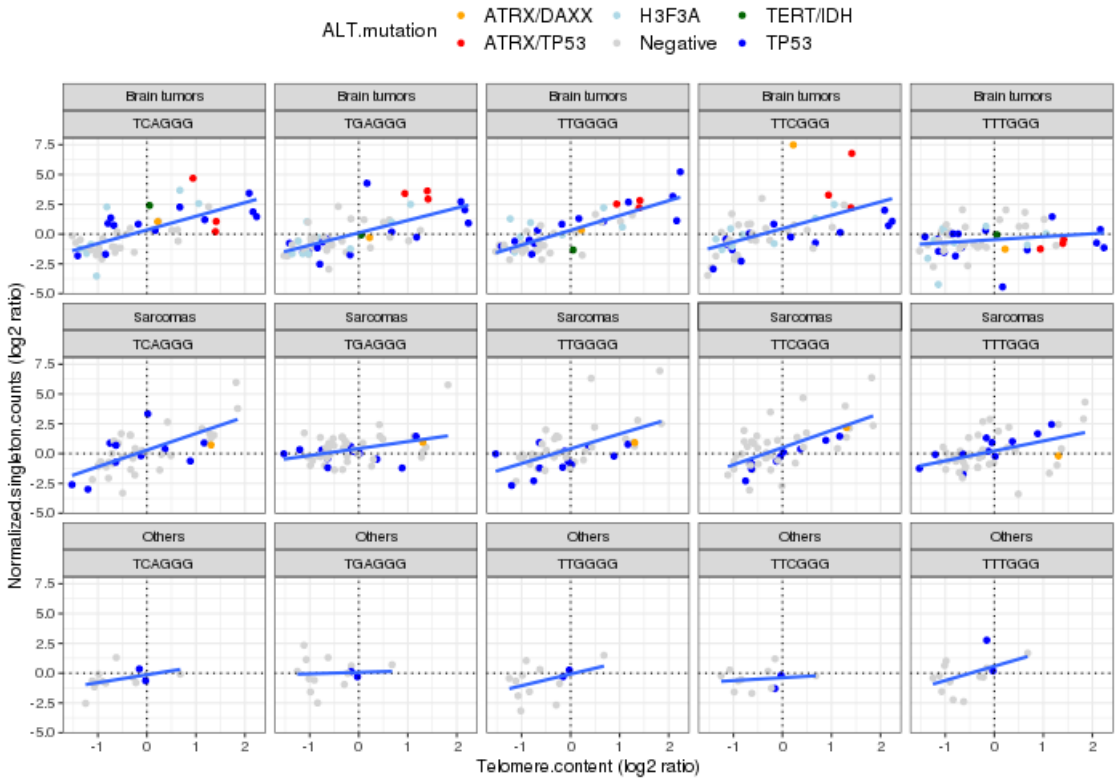


Figure 47: Enrichment of telomere variant repeats of tumor samples in the pediatric cohort. ATRX/-DAXX (orange) represents samples with a mutation in either ATRX or DAXX. ATRX/TP53 (red) are samples with mutations in both ATRX and TP53. Samples with a mutation in TERT or IDH1/IDH2 were named TERT/IDH (dark green).

4.5.2 Telomere content is decreasing in most of patient's cfDNA

Since there were difference in the sample preparation process between cfDNA and tumor samples, we presumed that the telomeric region could be affected by these factors. We applied TelomereHunter to 146 lcWGS datasets of cfDNA and compared their telomere content (number of intratelomeric reads per million reads with telomeric GC content) with matched tumor and control samples. As expected, the telomere content of cfDNA (median=235) is significantly lower than control (median=652) and tumor samples (median=464) (Supplement Figure S4B). Using individual-matched cfDNA/control as the inputs to TelomereHunter might thus not be suitable. Therefore, we analyzed telomeric aberrations in cfDNA without individual-matched control for downstream analysis. Since the coverage of cfDNA lcWGS was relatively low, we checked the correlation of lcWGS coverage with the estimated telomere content. Although showing a weak correlation, the telomere content tended to decline at below 0.4X genomic sequencing coverage (Supplement Figure S4A). Further integrative analysis with matching tumors is required to ensure that ultra-deep sequencing could affect the estimation of telomere content.

We compared the telomere content of cfDNA samples per tumor entity in the pediatric cohort. Using 3% tumor fraction as threshold, we classified 146 cfDNA samples into two classed: low ctDNA (n=124) and high ctDNA (n=22). Similar to tumor samples, telomere content of most cfDNA samples was decreasing compared to cfDNA from healthy donors (median=302; n=10) (Figure 48). Those telomere

contents of high ctDNA samples (median=187) also were declining when comparing them to low ctDNA samples (median=239). On the other hand, several samples from both low and high ctDNA have telomere content more than the median of healthy donors. We found cfDNA from 3 brain tumors, 4 sarcomas, and 1 other cancer with high telomere content (telomere content > 400). This indicates that cfDNA could harbor the evidence of shortening or elongation of telomere of pediatric cancers.

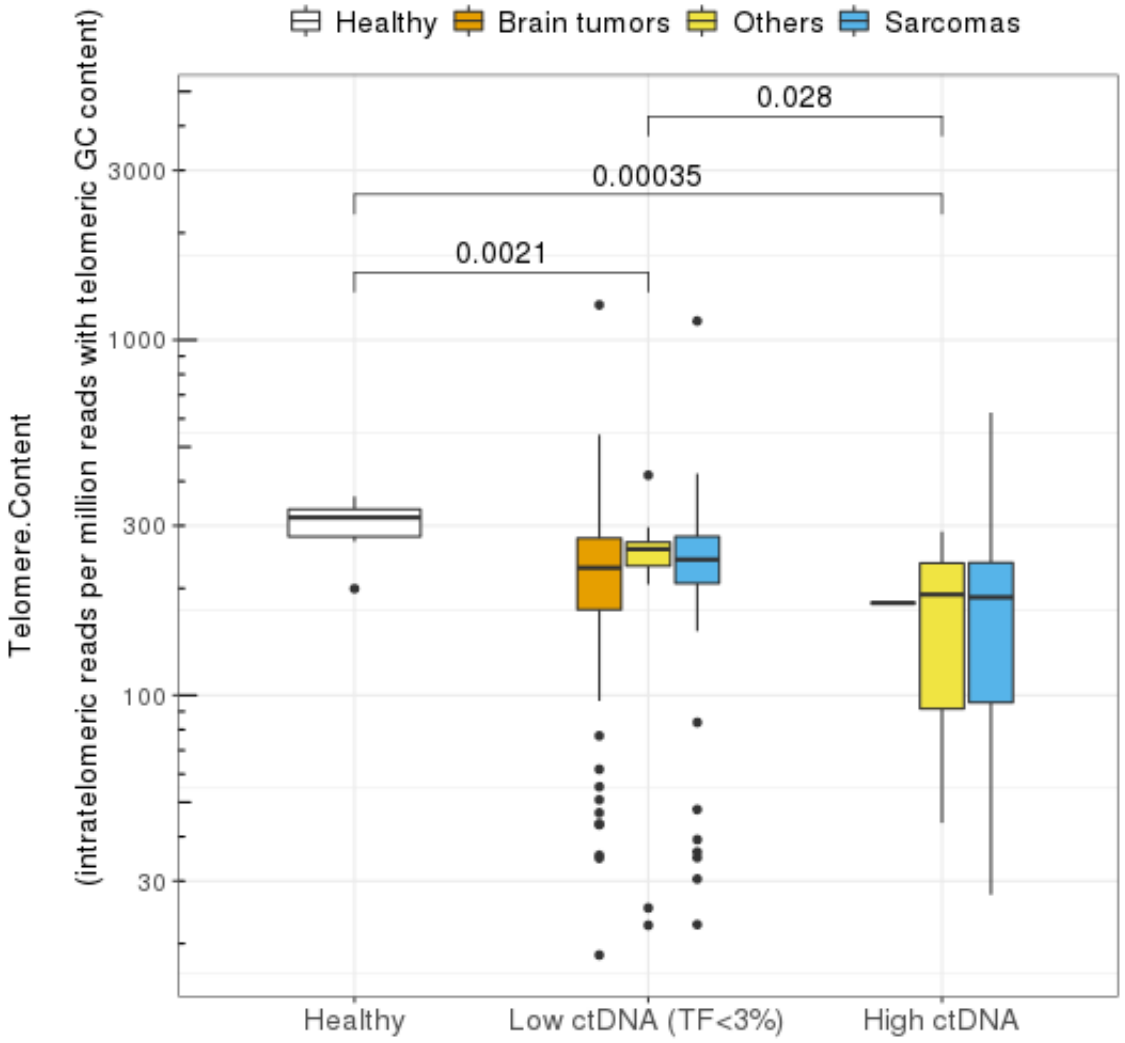


Figure 48: Telomere content of cfDNA in the pediatric cohort and additional healthy donors: Telomere content of high ctDNA samples had decreased significantly comparing to healthy donor (p-value=0.00035; Wilcoxon rank sum test) and low ctDNA samples (p-value=0.028; Wilcoxon rank sum test).

4.5.3 Integration of telomere variant repeats were detectable in plasma cfDNA

Without matched control given, TelomereHunter calculated the normalized count of TVR singletons in intratelomeric regions of cfDNA. The normalized count of TVR singletons also increased accordingly with the telomere content in a number of patient-derived cfDNA. Figure 49 plotted normalized count of five TVRs (TCAGGG, TGAGGG, TTGGGG,TTCGGG, and TTTGGG) against total telomere content per tumor entity. Among 10 healthy cfDNA, none of the TVRs were explicitly enriched along with the increasing telomere content. Meanwhile, all TVRs except TTCGGG were positively correlated with increasing telomere content in cfDNA of brain tumors and sarcomas. The frequently integrated TVRs in brain tumors and sarcomas cfDNA were: TCAGGG, TTGGGG, and TTTGGG. Meanwhile, none of the other cancer cfDNA showed any frequently integrated TVRs.

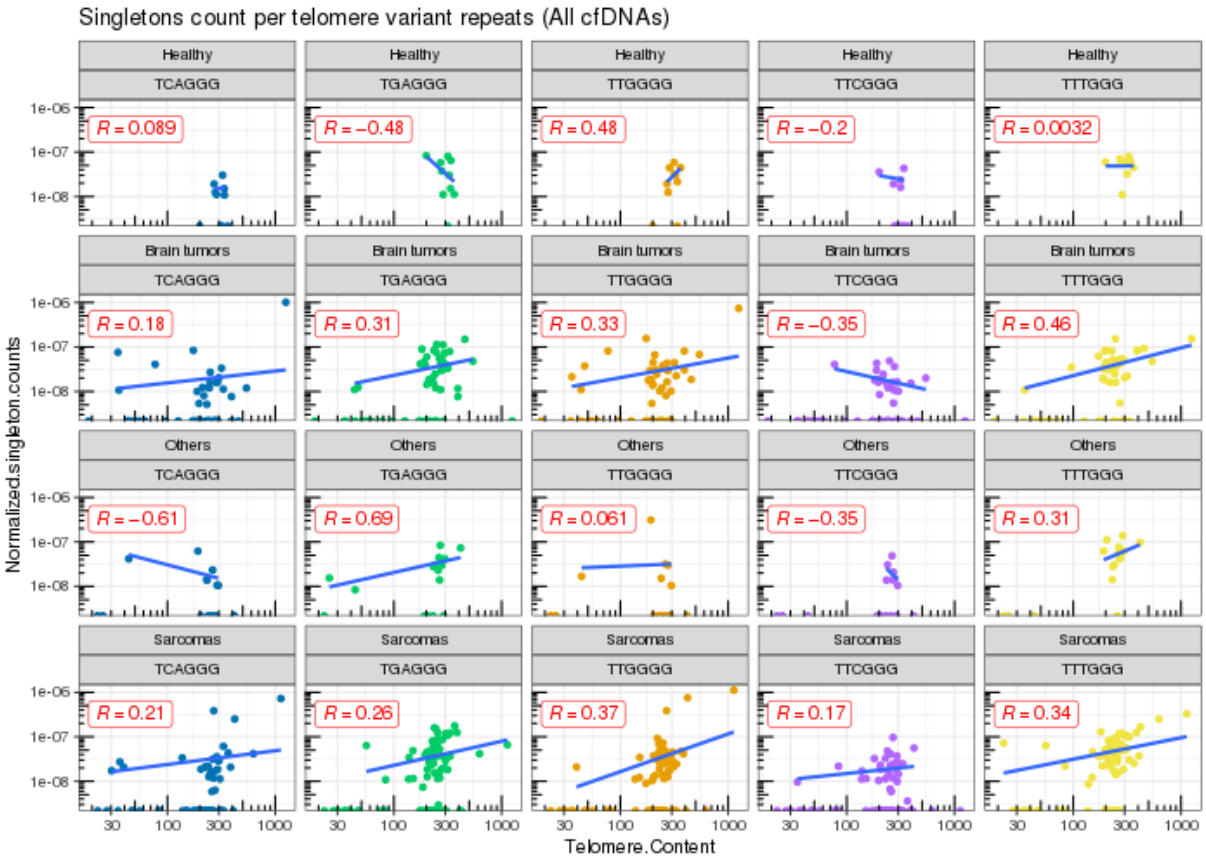


Figure 49: Normalized count of telomeric variant repeat in cfDNA : The normalized count of 5 TVR singleton (TCAGGG, TGAGGG, TTGGGG,TTCGGG, and TTTGGG) was plotted against telomere content. The enrichment of TVRs and telomeres was observed in brain tumor and sarcoma cfDNA samples. The correlation coefficient (in the red box) was calculated using Pearson correlation.

Among those cfDNA samples, 36 samples were derived from patients whose tumor harbors at least one ALT-associated point mutation (Figure 50). A cfDNA sample from HGG with an ATRX point mutation showed an increasing telomere content, but none of the TVR was increased, possibly due to the low tumor fraction. On the other hand, an increase of telomere content and TVR normalized counts was often found in the group of sarcoma patients. Most of them have a mutation in TP53 and commonly have a high estimated tumor fraction. The cfDNA from ATRX-mutated osteosarcoma did not show strong enrichment of TVR insertions nor telomere elongation, possibly due to low tumor fraction in the sample. It is interesting to find out which mutation could cause ALT and telomere elongation in pediatric sarcomas.

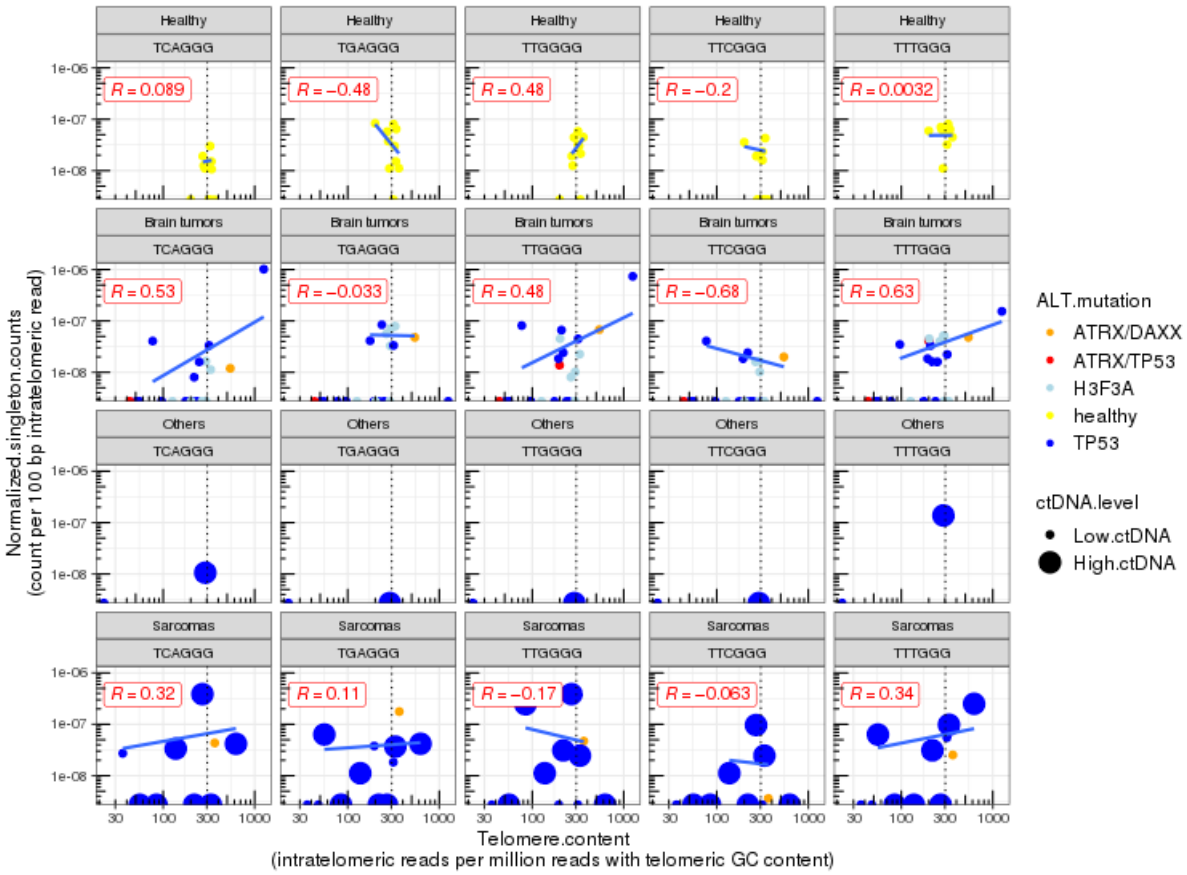


Figure 50: Enrichment of telomere variant repeats of cfDNA samples with ALT-associated point mutation. The ctDNA level was categorized regarding the tumor fraction (TF) estimation from ichorCNA where low.ctDNA are samples with TF $< 3\%$. The verticle dashed line denotes the median telomere content of healthy cfDNA.

4.6 CfDNA Analysis of Pediatric Cancer

4.6.1 Tumor entity influences the success of detection

The tumor entity seems to influence the success of CNV detection using lcWGS of plasma cfDNA. In this cohort, high-ctDNA samples were detected in sarcoma (28.6%), followed by other pediatric cancers (20%) and brain tumors (1.9%) (Figure 51A). With this rate, detecting tumor CNVs in sarcomas or other cancers is more likely than in brain tumors.

We determined how many tumor CNVs are detected in high-ctDNA (TF $\geq$ 3%) and low-ctDNA (TF $<$ 3%) samples (Figure 51B). A tumor CNV is considered as detected when at least 20% of the segment is overlapping with a cfDNA segment and both report the same CNV event (either amplification, neutral, or deletion). Among cfDNA from sarcoma patients, 79% (426/541) of tumor CNVs were detected from cfDNA with high-ctDNA whereas samples with low-ctDNA detected 28.5% of tumor CNVs. The only high-ctDNA sample from brain tumors was derived from a patient diagnosed with metastatic medulloblastoma. This sample shows a similar CNV profile to the matched tumor and allowed detection of 55% (10/18) of tumor CNVs. Considering high-ctDNA samples of other childhood cancers, the detection rate is the lowest (20%) although they were reported to have very high TF (37%, 15%, and 3.1%). Low-ctDNA samples showed a similar detecting rate at approximately 30% of tumor CNVs in patients with sarcomas and other tumors, and 18% in patients with a brain tumor. Together, if we consider the detection rate of low-ctDNA as a background signal, cfDNA samples with TF $>$ 3% can detect approximately half of tumor CNVs.

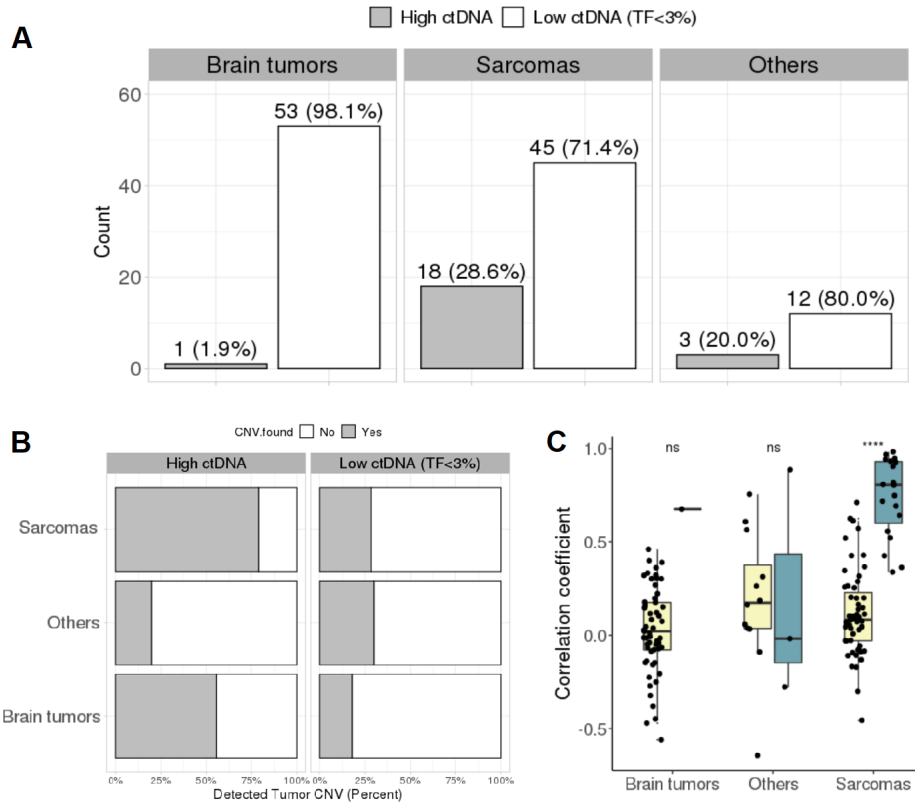


Figure 51: Estimated tumor-fraction in the pediatric cohort and correlation to tumor copy-number profile: (A) Number of high ctDNA and low ctDNA samples per tumor entity; (B) Number of CNVs detected per tumor entity; (C) Pearson correlation coefficient of genomic log2 ratio between tumor and cfDNA.

Not only considering the tumor CNVs, but we also calculated the correlation between the copy-number

log2 ratios of cfDNA and the matched tumor (Figure 51C). The correlation of cfDNA with low-ctDNA was quite varied because the correlation also consider the tumor with few or flat copy-number profiles. The correlation rises when comparing flat cfDNA with a flat tumor genome. In particular, we found a significantly high correlation among sarcoma patients when the cfDNA is high-ctDNA. This supports that the cfDNA shows a similar profile to the tumor genome. Among other pediatric cancers, the correlation of high-ctDNA samples indicates that two cfDNA shows a different profile from the respective tumor genome. One of the cfDNA is derived from a Wilms tumor patient and has low similarity to the matched tumor although the TF is high (37%). We wondered that this cfDNA contains tumor cfDNA secreted from other tumor cells located at other sites in the body. The result of the investigation is shown in the next section (Section 4.6.3).

4.6.2 Short-fragmented cfDNA are enriched in high-ctDNA samples

Using cfdnakit, fragment-length profiles have been generated from 13 high-ctDNA (TF $\geq$ 3%), and 81 low-ctDNA (TF $<$ 3%) samples from 15 different pediatric cancer types. We also analyzed additional 10 plasma samples from healthy controls. The ratio of short-fragmented cfDNA (size between 100 to 150 base pairs) over long-fragmented cfDNA (size between 151 to 250 base pairs) is calculated per cfDNA sample. The ratio is significantly higher in high-ctDNA samples than in those samples from healthy controls or with low-ctDNA (Figure 52A). The ratio of healthy controls ranges from 0.14 to 0.24 (median = 0.18). The ratio of cfDNA from cancer patient varies between 0.11 and 0.88 (median = 0.22) in low-ctDNA, and between 0.18 and 1.10 (median=0.43) in high-ctDNA samples. It is also possible that other tumor genetic alterations could contribute to short-fragmented cfDNA rather than copy-number aberrations. It thus seems that the enrichment of short-fragmented cfDNA is commonly associated with enrichment of ctDNA in the blood plasma of cancer patients.

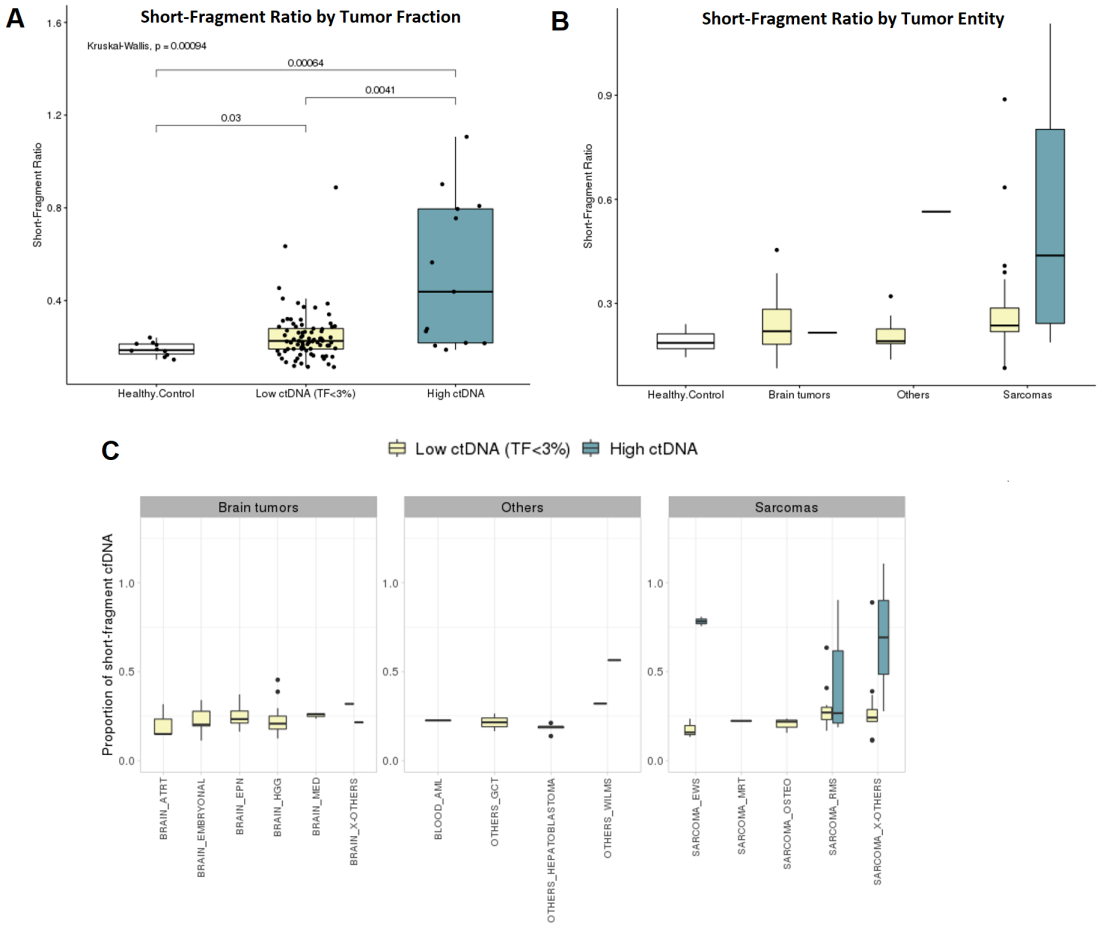


Figure 52: Short-fragment ratio of cfDNA in the pediatric cohort and association with estimated tumor fraction: (A) Distribution of short-fragment ratio shows that high ctDNA sample contains significantly more short-fragmented cfDNA than healthy donors and low ctDNA. (B) Enrichment of short-fragment cfDNA is frequently found in sarcoma. Enrichment in brain tumors and other childhood cancer is rare. (C) Distribution of short-fragment ratio per diagnostic group shows that enrichment is found in Ewing's sarcomas, rhabdomyosarcomas (RMS), and other sarcomas.

Regarding tumor entities, the majority of cfDNA with enrichment of short-fragments were from sarcoma patients where we found 11 high-ctDNA from 44 total sarcoma cfDNA (Figure 52B). The ratio of sarcoma high-ctDNA ranges from 0.18 to 1.1 (median = 0.43). We found a high short fragment ratio (0.56) in the sample from the bilateral Wilms tumor (Section 4.6.3). Nevertheless, the brain tumor with high-ctDNA has a short-fragment ratio of 0.21, which is much lower than in other high-ctDNA samples. We found enrichment of short-fragment cfDNA among a group of rhabdomyosarcomas, Ewing's sarcomas, and other sarcomas (Figure 52C).

4.6.3 Tumor spatial and temporal heterogeneity in plasma-derived cell-free DNA

We explored the potential benefits of cfDNA as a minimal-invasive liquid biopsy in a cancer management setting. A liquid biopsy should be able to inform the emergence of refractory tumors or the existence of clones locating at multiple sites. In this cohort, we have inspected spatial and temporal heterogeneity of the tumors as detected in the corresponding cfDNA.

A 5-year-old girl was diagnosed with a Wilms tumor, the most common type of kidney cancer in children, at both of her kidneys. A tumor biopsy from one of her kidneys was obtained. The genomic analysis of the tumor biopsy revealed amplification of MYC, a somatic mutation in TP53, and overexpression of KDM1A. The genome-wide copy-number profile shows amplifications at chromosome 4p,

8q23.1, and 18q22.1 and deletions on chromosome 4q, 8q, 17, 18, 21, and 22 (Figure 53A). The tumor later progressed and spread to multiple locations including liver, lymph nodes, and abdominal wall. Multiple sampling of tumor tissues to get comprehensive genetic information might be difficult. A liquid biopsy has been obtained from the peripheral blood of the patient. Plasma cell-free DNA was extracted and submitted to multiple sequencing libraries includes lcWGS, WES, and Panel-seq.

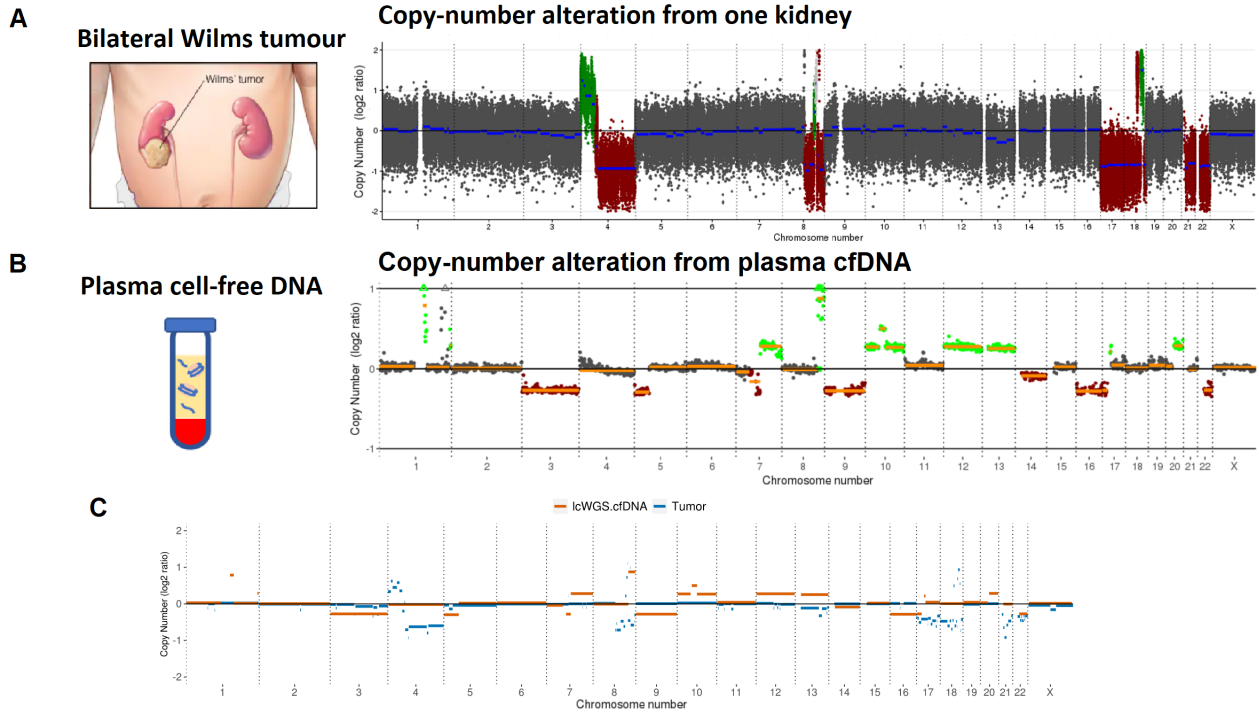


Figure 53: Tumor spatial heterogeneity was captured by cfDNA from a patient with metastasis bilateral Wilms tumor. (A) Genome-wide copy-number aberrations from a tumor tissue obtained from one kidney; The tumor image is reprinted with permission from MayoClinic.org (Copyright © 1998-2021 Mayo Foundation for Medical Education and Research (MFMER). All rights reserved.). (B) Genome-wide copy-number aberrations from plasma cfDNA; (C) Overlapping genomic segments from the tumor tissue (blue) and the plasma cfDNA (orange); Colors in a genomic profile represent CNV events (grey:neutral, red:deletion, green:gain (3N), light green:amplification ($>$ 3N)).

From the result of the lcWGS method, ichorCNA reported 37.4% estimated tumor fraction and detected copy-number aberrations in multiple loci (Figure 53B). Interestingly, the genome-wide copy-number profile of the cfDNA looks different from the tumor profile (Figure 53C). Genotyping has confirmed that those samples were derived from the same individual (Supplement FigureS5). Low-coverage whole-genome sequencing of cfDNA reveals only similarity in the deletion of chromosome 22. Aberrations in chromosome 4 still existed at the very low fraction. This shows that the majority of ctDNA was not released by the tumor population that we have obtained.

We annotated the aberrant regions with an in-house list of druggable genes. We found 105 genes that were exclusive to cfDNA, 64 exclusive to the tumor, and 20 common druggable genes. Supplement Table S3 shows druggable genes found exclusively in cfDNA. Among those cfDNA-exclusive alterations, a deletion of CTNNB1, gene encoding beta-catenin, was found. This gene is commonly mutated in Wilms tumors[155] and many types of cancer[156–159]. It is known as a major component of the Wnt signaling pathway and forming E-cadherin cell-cell adhesion systems[157]. The loss of E-cadherin adhesion in association with the epithelial–mesenchymal transition (EMT) occurs frequently during tumor metastasis[160]. However, this alteration might not be druggable since CTNNB1-targeted drugs, such

as TTK inhibitor, aim to suppress the activation of CTNNB1 that drives cell proliferation through the Wnt signaling pathway [161]. We are now looking forward to finding potential candidates for the next drug target that could cure the majority of tumors based on the evidence from the liquid biopsy.

We received plasma cfDNA samples in a time-series manner from 10 patients including 9 sarcomas and 1 pediatric hepatoblastoma. Although there are 2 samples per patient, we have identified 4 patients whose cfDNA contains high-ctDNA in at least 1 time-point (Figure 54A). We detected 3 high-ctDNA samples obtained at the first time point from patients with sarcoma. Their copy-number profile looks similar to their matched tumor CNV profiles. The estimated tumor fraction (TF) were 28.6%, 10.1% and 3.1%. Since we do not have the clinical record at the sampling time, it is possible that those liquid biopsies were obtained at the diagnosis time or before the surgery and contained a detectable amount of ctDNA. The second biopsy from a patient with an inflammatory myofibroblastic tumor (IMT) also contained a high level of ctDNA (TF = 5.2%) and also maintained the same CNV profile from as in first biopsy (TF = 3.1%).

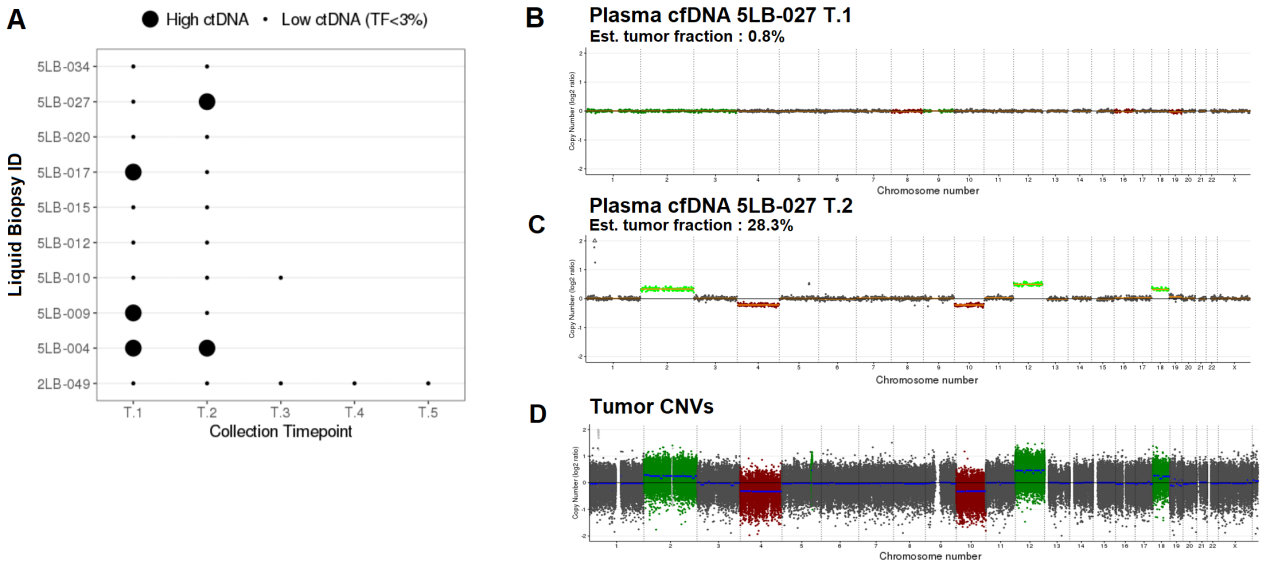


Figure 54: Time-series cfDNA biopsy captured refractory tumor in pediatric patients. (A) Tumor fraction estimation was performed in a time-series cfDNA collection of 5 patients. Samples with high TF (TF $\geq$ 3%) were highlighted with bigger dots. (B and C) Genome-wide copy-number aberrations plots of cfDNA collected at timepoint T.1 and T.2; (D) Genome-wide copy-number aberrations of tumor WES of the same patient; The copy-number profile of T.2 looks similar to the profile of tumor while T.1 does not shows any apparent CNVs. Colors in a genomic profile represent CNV events (grey:neutral, red:deletion, green:gain (3N), light green:amplification ($>$ 3N)).

We also detect a cfDNA sample that captures the refractory of the tumor at the second time point. The sample was obtained from a patient with rhabdomyosarcoma. The first cfDNA has not shown any copy-number aberration and ichorCNA reported 0.8% TF (Figure 54B). However, the second liquid biopsy was reported having TF 28.3% and contains multiple large CNVs including amplification of chromosome 2, 12 and 18, and deletion in chromosome 4 and 10 (Figure 54C). These CNVs were also found in the matched tumor biopsy (Figure 54D).

4.6.4 Estimated tumor fraction guides detection of targetable mutation in noncranial tumor

As described in Section 4.4.4, a positive correlation was found between the estimated tumor fraction, the mutational burden, the number of tumor mutations, and the percentage of detected tumor mutation. We hypothesize that when a high tumor fraction (TF $>$ 3%) is reported from lcWGS, it could suggest the utilization of WES that could provide more sensitive detection of point mutations to screen for clinically relevant or druggable mutations. We extracted copy-number aberrations and point mutations from 54 individual-matched tumor WES and plasma WES. The number and the proportion of tumor alterations detected in cfDNA were counted and calculated. We also track the number and the percentage of detected alteration of druggable genes.

As a result, cfDNA can detect the majority of tumor CNVs and point mutations with WES strategy when more than 3% tumor fraction was reported from the lcWGS (Figure 55). The cfDNA from 4 sarcomas and 2 other pediatric cancers were reported with high TF whereas none of the brain tumors reach 3% of the tumor fraction threshold. Being reported as high TF, cfDNA detected more than 70% of druggable CNVs and 80% of tumor mutations found in the tumor through the WES strategy. Only the cfDNA obtained from bilateral Wilms tumor detected only tumor mutations (35%) and druggable CNVs (21%) because of spatial heterogeneity (Section 4.6.3).

The detection rate of samples with low tumor fraction (TF $<$ 3%) was decreasing in detecting tumor mutations and druggable CNVs. The detection rate decreased to below 12 % on average among low TF samples. However, the detection of tumor mutations and druggable mutations were increased to above 20% in 12 samples. It shows that lcWGS ignores the existence of point mutation when very few or none of the copy-number aberrations exist.

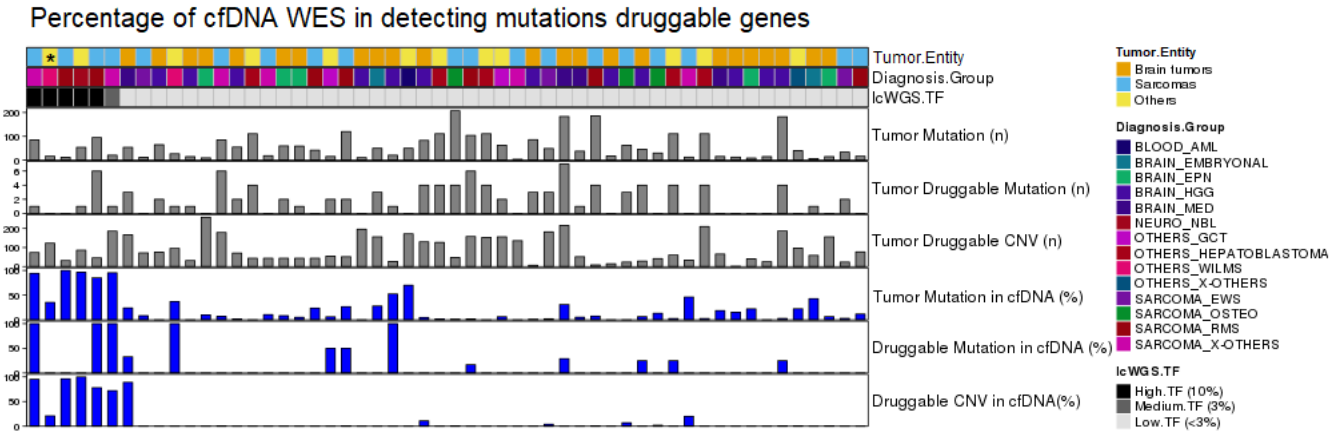


Figure 55: Number and percentage of targetable aberrations by the level of estimated tumor fraction. The first 3 rows of grey bar plots show the number of mutations, the number of druggable mutated genes, and the number of druggable genes with CNVs in a tumor. The 3 rows of blue bar plots below show the percentage of detected tumor mutations, percentage of detected tumor druggable genes, and percentage of detected druggable genes with CNVs in cfDNA using the WES strategy. The cfDNA sample obtained from the patient with bilateral Wilms tumor is highlighted (*).

The chance of detecting tumor mutations is higher when more than 3% TF is reported by lcWGS. Among cfDNA samples from noncranial tumors (sarcomas and other pediatric cancers), we detected 5 very high tumor fractions (TF $>$ 10%) cfDNA samples from 3 sarcomas including 2 embryonal rhabdomyosarcomas, 1 alveolar rhabdomyosarcoma, and 2 other pediatric cancers including a neuroblastoma

and a bilateral Wilms tumor (Figure 56). A substantial tumor fraction (TF > 3%) is derived from an inflammatory myofibroblastic tumor (IMT). WES successfully detects at least one druggable mutation in 4 out of 6 patients that have estimated TF > 3%. We detected a novel mutation in the targetable NOTCH2 gene, an oncogene that is overexpressed in a range of cancers [162], from the blood of a patient with bilateral Wilms tumor. This mutation could be secreted from a tumor that locates apart from the primary site.

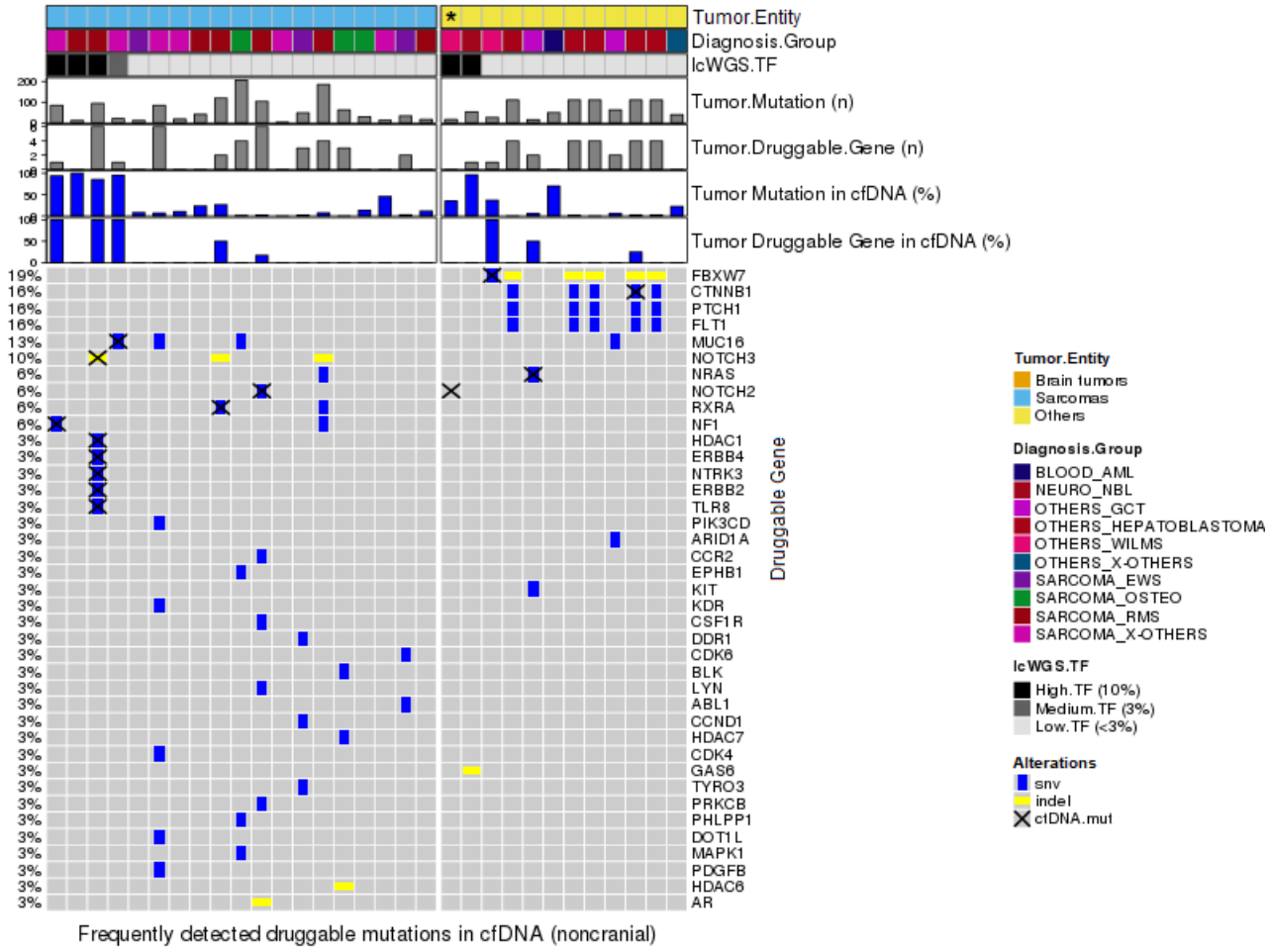


Figure 56: Mutation detection rate and detected druggable genes in noncranial tumors: the first 2 rows of the bar plots show the number of mutations and number of druggable mutated genes in tumor tissue. The following 2 rows of blue bar plots show the percentage of detected tumor mutation and druggable mutated genes in cfDNA. The matrix reports deleterious somatic mutation in druggable genes found in tumors (SNV in blue and INDEL in yellow) and in cfDNA (black cross).

Interestingly, we detected all 5 druggable mutations from a cfDNA sample (2LB-055.P01) of a patient with embryonal rhabdomyosarcoma (RME). Those mutated genes included NOTCH3, HDAC1, ERBB2, ERBB4, NTRK3 and TLR8. Recently known as HER2 and HER4, ERBB2 and ERBB4 encode receptor tyrosine kinases of the human epidermal growth factor receptor (EGFR) protein family [163]. These HER proteins are classified as oncogenes, causing tumorigenesis, tumor growth and progression through overexpression, mutation, truncation and gene amplification [164] The mutation of HER2 was found at low frequencies in many cancer types including RME and could be the target of HER2 targeted drugs [165–167]. NTRK3, one of the neurotrophic tyrosine kinase (NTRK) genes, promotes cell proliferation, differentiation, and survival via activation of several signaling pathways including JAK/STAT, PI3K/AKT, and SHC/RAS/MAPK [168]. Although FDA has approved the use of tropomyosin-related

kinase (TRK) inhibitors for patients with NTRK gene-fusion, mutations in the kinase domain have shown high sensitivity to type II inhibitors, including altiratinib, cabozantinib and foretinib [169]. NOTCH3, encodes one of NOTCH signaling transmembrane protein, and HDAC1, encodes one of histone deacetylase protein complex, regulate the activation of NOTCH targeted genes [170]. Those genes include HES and HEY protein families, CD25, cyclin D1 and c-MYC which implicate in cell-cycle progression, and cancer stemness [171, 172]. Many strategies have been introduced to target NOTCH3-mutated cancer including siRNA/shRNA and antibody-drug conjugates [172]. TLR8, a member of toll-like receptor encoding genes, were found highly expressed in cancers leading to tumor cell proliferation and chemoresistance [173].It is also being recognized as a potential target of cancer immunotherapy that could reverse the immune suppressive function leading to strong tumor inhibition [174].

Among cfDNA samples from brain tumor patients, no sample reaches 3% TF (Figure 57). The detection rate of tumor mutations is approximately 12%, and only 8 samples contain at least one mutation in druggable genes. We found a sample from high-grade glioma to contain 15 novel mutations and 1 tumor mutation locating in druggable genes. We have checked base quality, genotypic fingerprint, and oxidative sequencing artifact; no sample problem is detected. The overall mutation burden of this cfDNA sample is 628 nucleotide variants which are even higher than the primary tumor. It is not clear what could be a possible source of these druggable mutations.

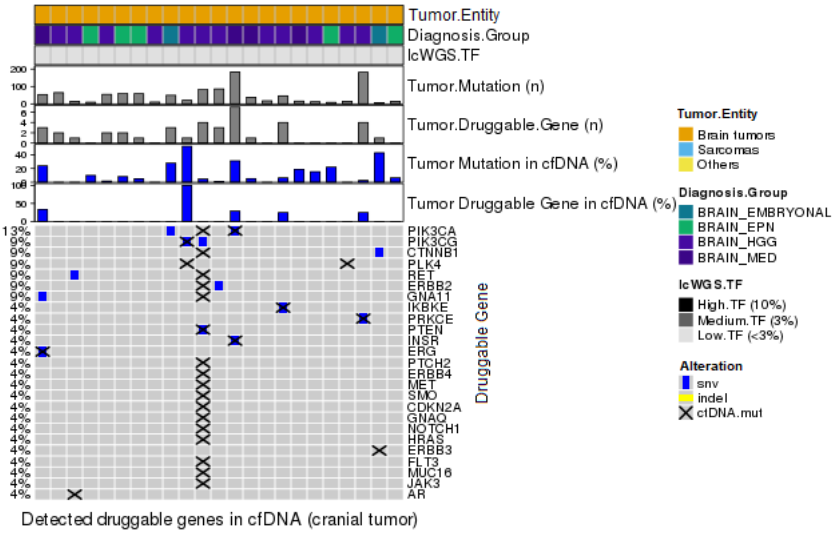


Figure 57: Druggable mutation in cfDNA from brain tumor patients

4.7 Summary

This thesis demonstrated a comprehensive utility of cfDNA in pediatric cancer based on information derived from three next-generation sequencing techniques. The proportions of cancer types in this cohort also represents the incidence of childhood solid tumors where brain tumors and sarcomas represented the majority of data in the dataset, followed by other childhood-specific entities. Based on this cohort, we found that the success of using plasma cfDNA in detecting tumor DNA from blood circulation (circulating tumor cell-free DNA or ctDNA) is affected by two factors. First, the different sequencing methods of choice have their limitation of sensitivity and comprehensiveness. The broader the genome-scale level of the method, the lower the sensitivity towards single-base alterations of the method is. The higher coverage could provide more sensitivity at the single-base resolution but will lose the ability to detect comprehensive alterations across the genome. Second, the tumor location influences the probability of

tumor-derived cfDNA being shed into blood circulation. We observed the rate at which mutations from brain tumors were detected in plasma cfDNA to be very small, while more alterations from non-cranial tumors can be detected.

It has been widely observed in many adult cancer studies that ctDNA is shorter than cfDNA shed by non-malignant cells. We explored the size of cfDNA in the patient-derived xenograft mouse experiment and the pediatric cohort. The enrichment of short-fragmented cfDNA was observed in human-derived cfDNA in the xenograft, supporting the result of a previous study in an ovarian cancer experiment. The enrichment of short-fragmented cfDNA and number of copy-number alterations were positively correlated. In-silico size-selection can enhance the copy-number alteration and increase the tumor fraction estimates. In the pediatric cohort, the enrichment is also more prevalent among non-cranial tumor cfDNA than in brain tumor patients. We found that the amount of short-fragmented cfDNA not only correlates with the copy-number alteration status but also with the overall mutation burden and with the amount of ctDNA.

The utility of plasma cfDNA in detecting telomeric aberrations including telomeric elongation and integration of telomeric variant repeats has been demonstrated in this study with lcWGS assay. The estimated telomere content of cfDNA derived from patients was decreased comparing to cfDNA of adult healthy donors. The integration of telomere variant repeats in the intratelomeric regions could be detected in plasma cfDNA. A positive correlation was found between the telomere content and the frequency of TVRs integration in brain tumors and sarcomas. However, the low concentration of tumor-derived cfDNA from brain tumors provided variability in the result. The association between ALT-associated mutations with particular TVR integration could not be found. Different quantification strategies and additional enrichment methods could provide higher sensitivity and specificity to the experiment in the future.

The thesis demonstrated that plasma cfDNA can reveal spatial and temporal tumor heterogeneity which commonly complicate the success of therapy. These findings have shown the potential benefit of liquid biopsy to pediatric cancer patient management.

5 DISCUSSION

5.1 Efficacy of Low-coverage Whole-genome Sequencing (lcWGS) in Detecting Tumor-derived CfDNA

5.1.1 Sequencing cfDNA with lcWGS shows a comprehensive copy-number profile and allows estimation of tumor fraction.

When applying cfDNA as a liquid biopsy material, the type of genetic alteration to be used as a marker would play an important role in the success of the detection. This study demonstrates the detection of large copy-number alterations and estimates tumor fraction (TF) from lcWGS (median coverage 1.32x) using the ichorCNA bioinformatics tool. We measured the sensitivity and specificity of lcWGS in detecting copy-number variants (CNV) at different tumor fractions found in the cohort. The sensitivity rises to approximately 75% when the TF reaches 3% and becomes stable at 80% when the TF reaches 5%. On the other hand, the specificity when the TF reaches 3% was stable around 68% and reaches approximately 80% at 9% TF. The evaluation result is similar to the result provided by the developer of ichorCNA who performed several comprehensive benchmarkings by an in-silico mixture approach at 1x genome coverage [102]. They also found the lower limit of 0.03 TF for detecting the presence of chormosome-arm aberration (>100 Mb). This indicates that the genome-wide copy-number profile of cfDNA should look very similar to the tumor when the tumor with large CNVs sheds enough DNA into the blood circulation. However, it is important to mention that this study considered CNVs in the tumor as a ground truth. We should not overlook the fact that cfDNA might contain CNVs originating from a tumor population that has not been captured by sequencing the primary tumor. The benchmarking by in-silico mixture approach using tumor DNA and health donor cfDNA was not performed in this thesis. Since CNVs are the alterations commonly found in pediatric cancers, comprehensive screening for CNVs from the liquid biopsy with lcWGS could further indicate the use of a more targeted and sensitive sequencing approach (e.g. whole-exome sequencing, gene-panel sequencing, or PCR).

Regarding the detection of point mutations, this study has shown that the detection rate from lcWGS is less than 15% of total tumor point mutations in the cohort even though the TF was higher than 3%. This implies that lcWGS cannot provide enough coverage to detect the tumor-derived cfDNA when the tumor is driven by point mutations. This is the major limitation of implementing lcWGS to detect tumors at the early stage especially in pediatric cancer that ≈10% of them harbor few mutations in cancer predisposition genes [175]. This problem can be solved by the whole-exome sequencing (WES), where point mutations can be detected at 5% lower limit of detection [96]. Combining both advantages of lcWGS and WES could provide comprehensive information regarding both CNVs and point mutation and increase the success of detection for all pediatric cancers. Further development of DNA extraction, isolation, and preparation are required to obtain enough cfDNA material for generating lcWGS and WES libraries from a limited DNA of blood collection from a child.

Since lcWGS could only detect CNVs when a sample reaches approximately 3% of TF, detecting tumor CNV at low TF might be difficult. In-vitro or in-silico size-selection could enriched tumor-derived cfDNA [110] and has been demonstrated in Section 4.4.2. The success rate of this strategy also depends on the initial concentration of tumor-derived cfDNA. The in-vitro could better enrich the detection of CNVs than the in-silico approach [110]. However, our samples have been already sequenced when this study begin and sample re-processing is not possible. The other sequencing strategy to detect early detection of a relapsed tumor is gene-panel sequencing or personalized panel-sequencing. They provide both sensitivity and specificity in detecting point mutations with the lower limit of detection at variant allele frequency at 0.1% [96, 176]. This thesis also performed the analysis of panel-sequencing (Section 4.3.4) and discussed in Section 5.2.2.

5.1.2 The location of the primary tumor influence the success of detection by plasma cfDNA sequencing.

Selecting the source of liquid biopsy is the most important decision that could already determine the success of capturing the tumor marker. Obtaining an unsuitable source of liquid biopsy often leads to high detection failure. In this study, we compare the success of detecting tumor CNVs from plasma cfDNA with lcWGS between cranial (brain tumors) and non-cranial (sarcomas and other childhood-specific tumors) (Section 4.6.1). Using 3% TF as the threshold of success detection, detecting CNV from a cranial tumor is very rare. Only 1 out of 54 cfDNA samples from brain tumor patients could reach the threshold. In total, plasma cfDNA detected only 9% of brain tumor CNVs and would rather be false-positive results because the low specificity was commonly found among samples with TF $<$ 3%. The more successful detection was observed among cfDNA from patients with non-cranial tumors. Approximately 25% of samples contained more than 3% TF and had 90% of tumor CNVs detected.

The success rate of using plasma cfDNA to capture ctDNA based on CNVs is influenced by the concentration of tumor-derived cfDNA in the liquid biopsy sample. The rare success rate among brain tumor patients could be explained by the location of the primary tumor where the blood-brain barrier blocks the release of tumor DNA into the blood circulation. The ideal source of liquid biopsy for detecting ctDNA from brain tumor patients is cerebrospinal fluid (CSF), which provides necessary nutrients and removes waste in the central nervous system. It has been demonstrated that CSF could lead to detection of genetic aberrations in patients with leptomeningeal metastases of non-small-cell lung cancer [82]. The extended dataset of this cohort, not included in this study, contains 33 CSF samples from pediatric brain tumor patients. Almost half of them are estimated to have more than 3% TF (Supplement Figure S6). Supplement Figure S7 shows an exemplary result of CSF in detecting CNVs from a patient with medulloblastoma. The further evaluation analysis of these CSF samples is beyond the scope of this thesis. On the other hand, the success rate of detection in this cohort shows that blood plasma is the possible source of liquid biopsy for patients with non-cranial tumors. Although we detected only 25% of samples having a high level of ctDNA (TF $>$ 3%), the related clinical status has been blind to us at most of the time in this study. Additional clinical information such as the stage or size of the tumor, time point of treatment when the sample was taken, or RECIST status could help us to understand the relationship between the progression of a tumor and the detection rate of plasma cfDNA. Overall, selecting a suitable source of liquid biopsy influences enormously the success of detection using any tumor marker. Keeping the correct sample could still provide a chance of trying different detection strategies while an incorrect source of sample will not be a suitable starting meterial.

5.2 Efficacy of Whole-exome Sequencing (WES) and Panel-seq in Detecting Alterations at Higher-resolution

5.2.1 Deep and broad coverage of WES allows interogation of point mutations.

Compared to adult cancers, childhood cancers typically have fewer somatic mutations but a higher prevalence of germline mutations in cancer predisposition genes [15]. Approximately 50% of pediatric cancers harbored at least 1 potentially druggable alteration, and one-third of them retain the potentially druggable alteration at the time of relapse. Obtaining a tumor biopsy from a patient allows us to extract its molecular profile which could guide the therapeutic selection. Obtaining multiple or serial biopsies could track the mutational dynamics during and after the course of treatment. However, it poses several challenges including patient's discomfort and overlooking of tumor clone at an adjacent or remote site. Although lcWGS of plasma cfDNA allow us to detect large CNVs and estimate the tumor fraction, it lacks the power to detect mutations at single-base resolution. Nevertheless, WES and gene-panel sequencing (Panel-seq) offer sequencing depth power to detect point mutations in targeted regions with

cfDNA material.

WES provides both breadth and depth sequencing to detect functional somatic point mutations with a lower limit of detection of 5% [96] . In this study, we detected at least one tumor point mutation in 90%, 85%, and 75% in cfDNA of sarcomas, other childhood cancers, and brain tumors, respectively. We extracted somatic functional point mutations in 367 genes that could be candidates for targeted therapy in pediatric cancer patients from tumor and cfDNA WES (Supplement Table S2). We found that 30% of cfDNA in this cohort contained at least 1 druggable mutation. Mutations detected from non-cranial tumor cfDNA show a higher concordance to the matching tumor than cfDNAs from a brain tumor. This implies that tumor type also affects the detection rate of cfDNA WES similarly to lcWGS.

The source of mutations that exclusively exist in cfDNA is unclear. Most of these samples gain 1 extra druggable mutation. It could originate from adjacent tumor populations that have not been captured by tumor biopsy. Alternatively, the mutation could arise from the subclone of the primary tumor during the course of treatment. The tumor spatial heterogeneity could explain the source of multiple exclusive point mutations ($>$ 4 mutations) in cfDNA. It could be the tumor population that seeds at a distance site away from the primary tumor. The local environment applies a different selective pressure that drives the continuous development of distinct clones and shed DNA into blood circulation. A study in a group of non-small cell lung cancer, known as TRACERx, used multi-region exome sequencing to construct phylogenetic tumor branches[177]. Multiplex-PCR assay panels were designed per patient targeting clonal and subclonal SNVs to track the phylogenetic tumor branches in plasma cfDNA [176]. They found that a median of 27% of subclonal SNVs were detected in 68% of ctDNA-positive patients. Many of these subclone SNVs existed only in a particular region. In our study, we have identified a patient with bilateral Wilm tumor whose primary tumor genome does not have druggable point mutation. However, we detected a point mutation in NOTCH2 with additional druggable CNVs in the plasma cfDNA of the patient. This case has been confirmed afterward having multiple metastasis sites including the liver, lymph nodes, and abdominal wall. We assumed that the source of cfDNA-exclusive mutation was derived from those sites (Discussed in section 5.6.1).

The main limitation of WES in cfDNA is the limited sensitivity to detect segmental loss of heterozygosity (LOH) from B-allele frequency (BAF). Most bioinformatics tools would find it challenging in the sample with a low concentration of tumor cfDNA [178]. For example, our CNV calling workflow for tumor WES (Method Section 2.3) classifies a segment with the global maximum between 0.45–0.55 ($\sim$ 5% alternative allele frequency) as a balanced segment (no LOH). The WES can accuratly identify regions with LOH when compared with gold standard whole-genome SNP6 microarray in tumors of 40-60% purity [179]. This study also found that PureCN also provided ambiguous segmentation of BAF when the sample with low estimated tumor fraction. Excluding higher tumor ploidy (ploidy 4 and more) from solution searching parameters of CNV calling software could eliminate the ambiguous result of absolute copy-number in low TF samples. The future evaluation of cfDNA with WES could also include detection of LOH at different tumor fraction.

5.2.2 Customised Panel-seq provides a detection with more sentivity but limited breadth.

Panel-seq provides a more sensitive detection but at a limited number of genomic loci. The customized gene-panel has designed to capture 130 genes which are recurrently altered in brain tumors, focusing on coding regions and selected intronic and promoter regions [115]. Based on tumor WES, the gene-panel could capture at least 1 somatic deleterious point mutation in around half of the tumor DNA samples. We interrogated these mutations from the matched cfDNA samples and calculated the tumor variant allele frequency (VAF) of the detected variants. We found that only one-fourth of plasma cfDNA from brain tumor patients can detect at least one point mutation at the VAF ranging from 0.04% to 1%. This range of tumor fractions is below the limit of detection of both lcWGS and WES processed by standard

pipelines [96]. Almost half of cfDNA from sarcomas and other pediatric tumors can detect at least one tumor point mutation. The maximum range of VAF reaches 63% in sarcomas and 30% in other pediatric cancers. This finding shows that the location of the tumor influences the detectability of plasma cfDNA although the customised Panel-seq already be able to detect mutation at very high sensitivity.

The design of the gene-panel limits the tracking of druggable mutations to regions of only 66 genes. Using Panel-seq, 71% (5/7), 40% (4/10), and 22% (4/18) of cfDNA can detect at least 1 druggable point mutation in sarcomas, other pediatric cancers, and brain tumor cases, respectively. Panel-seq narrows down the scope of mutation detection and decreases the detection rate of druggable genes comparing to WES. The rate of detecting at least one targetable mutation is comparable between WES and Panel-seq (Figure 34). WES would increase the higher chance of detecting more functional druggable mutations per sample. Comprehensive alteration detection using WES could be applied to disease monitoring of advanced-stage patients and suggest the next therapeutic option [151]. Because WES supports the characterization of the genomic profile of cfDNA, it expands the possibility to detect alterations that might exclusively be shed by tumor clones that locate at a distant site.

The implementation of WES or Panel-seq of liquid biopsy cfDNA in the clinical management of pediatric cancer should consider the clinical objective of the application. Early diagnosis would require a sensitive assay to notify the developing tumor in the body. Most childhood cancers have been found driven by only a single cancer-driving mutation rather than multiple hits on cancer-driving genes [15]. The customized gene-panel could be designed to capture the most frequently mutated genes among childhood cancers. Meanwhile, a personalized gene-panel could be beneficial for disease monitoring and detecting minimal residual disease in terms of sensitivity and specificity. In the TRACERx study [177], personalized multiplex-PCR panel sequencing was used in ctDNA profiling of non-small cell lung cancers (NSCLC)[176]. This study shows that the multiplex-PCR assay provide a sensitivity above 99% for the detection of tumor allele frequencies above 0.1% and 99.6% specificity of detecting a SNV. It also was estimated that a plasma VAF of 0.1% would correspond to a primary NSCLC burden of 302 million tumor cells or tumor volumn of 10 cm^3. With this power of detection, personalized panel sequencing had detected at least two SNVs in 93% of patients with tumor relapsed before or at clinical replase (median lead time = 70 days). This shows that the sensitivity of Panel-seq could provide clinical benefits for early detection of relapse tumor. Overall, the limit of detection and coverage of these two next-generation sequencing approaches (WES or Panel-seq) should be major points of concern.

5.3 Estimation of Tumor Fraction Guides the Use of Subsequent Sensitive Detection Method.

The detection of actionable somatic mutations in the plasma of pediatric cancer patients has made possible the minimal-invasive biopsy to guide therapy selection. WES and standard WGS can provide genomic profiles from the plasma cfDNA. However, the high cost of sequencing and the low tumor fraction limit the cost-effectiveness of those sequencing approaches. Gene-panel sequencing could be very sensitive but limited to only detecting mutations in clinically actionable regions by the assay design. Moreover, gene-panel sequencing cannot be used for the characterization of genomic features such as mutational signatures or mutational burden which could be used as a biomarker of checkpoint blockade immunotherapy[151]. LcWGS uses less DNA material but can provide a comprehensive genomic profile of plasma cfDNA. The correlation between tumor fraction estimated from lcWGS and the success in detecting actionable point mutation with WES has been previously demonstrated in metastatic adult solid tumors [102, 151]. However, pediatric cancers are known to have less tumor mutational burden than adult cancers [15]. The possibility to detect pediatric druggable mutation using lcWGS tumor fraction estimate as a guide has never been demonstrated.

In this study, we compared the detection rate of tumor point mutations and the detection rate of

druggable mutations to the TF reported by lcWGS. As expected, we found a positive correlation between TF of lcWGS, percent of tumor point-mutations detected, and mutational burden from WES of plasma cfDNA in the cohort. When lcWGS reaches the 3% TF threshold, the WES of the identical sample could detect more than 80% of tumor mutations and 70% of druggable CNVs. Interestingly, all druggable point mutations were detected by both WES and Panel-seq. Nevertheless, in samples with lower TF only 12 % of tumor point mutations were detected on average and very few druggable mutations and CNVs. At least 1 druggable point mutation was detected in 30% of WES and 10% Panel-seq. This suggests that the estimation of TF by using lcWGS can relatively guide the success of detecting using a more sensitive sequencing technique.

5.4 Fragment-length Analysis of CfDNA in Pediatric Cancers

5.4.1 Pediatric cancers shed short-fragmented cfDNA into the blood circulation.

Even though the underlying mechanism of the generation of cfDNA is not fully understood, the fragmentation analysis has found that cfDNA fragments were generated by mostly endonuclease activity as a part of the cell apoptosis process [72]. A previous experiment of xenografted human ovarian cancer have observed that human-derived (tumor) cfDNA is shorter than mouse-derived (non-tumor) cfDNA [110]. We assumed that pediatric cancer also releases cfDNA into blood circulation through a similar mechanism. We developed cfdnakit, a bioinformatics tool specialized in the fragment-length analysis of cfDNA. Using this package, we extracted the length of cfDNA and compare the sample fragment-length profile of human-derived and mouse-derived cfDNA. In this study of pediatric cancer, we found that tumor-derived cfDNA was shorter than cfDNA shed by non-malignant cells in the patient-derived xenograft (PDX) experiment. This implies that the secretion of DNA into the bloodstream of pediatric cancers and adult cancers is driven by the same underlying mechanism. The fragment length of tumor cfDNA in the PDX experiment has shown that the tumor cells always secrete shorter fragment lengths mainly 142 bases long on average which is the size of DNA wrapping around 1 unit of mononucleosome. The cause of this fragmentation pattern in tumor cfDNA has not been fully understood. It could be related to the differentiation stage of the tumor where chromatin repositioning and destabilization is common [180, 181].

5.4.2 Short-fragment cfDNA is enriched in cfDNA with high tumor-derived cfDNA.

Fragment lengths of plasma cfDNA have been mainly explored in adult cancers [110]. The number of short-fragment cfDNA increases accordingly with the concentration of tumor-derived cfDNA. Many studies have tried to explore the utilization of short-fragmented cfDNA as a quantitive measurement of tumor-derived cfDNA [110, 151, 182]. As previously mentioned, we found that pediatric tumors also release short-fragmented cfDNA into the blood circulation, we thus expected the enrichment of short-fragmented cfDNA to also correlate with the estimated tumor fraction reported by ichorCNA. In this study, we explored the fragment-length profile of cfDNA lcWGS in the pediatric cohort using cfdnakit. We observed an enrichment of short-fragmented cfDNA ($<$150 bases) among sarcomas and other pediatric cancers. In particular, it is because of the high prevalence of high-TF cfDNA among sarcomas and other pediatric cancers. CfDNA with a low tumor fraction contains a similar amount of short-fragmented cfDNA to cfDNA of healthy doners. The abundance of short cfDNA of a particular genomic region has been shown to correlate with absolute copy-number aberration found in the cfDNA. Moreover, selecting only short-fragmented cfDNA in-silico can enhance the detection of CNVs and increase the estimated tumor fraction in the pediatric cancer cohort. This finding is similar to previous experiments of adult pan-cancer [110]. However, the relationship between TF estimates and short-fragmented ratio (ratio of short-fragmented cfDNA over longer-fragmented cfDNA) is still unclear. Some low TF samples contain a relatively high short-fragmented ratio. The most possible cause could be that the tumor secretes

only small or point mutations rather than large CNVs. This means that using TF to infer the overall concentration of ctDNA with lcWGS could have overlooked the sample with an excessive point mutation rate.

In this study, we observed the overall short-fragment in cfDNA is associated with both copy-number aberrations and tumor mutational burden. The developed CPA-Score has done better to predict cfDNA samples with high mutation burden and likely to contains tumor-derived cfDNA by using lcWGS data than TF estimates. It could be used as a guiding measurement to increase the chance of detecting tumor point mutations with WES. The limitation of further evaluation of CPA score is the fact that cancer accumulates mutations through a lifetime thus the childhood cancers frequently have a lower mutation rate than adult cancers. Further evaluation should be performed in adult cancers in which more somatic mutations are acquired during lifetime and from exposure. It could also guide the utilization of WES or WGS to perform characterization of genomic features such as mutational signature analysis.

5.5 Detecting Telomeric Aberration and Insertion of Variant Repeats

A previous study has suggested that cfDNA originates from somatic cells [183]. The decreasing level of plasma telomeric cfDNA is associated with age in healthy individuals. Moreover, the level of telomeric cfDNA is decreasing among baseline breast cancer [183] and gastric cancer patients[184]. We assumed that plasma telomeric cfDNA might be able to indicate the alternative lengthening of telomeres (ALT) and the integration of telomeric variant repeat (TVR) into intratelomeric regions.

In this study, we firstly explored the telomeric alteration of tumors in the pediatric cohort with lcWGS. While most tumors had a decreasing telomere content, high-grade gliomas and osteosarcomas had increasing telomere content compared to their matched control. The correlation between telomere content and normalized count of TVRs was observed in brain tumors and sarcomas. Samples with ATRX functional mutation had an increased telomere content and frequently showed an integrated a pattern of variant repeats. These findings are in line with the analysis of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium dataset [21]. In addition to the PCAWG study, we found that high-grade gliomas with mutations in both ATRX and TP53 (Figure 47) had increased telomere contents. TGAGGG and TTCGGG singletons were frequently integrated into their intratelomeric regions.

Similar to previous studies [183, 184], we observed decreasing telomere content in most cfDNA samples when compared to a group of adult healthy donors. The pattern of TVR integration in cfDNA were ambiguous although they were still positively correlated with increasing telomere content. In brain tumors and sarcoma cfDNA, all TVR except TTCGGG were found to be integrated into elongating telomeres. Without normalization with matched control, it might not be suitable to compare the normalized count of telomere content between samples because telomeres shortens with increasing age. Therefore, the result of TVR plotting against telomere content might not correctly indicate the pattern and the frequency of TVR integration per increment of telomere content. This is the limitation of using sequencing data in this analysis.

Telomeric DNA might be depleted in plasma cell-free DNA as we found a significant decrease of telomere content comparing to control and tumor samples (Figure S4B). The mechanism behind the secretion of telomeric DNA into the circulation still remain elusive. In general, endonuclease activity of DFFB (DNA fragmentation factor sub-unit β) and DNASE1L3 (deoxyribonuclease 1-like3) as a part of the cell death program would cleave open-chromatin regions into highly fragmented cfDNA while leaving closed-chromatin regions mostly intact [72]. Therefore, it is possible that most of telomere DNA is released into the blood circulation as a large telomere-protein complex unit. The plasma cfDNA mostly contains small molecular units such as short-fragmented cfDNA which contains more tumor genetic mutation [185]. On the other hand, serum cfDNA has been shown to contain larger DNA fragments [185] and more telomeric cell-free DNA when treated with DNase [186]. It could be an opportunity to

isolate telomeric cell-free DNA from serum as a better source than plasma.

Although we could demonstrate that plasma cfDNA harbor telomeric alterations including elongation and integration of TVRs with lcWGS, the low concentration of tumor-derived cfDNA and the limited number of samples with ALT mutation hindered the clear interpretation of the results of this study. Sequencing lcWGS might not be a suitable strategy for detecting reads with TVRs since the power of detection is not high enough for low-ctDNA samples. Several studies recommended the enrichment for telomeric DNA and utilization of PCR-based detection strategies.

5.6 Application to Pediatric Cancer Patient Management

5.6.1 CfDNA reveals spatial tumor heterogeneity in a patient with bilateral Wilms tumor.

cfDNA in blood circulation is contributed from cells including from all tumor mass in the body. In this study, a cfDNA sample obtained from a patient with bilateral Wills tumor has shown discordance genomic profile with the primary tumor obtained from one of the patient's kidneys. We hypothesize that the source of aberration could be tumor mass located in another kidney where obtaining the tumor sample could have been complicated. The clinical data confirmed that this patient also suffered from multiple tumor metastase in the liver, lymph nodes, and abdominal wall. Recently, we have obtained additional tumor biopsies from another kidney and other tumor sites and are looking forward to generating their genomic profiles. With this information, we could clarify the origin of CNVs found in the cfDNA. We might have better evidence to support the utility of cfDNA as a surveillance liquid biopsy assay in pediatric cancer management. The information could be used in the therapeutic decision or suggesting the utilization of additional liquid biopsy assays such as single-cell sequencing of circulating tumor cells to precisely identify a druggable target. However, analyzing upcoming genomic data could not be performed in the time frame of this thesis.

5.6.2 Time-series liquid biopsy of cfDNA allows a tracking of tumor progression over a period of time.

The primary advantage of liquid biopsy is the non-invasiveness compared to tumor biopsy operations. It can be a source of tumor markers throughout the course of treatment for tracking the response of the tumor and notify the refractory of disease. In this study, we obtained several cfDNA samples collected in a pseudo-time-series manner. WES and lcWGS informed us about the detected mutations, comprehensive copy-number variants, and tumor fraction estimated from the cfDNA sample over the studied period. The lcWGS can notify us of the rising tumor clone without information on the tumor genome. It could be a cost-effective strategy when applied in clinical routine. However, their prognostic value needs to be evaluated per disease and clinical setting. Because of the lack of overlapping samples, our study could not combine the finding in lcWGS with mutation detection from WES to show how many point mutations were detected when the TF is high. Moreover, additional information at the time of biopsy such as clinical status, time relative to the start of therapy, or size of the tumor could give us a complete picture of the benefit of cfDNA in pediatric cancer management.

5.7 Limitations of the Study

First of all, plasma cfDNA samples were collected from the pediatric cohort where the tumor molecular diagnostic process has been well-established and the clinical status of tumors is mostly available. However, the information regarding the clinical status of patients per liquid biopsy sample was not always available to us. There was also variability in the clinical status of patients, the tumor diagnostical types, the time interval between tumor and liquid biopsy, and the treatment a patient received. This study mainly focuses on the technical development of detecting genomic aberrations from cfDNA. Future prospective studies of specific cancer types to evaluate the utility of cfDNA will require the information

of stages of disease progression and the precise information of liquid biopsy time interval.

Second, the number of collected liquid biopsies per diagnostical disease type varied and might not be enough to perform a comprehensive evaluation. The extended dataset of the pediatric cohort contains an additional number of samples obtained from brain tumors and rhabdomyosarcoma which could not be included into this thesis, however.

6 REFERENCES

1. Cunningham, R. M., Walton, M. A. & Carter, P. M. The major causes of death in children and adolescents in the United States. *New England Journal of Medicine* **379,** 2468–2475 (2018).
2. Steliarova-Foucher, E. *et al.* International incidence of childhood cancer, 2001–10: a population-based registry study. *The Lancet Oncology* **18,** 719–731 (2017).
3. Johnston, W. *et al.* Childhood cancer: estimating regional and global incidence. *Cancer epidemiology* **71,** 101662 (2021).
4. Siegel, D. A. *et al.* Pediatric cancer mortality and survival in the United States, 2001-2016. *Cancer* **126,** 4379–4389 (2020).
5. Curtin, S. C., Minino, A. M. & Anderson, R. N. Declines in cancer death rates among children and adolescents in the United States, 1999–2014. *NCHS data brief* **257,** 1–8 (2016).
6. Zwaan, C. M. *et al.* Collaborative efforts driving progress in pediatric acute myeloid leukemia. *Journal of clinical oncology* **33,** 2949 (2015).
7. Perkins, S. M., Shinohara, E. T., DeWees, T. & Frangoul, H. Outcome for children with metastatic solid tumors over the last four decades. *PloS one* **9,** e100396 (2014).
8. Dragomir, M. D. *OC-85 Early diagnosis in chilhood cancer saves lives* 2017.
9. Hawkes, N. *Cancer survival data emphasise importance of early diagnosis* 2019.
10. Organization, W. H. *et al.* Cancer control: early detection. *WHO guide to effective programmes. Geneva: World Health Organization* (2007).
11. *Childhood cancer* https://www.who.int/news-room/fact-sheets/detail/cancer-in-children. Accessed: 2021-08-05.
12. Singh, E., Naidu, G., Davies, M.-A. & Bohlius, J. HIV-associated malignancies in children. *Current Opinion in HIV and AIDS* **12,** 77 (2017).
13. Indolfi, G. *et al.* Hepatitis B virus infection in children and adolescents. *The lancet Gastroenterology & hepatology* **4,** 466–476 (2019).
14. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
15. Gröbner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555,** 321–327 (2018).
16. Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555,** 371–376 (2018).
17. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *cell* **100,** 57–70 (2000).
18. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674. ISSN: 0092-8674. `https://www.sciencedirect.com/science/article/pii/S0092867411001279` (2011).
19. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology* **15,** 81–94 (2018).
20. Sequist, L. V. *et al.* Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Science translational medicine* **3,** 75ra26–75ra26 (2011).

21. Sieverling, L. *et al.* Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nature communications* **11,** 1–13 (2020).

22. Henson, J. D. *et al.* DNA C-circles are specific and quantifiable markers of alternative-lengthening-of-telomeres activity. *Nature biotechnology* **27,** 1181–1185 (2009).

23. Lee, M. *et al.* Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic acids research* **42,** 1733–1746 (2014).

24. Henson, J. D. *et al.* The C-Circle Assay for alternative-lengthening-of-telomeres activity. *Methods* **114,** 74–84 (2017).

25. Arechederra, M., Ávila, M. A. & Berasain, C. Liquid biopsy for cancer management: a revolutionary but still limited new tool for precision medicine. *Advances in Laboratory Medicine/Avances en Medicina de Laboratorio* **1** (2020).

26. Hashim, O. H., Jayapalan, J. J. & Lee, C.-S. Lectins: an effective tool for screening of potential cancer biomarkers. *PeerJ* **5,** e3784 (2017).

27. Baid, A. ELISA-a mini review. *Res. Rev. J. Pharm. Anal* **5** (2016).

28. Gan, S. D., Patel, K. R., *et al.* Enzyme immunoassay and enzyme-linked immunosorbent assay. *J Invest Dermatol* **133,** e12 (2013).

29. Ashworth, T. A case of cancer in which cells similar to those in the tumours were seen in the blood after death. *Aust Med J.* **14,** 146 (1869).

30. Racila, E. *et al.* Detection and characterization of carcinoma cells in the blood. *Proceedings of the National Academy of Sciences* **95,** 4589–4594. ISSN: 0027-8424. eprint: `https://www.pnas.org/content/95/8/4589.full.pdf`. `https://www.pnas.org/content/95/8/4589` (1998).

31. Allard, W. J. *et al.* Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clinical cancer research* **10,** 6897–6904 (2004).

32. Foy, V., Fernandez-Gutierrez, F., Faivre-Finn, C., Dive, C. & Blackhall, F. The clinical utility of circulating tumour cells in patients with small cell lung cancer. *Translational lung cancer research* **6,** 409 (2017).

33. Gallo, M. *et al.* Clinical utility of circulating tumor cells in patients with non-small-cell lung cancer. *Translational lung cancer research* **6,** 486 (2017).

34. Zhang, T. & Armstrong, A. J. Clinical utility of circulating tumor cells in advanced prostate cancer. *Current oncology reports* **18,** 3 (2016).

35. Zhang, X. *et al.* Analysis of circulating tumor cells in ovarian cancer and their clinical value as a biomarker. *Cellular Physiology and Biochemistry* **48,** 1983–1994 (2018).

36. Chou, W.-C. *et al.* A prognostic model based on circulating tumour cells is useful for identifying the poorest survival outcome in patients with metastatic colorectal cancer. *International journal of biological sciences* **14,** 137 (2018).

37. Wang, W.-C. *et al.* Survival mechanisms and influence factors of circulating tumor cells. *BioMed research international* **2018** (2018).

38. Habli, Z., AlChamaa, W., Saab, R., Kadara, H. & Khraiche, M. L. Circulating tumor cell detection technologies and clinical utility: Challenges and opportunities. *Cancers* **12,** 1930 (2020).

39. Cubero, M. A. *et al.* in *Circulating Tumor Cells* 283–303 (Springer, 2017).

40. Dong, Y. *et al.* Microfluidics and circulating tumor cells. *The Journal of Molecular Diagnostics* **15,** 149–157 (2013).

41. Kamande, J. W. *et al.* Modular microsystem for the isolation, enumeration, and phenotyping of circulating tumor cells in patients with pancreatic cancer. *Analytical chemistry* **85,** 9092–9100 (2013).

42. Miller, M. C., Robinson, P. S., Wagner, C. & O'Shannessy, D. J. The Parsortixâ¢ cell separation systemâA versatile liquid biopsy platform. *Cytometry Part A* **93,** 1234–1239 (2018).

43. Schuur, E. R. Rapid and simple isolation of circulating tumor cells for clinical and research applications using ScreenCell®. *ScreenCell® devices: a flexible ctc platform* (2012).

44. Hillig, T. *et al.* In vitro detection of circulating tumor cells compared by the CytoTrack and CellSearch methods. *Tumor Biology* **36,** 4597–4601 (2015).

45. Somlo, G. *et al.* Multiple biomarker expression on circulating tumor cells in comparison to tumor tissues from primary and metastatic sites in patients with locally advanced/inflammatory, and stage IV breast cancer, using a novel detection technology. *Breast cancer research and treatment* **128,** 155–163 (2011).

46. Ogle, L. F. *et al.* Imagestream detection and characterisation of circulating tumour cells–A liquid biopsy for hepatocellular carcinoma? *Journal of hepatology* **65,** 305–313 (2016).

47. Gupta, V. *et al.* ApoStreamâ¢, a new dielectrophoretic device for antibody independent isolation and recovery of viable cancer cells from blood. *Biomicrofluidics* **6,** 024133 (2012).

48. Di Trapani, M., Manaresi, N. & Medoro, G. DEPArrayâ¢ system: An automatic image-based sorter for isolation of pure circulating tumor cells. *Cytometry Part A* **93,** 1260–1266 (2018).

49. Gabriel, M. T., Calleja, L. R., Chalopin, A., Ory, B. & Heymann, D. Circulating tumor cells: a review of non–EpCAM-based approaches for cell enrichment and isolation. *Clinical chemistry* **62,** 571–581 (2016).

50. Dai, J. *et al.* Exosomes: Key players in cancer and potential therapeutic strategy. *Signal Transduction and Targeted Therapy* **5,** 1–10 (2020).

51. Valencia, K. & Montuenga, L. M. Exosomes in Liquid Biopsy: The Nanometric World in the Pursuit of Precision Oncology. *Cancers* **13,** 2147 (2021).

52. Keerthikumar, S. *et al.* ExoCarta: a web-based compendium of exosomal cargo. *Journal of molecular biology* **428,** 688–692 (2016).

53. Ponti, G., Manfredini, M. & Tomasi, A. Non-blood sources of cell-free DNA for cancer molecular profiling in clinical pathology and oncology. *Critical reviews in oncology/hematology* **141,** 36–42 (2019).

54. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164,** 57–68 (2016).

55. Wan, J. C. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer* **17,** 223–238 (2017).

56. Kustanovich, A., Schwartz, R., Peretz, T. & Grinshpun, A. Life and death of circulating cell-free DNA. *Cancer biology & therapy* **20,** 1057–1067 (2019).

57. Yu, D. *et al.* Diagnostic value of concentration of circulating cell-free DNA in breast cancer: a meta-analysis. *Frontiers in oncology* **9,** 95 (2019).

58. Miao, Y., Fan, Y., Zhang, L., Ma, T. & Li, R. Clinical value of plasma cfDNA concentration and integrity in breast cancer patients. *Cellular and Molecular Biology* **65,** 64–72 (2019).

59. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early-and late-stage human malignancies. *Science translational medicine* **6,** 224ra24–224ra24 (2014).

60. Yeh, P. *et al.* Circulating tumour DNA reflects treatment response and clonal evolution in chronic lymphocytic leukaemia. *Nature communications* **8,** 1–7 (2017).

61. Mandel, P. Les acides nucleiques du plasma sanguin chez 1 homme. *CR Seances Soc Biol Fil* **142,** 241–243 (1948).

62. Tan, E., Schur, P., Carr, R., Kunkel, H., *et al.* Deoxybonucleic acid (DNA) and antibodies to DNA in the serum of patients with systemic lupus erythematosus. *The Journal of clinical investigation* **45,** 1732–1740 (1966).

63. Leon, S., Shapiro, B., Sklaroff, D. & Yaros, M. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer research* **37,** 646–650 (1977).

64. Stroun, M. *et al.* Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* **46,** 318–322 (1989).

65. Sorenson, G. D. *et al.* Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiology and Prevention Biomarkers* **3,** 67–71 (1994).

66. Vasioukhin, V. *et al.* Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *British journal of haematology* **86,** 774–779 (1994).

67. Chen, X. Q. *et al.* Microsatellite alterations in plasma DNA of small cell lung cancer patients. *Nature medicine* **2,** 1033–1035 (1996).

68. Nawroz, H., Koch, W., Anker, P., Stroun, M. & Sidransky, D. Microsatellite alterations in serum DNA of head and neck cancer patients. *Nature medicine* **2,** 1035–1037 (1996).

69. Oliveira, K. C. *et al.* Current perspectives on circulating tumor DNA, precision medicine, and personalized clinical management of cancer. *Molecular Cancer Research* **18,** 517–528 (2020).

70. Chen, S., Liu, M. & Zhou, Y. in *Computational Systems Biology: Methods and Protocols* (ed Huang, T.) 67–95 (Springer New York, New York, NY, 2018). ISBN: 978-1-4939-7717-8. `https://doi.org/10.1007/978-1-4939-7717-8_5`.

71. Zhang, L. *et al.* The interplay of circulating tumor DNA and chromatin modification, therapeutic resistance, and metastasis. *Molecular cancer* **18,** 1–20 (2019).

72. Corcoran, R. B. & Chabner, B. A. Application of cell-free DNA analysis to cancer treatment. *New England Journal of Medicine* **379,** 1754–1765 (2018).

73. Lo, Y. D. *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Science translational medicine* **2,** 61ra91–61ra91 (2010).

74. De Vlaminck, I. *et al.* Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Science translational medicine* **6,** 241ra77–241ra77 (2014).

75. Jiang, P. *et al.* Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proceedings of the National Academy of Sciences* **115,** E10925–E10933 (2018).

76. Ulz, P. *et al.* Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nature communications* **10,** 1–11 (2019).

77. Li, M., Jia, Y., Xu, J., Cheng, X. & Xu, C. Assessment of the circulating cell-free DNA marker association with diagnosis and prognostic prediction in patients with lymphoma: a single-center experience. *Annals of hematology* **96,** 1343–1351 (2017).

78. Vittori, L. N., Tarozzi, A. & Latessa, P. M. in *Cell-free DNA as Diagnostic Markers* 183–197 (Springer, 2019).

79. Hammad, R. *et al.* Circulating cell-free DNA, peripheral lymphocyte subsets alterations and neutrophil lymphocyte ratio in assessment of COVID-19 severity. *Innate Immunity* **27,** 240–250 (2021).

80. Doherty, C. M. & Forbes, R. B. Diagnostic lumbar puncture. *The Ulster medical journal* **83,** 93 (2014).

81. Wang, Y. *et al.* Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas. *Science translational medicine* **7,** 293ra104–293ra104 (2015).

82. Li, Y. *et al.* Unique genetic profiles from cerebrospinal fluid cell-free DNA in leptomeningeal metastases of EGFR-mutant non-small-cell lung cancer: a new medium of liquid biopsy. *Annals of Oncology* **29,** 945–952 (2018).

83. Feil, G. & Stenzl, A. Tumor marker tests in bladder cancer. *Actas urologicas espanolas* **30,** 38–45 (2006).

84. Salvi, S. *et al.* The potential use of urine cell free DNA as a marker for cancer. *Expert review of molecular diagnostics* **16,** 1283–1290 (2016).

85. Yokota, M., Tatsumi, N., Tsuda, I., Takubo, T. & Hiyoshi, M. DNA extraction from human urinary sediment. *Journal of clinical laboratory analysis* **12,** 88–91 (1998).

86. Su, Y.-H. *et al.* Removal of high-molecular-weight DNA by carboxylated magnetic beads enhances the detection of mutated K-ras DNA in urine. *Annals of the New York Academy of Sciences* **1137,** 82 (2008).

87. Su, Y.-H. *et al.* Human urine contains small, 150 to 250 nucleotide-sized, soluble DNA derived from the circulation and may be useful in the detection of colorectal cancer. *The journal of molecular diagnostics* **6,** 101–107 (2004).

88. Lu, T. & Li, J. Clinical applications of urinary cell-free DNA in cancer: current insights and promising future. *American journal of cancer research* **7,** 2318 (2017).

89. Kimura, H. *et al.* EGFR mutation status in tumour-derived DNA from pleural effusion fluid is a practical basis for predicting the response to gefitinib. *British journal of cancer* **95,** 1390–1395 (2006).

90. Soh, J. *et al.* Usefulness of EGFR mutation screening in pleural fluid to predict the clinical outcome of gefitinib treated patients with lung cancer. *International journal of cancer* **119,** 2353–2358 (2006).

91. Kawahara, A. *et al.* Epidermal growth factor receptor mutation status in cell-free DNA supernatant of bronchial washings and brushings. *Cancer cytopathology* **123,** 620–628 (2015).

92. Shi, C. *et al. Analysis of mutation detection in cell-free DNA in ascites using comprehensive NGS panel.* 2019.

93. Husain, H. *et al.* Cell-free DNA from ascites and pleural effusions: molecular insights into genomic aberrations and disease biology. *Molecular cancer therapeutics* **16,** 948–955 (2017).

94. Werner, B. *et al.* Cell-free DNA is abundant in ascites and represents a liquid biopsy of ovarian cancer. *Gynecologic Oncology* (2021).

95. Imperiale, T. F. *et al.* Multitarget stool DNA testing for colorectal-cancer screening. *New England Journal of Medicine* **370,** 1287–1297 (2014).

96. Moding, E. J., Diehn, M. & Wakelee, H. A. Circulating tumor DNA testing in advanced non-small cell lung cancer. *Lung Cancer* **119,** 42–47 (2018).

97. Cimmino, F., Lasorsa, V. A., Vetrella, S., Iolascon, A. & Capasso, M. A targeted gene panel for circulating tumor DNA sequencing in neuroblastoma. *Frontiers in oncology* **10** (2020).

98. Hastings, R. *et al.* Longitudinal whole-exome sequencing of cell-free DNA unravels the metastatic evolutionary dynamics of BRCA2-mutated breast cancer. *bioRxiv* (2020).

99. Chicard, M. *et al.* Whole-exome sequencing of cell-free DNA reveals temporo-spatial heterogeneity and identifies treatment-resistant clones in neuroblastoma. *Clinical Cancer Research* **24,** 939–949 (2018).

100. Manier, S. *et al.* Whole-exome sequencing of cell-free DNA and circulating tumor cells in multiple myeloma. *Nature communications* **9,** 1–11 (2018).

101. Seaby, E. G., Pengelly, R. J. & Ennis, S. Exome sequencing explained: a practical guide to its clinical application. *Briefings in functional genomics* **15,** 374–384 (2016).

102. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature communications* **8,** 1–13 (2017).

103. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods* **11,** 163–166 (2014).

104. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research* **27,** 491–499 (2017).

105. Kim, M.-J., Kim, S.-C. & Kim, Y.-J. A Universal Analysis Pipeline for Hybrid Capture-Based Targeted Sequencing Data with Unique Molecular Indexes. *Genomics & informatics* **16** (2018).

106. Huang, C.-C., Du, M. & Wang, L. Bioinformatics analysis for circulating cell-free DNA in cancer. *Cancers* **11,** 805 (2019).

107. Yang, Y. *et al.* Detecting Ultralow Frequency Mutation in Circulating Cell-Free DNA of Early-Stage Nonsmall Cell Lung Cancer Patients with Unique Molecular Identifiers. *Small Methods* **3,** 1900206 (2019).

108. Ng, H. I. *et al.* Analysis of fragment size distribution of cell-free DNA: A potential non-invasive marker to monitor graft damage in living-related liver transplantation for inborn errors of metabolism. *Molecular genetics and metabolism* **127,** 45–50 (2019).

109. Shi, J., Zhang, R., Li, J. & Zhang, R. Size profile of cell-free DNA: A beacon guiding the practice and innovation of clinical testing. *Theranostics* **10,** 4737 (2020).

110. Mouliere, F. *et al.* Enhanced detection of circulating tumor DNA by fragment size analysis. *Science translational medicine* **10** (2018).

111. Thierry, A. R. *et al.* Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. *Nucleic acids research* **38,** 6159–6175 (2010).

112. Underhill, H. R. *et al.* Fragment length of circulating tumor DNA. *PLoS genetics* **12,** e1006162 (2016).

113. Mouliere, F., El Messaoudi, S., Pang, D., Dritschilo, A. & Thierry, A. R. Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. *Molecular oncology* **8,** 927–941 (2014).

114. Worst, B. C. *et al.* Next-generation personalised medicine for high-risk paediatric cancer patients–The INFORM pilot study. *European journal of cancer* **65,** 91–101 (2016).

115. Sahm, F. *et al.* Next-generation sequencing in routine brain tumor diagnostics enables an integrated diagnosis and identifies actionable targets. *Acta neuropathologica* **131,** 903–910 (2016).

116. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

117. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31,** 2032–2034 (2015).

118. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

119. Rimmer, A. *et al.* Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics* **46,** 912–918 (2014).

120. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38,** e164–e164 (2010).

121. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22,** 1760–1774 (2012).

122. Reisinger, E. *et al.* OTP: An automatized system for managing and processing NGS data. *Journal of biotechnology* **261,** 53–62 (2017).

123. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS computational biology* **12,** e1004873 (2016).

124. Warsow, G. *et al.* Genomic features of renal cell carcinoma with venous tumor thrombus. *Scientific reports* **8,** 1–12 (2018).

125. Kleinheinz, K. *et al.* ACEseq–allele specific copy number estimation from whole genome sequencing. *BioRXiv,* 210807 (2017).

126. Fulcrum Genomics. *Tools for working with genomic and high throughput sequencing data* `http://fulcrumgenomics.github.io/fgbio/`. (Accessed: 2020/12/17; version 1.1.0).

127. Broad Institute. *Picard Tools* `http://broadinstitute.github.io/picard/`. (Accessed: 2018/02/21; version 2.17.8).

128. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research* **41,** e67–e67 (2013).

129. Johansson, L. F. *et al.* NIPTeR: an R package for fast and accurate trisomy prediction in non-invasive prenatal testing. *BMC bioinformatics* **19,** 1–5 (2018).

130. Riester, M. *et al.* PureCN: copy number calling and SNV classification using targeted short read sequencing. *Source code for biology and medicine* **11,** 1–13 (2016).

131. Benjamin, D. *et al.* Calling somatic SNVs and indels with Mutect2. *Biorxiv,* 861054 (2019).

132. Feuerbach, L. *et al.* TelomereHunter–in silico estimation of telomere content and composition from cancer genomes. *BMC bioinformatics* **20,** 1–11 (2019).

133. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570,** 385–389 (2019).

134. Liu, Y. *et al.* Increased detection of circulating tumor DNA by short fragment enrichment. *Translational lung cancer research* **10,** 1501 (2021).

135. Liu, X. *et al.* Fragment enrichment of circulating tumor DNA with low-frequency mutations. *Frontiers in genetics* **11,** 147 (2020).

136. Morgan, M., Pagès, H., Obenchain, V. & Hayden, N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. *R package version* **1,** 677–689 (2016).

137. Markus, H. *et al.* Sub-nucleosomal organization in urine cell-free DNA. *BioRxiv,* 696633 (2019).

138. Mouliere, F. *et al.* Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. *EMBO molecular medicine* **10,** e9323 (2018).

139. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Scientific reports* **9,** 1–5 (2019).

140. Karolchik, D. *et al.* The UCSC genome browser database. *Nucleic acids research* **31,** 51–54 (2003).

141. Scheinin, I. *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. *Genome research* **24,** 2022–2032 (2014).

142. Olshen, A. B., Venkatraman, E., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5,** 557–572 (2004).

143. Seshan, V. E., Olshen, A., *et al.* DNAcopy: DNA copy number data analysis. *R package version* **1** (2016).

144. Olshen, A. B. *et al.* Parent-specific copy number in paired tumor–normal studies using circular binary segmentation. *Bioinformatics* **27,** 2038–2046 (2011).

145. Raman, L. *et al.* Shallow whole-genome sequencing of plasma cell-free DNA accurately differentiates small from non-small cell lung carcinoma. *Genome medicine* **12,** 1–12 (2020).

146. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome biology* **14,** 1–20 (2013).

147. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).

148. Joseph, C. *et al.* Elevated MMP9 expression in breast cancer is a predictor of shorter patient survival. *Breast cancer research and treatment* **182,** 267–282 (2020).

149. Du, R., Huang, C., Liu, K., Li, X. & Dong, Z. Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy. *Molecular Cancer* **20,** 1–27 (2021).

150. Carroll, M. & Borden, K. L. The oncogene eIF4E: using biochemical insights to target cancer. *Journal of Interferon & Cytokine Research* **33,** 227–238 (2013).

151. Tsui, D. W. *et al.* Tumor fraction-guided cell-free DNA profiling in metastatic solid tumor patients. *Genome medicine* **13,** 1–15 (2021).

152. Minasi, S. *et al.* Alternative lengthening of telomeres in molecular subgroups of paediatric high-grade glioma. *Child's Nervous System* **37,** 809–818 (2021).

153. Lee, J., Solomon, D. A. & Tihan, T. The role of histone modifications and telomere alterations in the pathogenesis of diffuse gliomas in adults and children. *Journal of neuro-oncology* **132,** 1–11 (2017).

154. Mukherjee, J. *et al.* Mutant IDH1 cooperates with ATRX loss to drive the alternative lengthening of telomere phenotype in glioma. *Cancer research* **78,** 2966–2977 (2018).

155. Ruteshouser, E. C., Robinson, S. M. & Huff, V. Wilms tumor genetics: mutations in WT1, WTX, and CTNNB1 account for only about one-third of tumors. *Genes, Chromosomes and Cancer* **47,** 461–470 (2008).

156. Zhou, C., Li, W., Shao, J., Zhao, J. & Chen, C. Analysis of the clinicopathologic characteristics of lung adenocarcinoma with CTNNB1 mutation. *Frontiers in genetics* **10,** 1367 (2020).

157. Arnold, A. *et al.* The majority of β-catenin mutations in colorectal cancer is homozygous. *BMC cancer* **20,** 1–10 (2020).

158. Kurnit, K. C. *et al.* CTNNB1 (beta-catenin) mutation identifies low grade, early stage endometrial cancer patients at increased risk of recurrence. *Modern Pathology* **30,** 1032–1041 (2017).

159. Van Schie, E. H. & van Amerongen, R. Aberrant WNT/CTNNB1 signaling as a therapeutic target in human breast Cancer: weighing the evidence. *Frontiers in cell and developmental biology* **8,** 25 (2020).

160. Na, T.-Y., Schecterson, L., Mendonsa, A. M. & Gumbiner, B. M. The functional activity of E-cadherin controls tumor cell metastasis at multiple steps. *Proceedings of the National Academy of Sciences* **117,** 5931–5937 (2020).

161. Zaman, G. J. *et al.* TTK inhibitors as a targeted therapy for CTNNB1 (β-catenin) mutant cancers. *Molecular cancer therapeutics* **16,** 2609–2617 (2017).

162. Xiu, M.-X. & Liu, Y.-M. The role of oncogenic Notch2 signaling in cancer: a novel therapeutic target. *American journal of cancer research* **9,** 837 (2019).

163. Paul, M. D. & Hristova, K. The RTK interactome: overview and perspective on RTK heterointeractions. *Chemical reviews* **119,** 5881–5921 (2018).

164. Roskoski Jr, R. Small molecule inhibitors targeting the EGFR/ErbB family of protein-tyrosine kinases in human cancers. *Pharmacological research* **139,** 395–411 (2019).

165. Triulzi, T., Bianchi, G. V. & Tagliabue, E. Predictive biomarkers in the treatment of HER2-positive breast cancer: an ongoing challenge. *Future Oncology* **12,** 1413–1428 (2016).

166. Wen, W. *et al.* Mutations in the kinase domain of the HER2/ERBB2 gene identified in a wide variety of human cancers. *The Journal of Molecular Diagnostics* **17,** 487–495 (2015).

167. De Giovanni, C. *et al.* HER Tyrosine Kinase Family and Rhabdomyosarcoma: Role in Onset and Targeted Therapy. *Cells* **10,** 1808 (2021).

168. Vaishnavi, A., Le, A. T. & Doebele, R. C. TRKing down an old oncogene in a new era of targeted therapy. *Cancer discovery* **5,** 25–34 (2015).

169. Somwar, R. *et al.* NTRK kinase domain mutations in cancer variably impact sensitivity to type I and type II inhibitors. *Communications biology* **3,** 1–13 (2020).

170. Aburjania, Z. *et al.* The role of notch3 in cancer. *The oncologist* **23,** 900 (2018).

171. Borggrefe, T. & Oswald, F. The Notch signaling pathway: transcriptional regulation at Notch target genes. *Cellular and Molecular Life Sciences* **66,** 1631–1646 (2009).

172. Xiu, M. *et al.* The Role of Notch3 Signaling in Cancer Stemness and Chemoresistance: Molecular Mechanisms and Targeting Strategies. *Frontiers in Molecular Biosciences* **8,** 545 (2021).

173. Grimmig, T. *et al.* TLR7 and TLR8 expression increases tumor cell proliferation and promotes chemoresistance in human pancreatic cancer. *International journal of oncology* **47,** 857–866 (2015).

174. Wang, Y. *et al.* Development of A Novel Highly Selective TLR8 Agonist for Cancer Immunotherapy. *bioRxiv* (2020).

175. Brodeur, G. M., Nichols, K. E., Plon, S. E., Schiffman, J. D. & Malkin, D. Pediatric cancer predisposition and surveillance: an overview, and a tribute to Alfred G. Knudson Jr. *Clinical Cancer Research* **23,** e1–e5 (2017).

176. Abbosh, C. *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545,** 446–451 (2017).

177. Jamal-Hanjani, M. *et al.* Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine* **376,** 2109–2121 (2017).

178. Xia, R., Vattathil, S. & Scheet, P. Identification of allelic imbalance with a statistical model for subtle genomic mosaicism. *PLoS computational biology* **10,** e1003765 (2014).

179. Oh, S. *et al.* Reliable analysis of clinical tumor-only whole-exome sequencing data. *JCO clinical cancer informatics* **4,** 321–335 (2020).

180. Wolffe, A. P. Chromatin remodeling: why it is important in cancer. *Oncogene* **20,** 2988–2990 (2001).

181. Amatori, S., Tavolaro, S., Gambardella, S. & Fanelli, M. The dark side of histones: genomic organization and role of oncohistones in cancer. *Clinical Epigenetics* **13,** 1–21 (2021).

182. Lapin, M. *et al.* Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer. *Journal of translational medicine* **16,** 1–10 (2018).

183. Wu, X. & Tanaka, H. Aberrant reduction of telomere repetitive sequences in plasma cell-free DNA for early breast cancer detection. *Oncotarget* **6,** 29795 (2015).

184. Shi, Y. *et al.* Telomere Length of Circulating Cell-Free DNA and Gastric Cancer in a Chinese Population at High-Risk. *Frontiers in oncology* **9,** 1434 (2019).

185. Lee, J.-S. *et al.* Plasma vs. serum in circulating tumor DNA measurement: characterization by DNA fragment sizing and digital droplet polymerase chain reaction. *Clinical Chemistry and Laboratory Medicine (CCLM)* **58,** 527–532 (2020).

186. Zinkova, A., Brynychova, I., Svacina, A., Jirkovska, M. & Korabecna, M. Cell-free DNA from human plasma and serum differs in content of telomeric sequences and its ability to promote immune response. *Scientific reports* **7,** 1–8 (2017).

187. Oliveros, J. *Venny. An interactive tool for comparing lists with Venn's diagrams. 2007–2015* (Accessed: 2021/09/19; version 2.1).

7 AUTHOR'S PUBLICATIONS, POSTERS AND TALKS

Manuscripts

From Sampling to Sequencing: A Liquid Biopsy Pre-Analytic Workflow to Maximize Multi-Layer Genomic Information from a Single Tube.

Kendra K. Maass*, Paulina S. Schad, Agnes ME Finster, **Pitithat Puranachot**, Fabian Rosing, Tatjana Wedig, Nathalie Schwarz, Natalie Stumpf, Stefan M. Pfister, and Kristian W. Pajtler.

(Cancers 13, no. 12 (2021): 3002.)

Orthogonal Comparison of Next Generation Sequencing Approaches to Benchmark Liquid Biopsy Analysis for Pediatric Tumor Patients – The INFORM Experience

Kendra K. Maass*, **Pitithat Puranachot***, Paulina S. Schad, Agnes ME Finster, Stefanie Zimmermann, Gnana Prakash Balasubramanian, Benedikt Brors, Kristian W. Pajtler

(in preparation)

Poster Presentations

A Novel Tool for Quantitative Measures of Genome wide Cell free DNA Fragmentation in Cancer Patient

Pitithat Puranachot, Steffen Dietz, Holger Sültmann, Benedikt Brors

(The German Conference on Bioinformatics 2019 (GCB), 2019)

Tracking chromosomal instability by short-fragmented circulating cell-free DNA

Pitithat Puranachot, Steffen Dietz, Holger Sültmann, Benedikt Brors

(DKFZ PhD Poster Presentation, 2019)

Conference Talks

Demonstration of Cancer Bioinformatics

Gnana Prakash Balasubramanian, **Pitithat Puranachot**

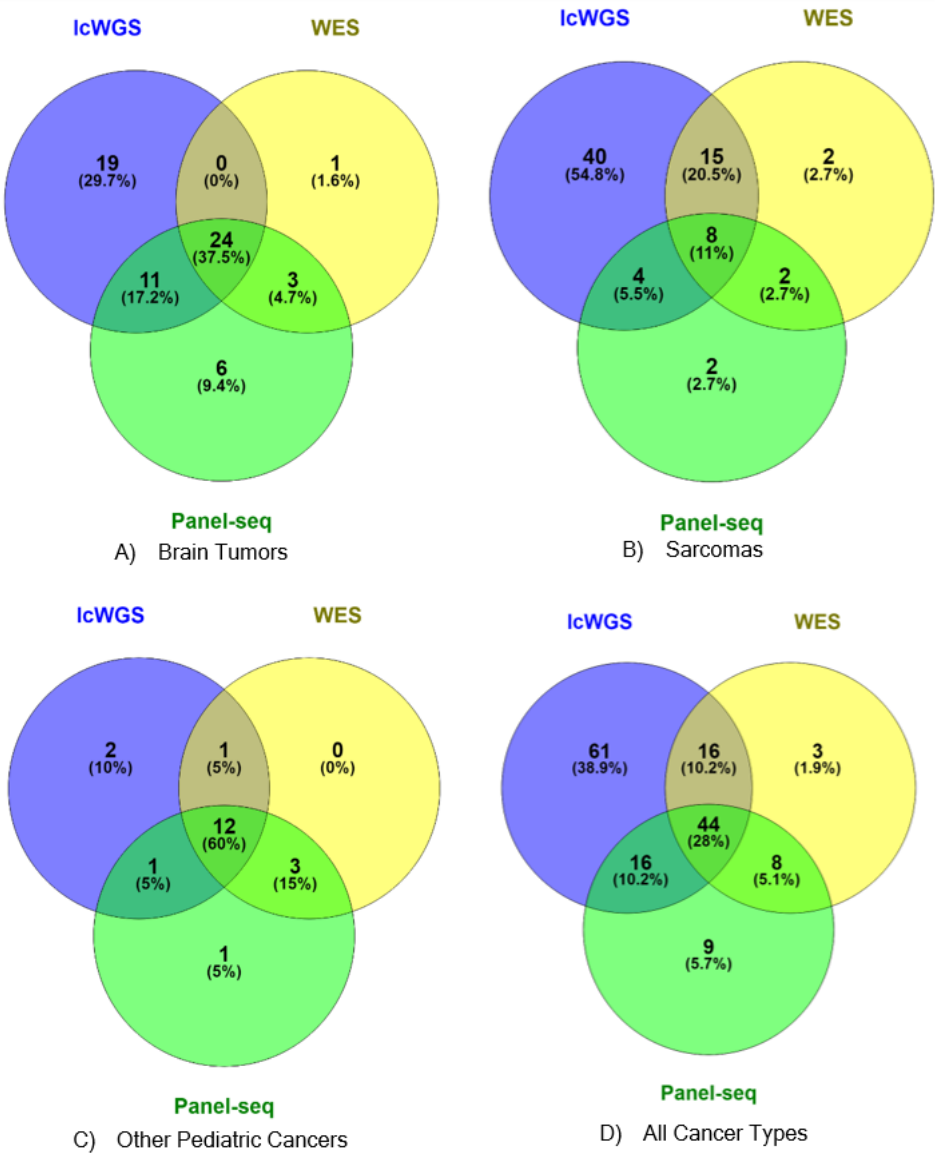
(Pre-Congress Programs, Princess Chulabhorn International Oncology Conference 2019)

8 ACKNOWLEDGEMENTS

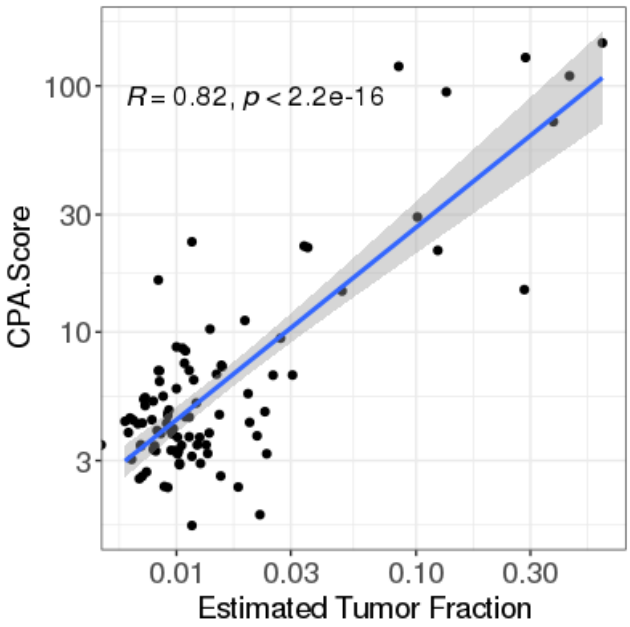
First and foremost, I would like to thank my supervisor, Prof. Benedikt Brors, for the opportunity to join the ABI research group and his kind supervision during my Ph.D. He had always given helpful advice, comments, and suggestion when I ready need. I would like to show gratitude to Prof. Stefan Fröhling and Simon Anders for being my Thesis Advisory Committee (TAC) and their helpful suggestions. I would like to thanks my collaborators Kendra Maaß, Paulina Schad, Agnes Finster, and Kristian Pajtler from KiTZ Hopp Children's Cancer Center in Heidelberg for working together with great effort and energy. I would also like to thanks all my friends and colleagues from ABI, and former BODA for all their suggestions, encouragement, and friendly working environment. I would like to thank Siao-Han Wong for her advice on writing and planning the thesis, and Birgit Vey and Corinna Sprengart for helping with administrative matters and support. Additional thanks is given to Charles Imbusch who helps me translating the abtract of this dissertation from English to German Zusammenfassung. I am grateful to two internship students, Diba Rafii and Ronja Völk, for their participation and good piece of works. I would thanks Nattharat Punyasu for her moral support through a hard time, and to Ms. Srirakit for being a very supportive flatmate. I would like to give a special thanks to Sasithorn Chotewutmontri for giving me a recommendation to DKFZ. I would give my full gratitude to my family for their support, understanding, and patience during my entire education. Lastly, special gratitude is expressed to Chulabhorn Royal Academy for granting the Scholarship in Commemoration of HM King Bhumibol Adulyadej's 90th Birthday Anniversary.

9 APPENDICES

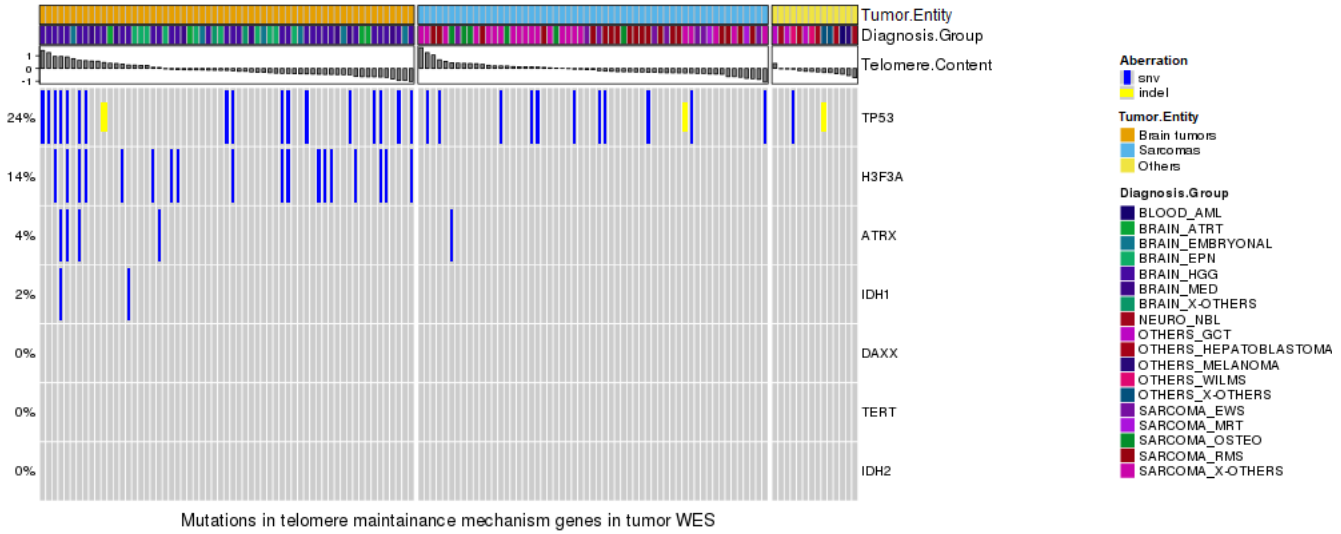
9.1 Supplementary Figures



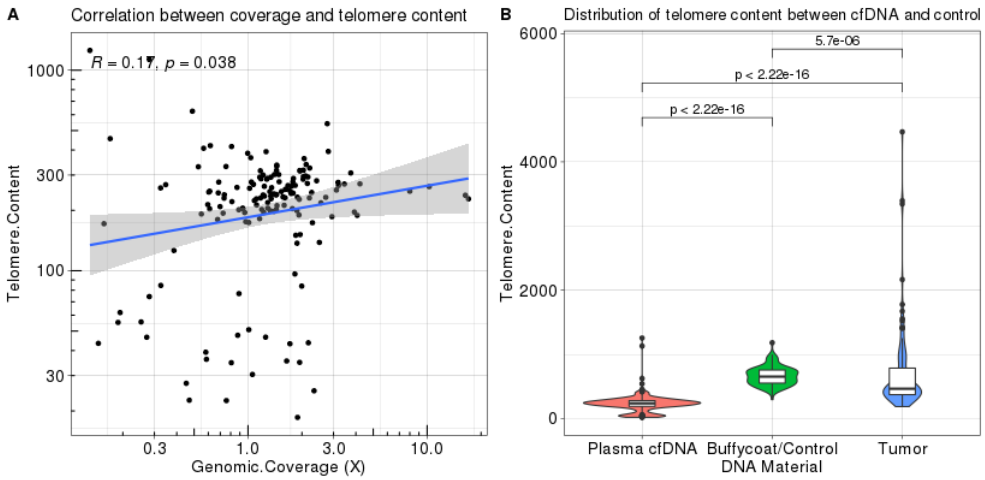
Supplementary Figure S1: The number of cfDNA next-generation sequencing data and the overlapping by tumor entity: A) Brain tumors B) Sarcomas C) Other Pediatric Cancers and D) All Cancer Types; The venn diagrams were generated by using Venny 2.1 [187].



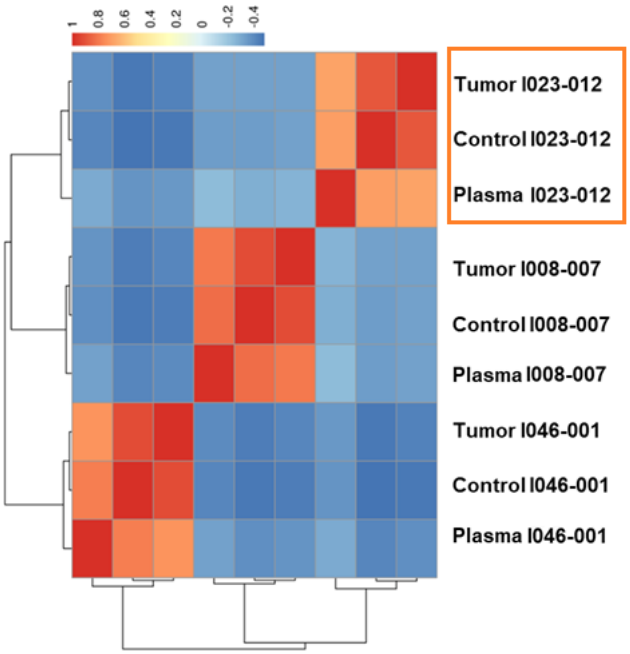
Supplementary Figure S2: Correlation between estimated tumor fraction and CPA Score of the pediatric cohort samples



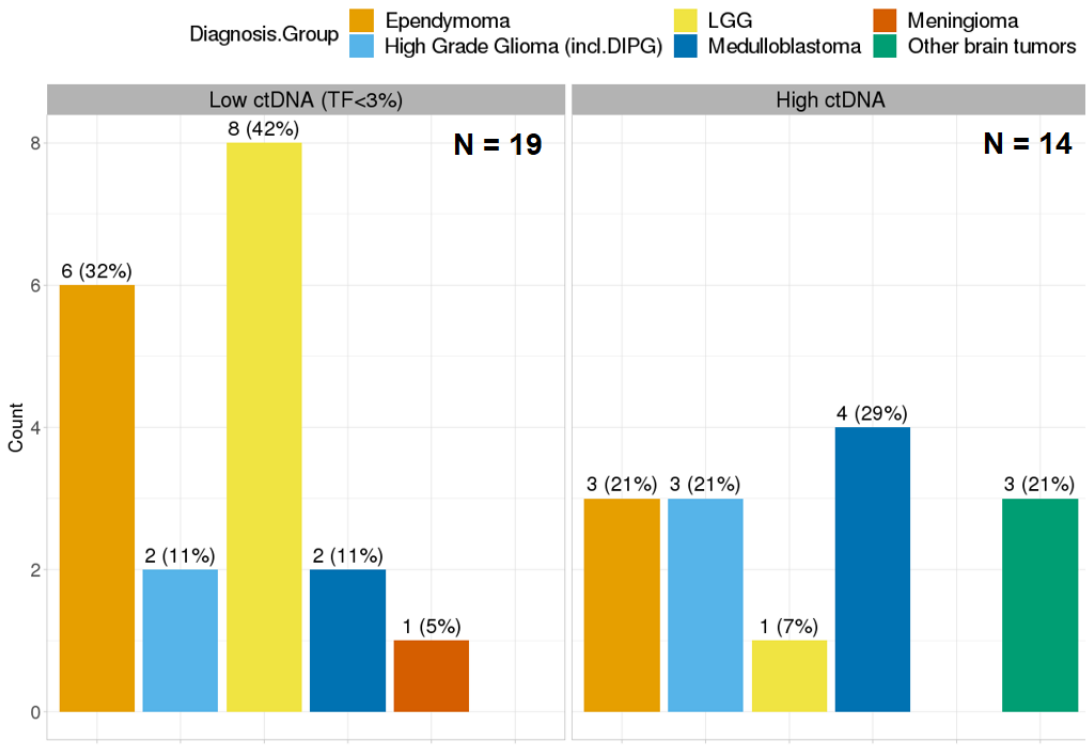
Supplementary Figure S3: Tumor point mutations in telomeric maintainance mechanism genes: Point mutations were detected from tumor WES. Telomore contents were extimated by TelomereHunter using tumor lcWGS.



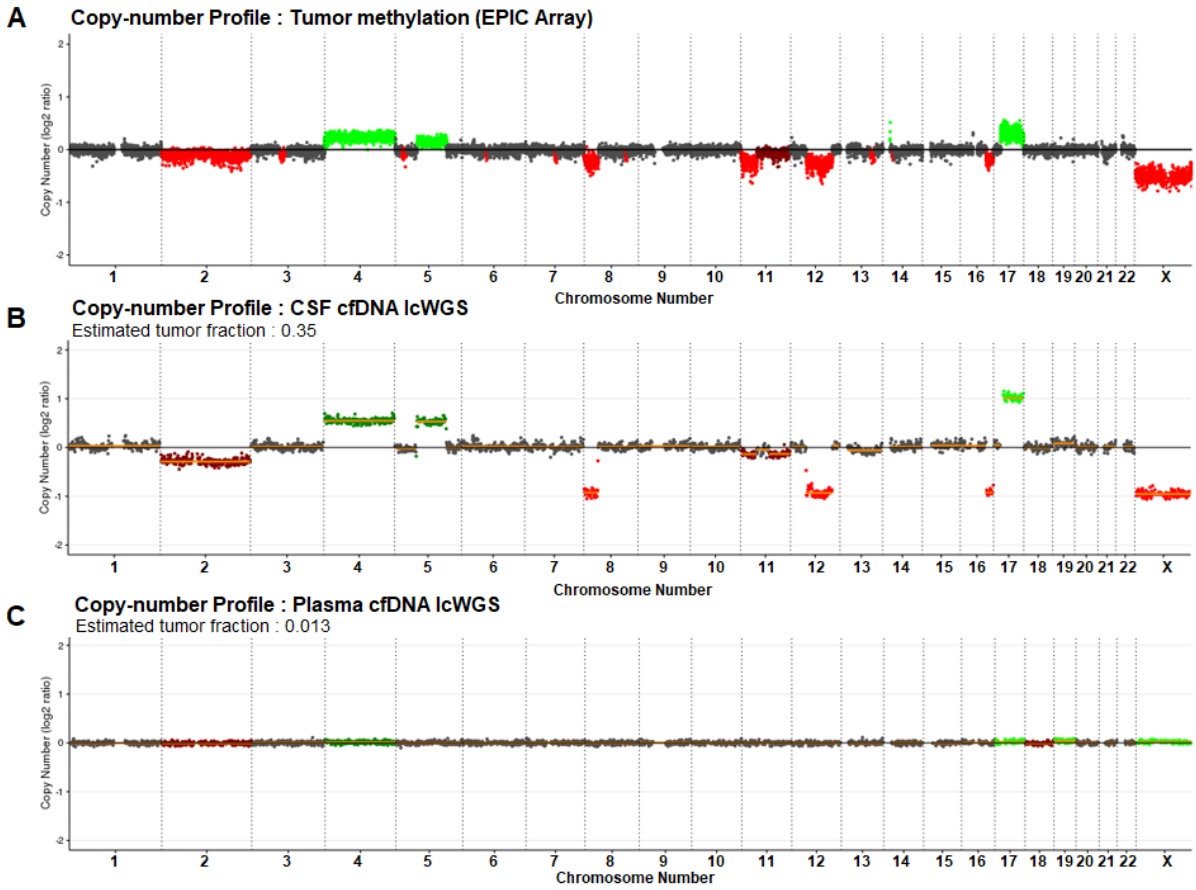
Supplementary Figure S4: Estimated telomere content in cell-free DNA with lcWGS: (A) The telomere content estimated from lcWGS of cfDNA samples (n=156) is weakly correlated with the sequencing coverage (Pearson correlation coefficient 0.17). (B) Telomere content in cfDNA is significantly lower than in buffycoat/control and tumor samples (Wilcoxon rank sum test).



Supplementary Figure S5: Genotypic fingerprint checking of the bilateral wilms tumor DNA The heatmap shows genotyping correlation matrix of samples derived from 3 different individuals. The correlation coefficient between samples from the patient with bilateral wilms tumor (in orange square) are clustered together. This confirms that the cfDNA were derived from the patient rather than another individual.



Supplementary Figure S6: Estimated tumor-fraction from cfDNA in cerebrospinal fluid of brain tumor patients: Number of high ctDNA (right panel) and low ctDNA (left panel) samples per tumor diagnostic group; Low Grade Glioma (LGG); Tumor Fraction (TF)



Supplementary Figure S7: Detecting CNVs and estimating tumor fraction from cerebrospinal fluid of a medulloblastoma patient: CfDNA samples were collected from cerebrospinal fluid (CSF) and blood plasma of the patient. Most of tumor CNVs were detected in CSF rather than plasma cfDNA. (A) Genomic CNV profile of tumor genome; (B) Genomic CNV profile of CSF cfDNA; (C) CNV profile of plasma cfDNA

9.2 Supplementary Tables

	lcWGS	WES	Panel-seq	Total
Brain Tumors	54	28	44	126
Sarcomas	67	27	17	110
Others Pediatric Cancers	16	16	17	49
Total	137	71	77	285

Supplementary Table S1: Total number of cfDNA next-generation sequencing dataset of the INFORM cohort

Supplementary Table S2: List of pediatric cancer druggable genes

chr	start	end	strand	gene
1	1138888	1142071	-1	TNFRSF18
1	1146706	1149518	-1	TNFRSF4
1	7979907	8000926	-1	TNFRSF9
1	8064464	8086368	-1	ERRFI1
1	9711790	9789172	1	PIK3CD
1	11166592	11322564	-1	MTOR
1	12123434	12204264	1	TNFRSF8
1	16450832	16482582	-1	EPHA2
1	23037332	23241818	1	EPHB2
1	23345941	23410182	1	KDM1A
1	26644448	26647014	1	CD52
1	26856252	26901521	1	RPS6KA1
1	27022524	27108595	1	ARID1A
1	27938575	27961788	-1	FGR
1	32479430	32526451	1	KHDRBS1
1	32716840	32751766	1	LCK
1	32757687	32799236	1	HDAC1
1	43803478	43818443	1	MPL
1	45285516	45308735	-1	PTCH2
1	51426417	51440305	1	CDKN2C
1	59041099	59043166	-1	TACSTD2
1	65298912	65432187	-1	JAK1
1	92414928	92479983	1	BRDT
1	110452864	110473614	1	CSF1
1	112025970	112106584	-1	ADORA3
1	115247090	115259515	-1	NRAS
1	120454176	120612240	-1	NOTCH2
1	150547032	150552066	-1	MCL1
1	154377669	154441926	1	IL6R
1	155158300	155162707	-1	MUC1
1	156785432	156851642	1	NTRK1
1	160709037	160724611	1	SLAMF7
1	161040785	161059389		NECTIN4
1	162601163	162757190	1	DDR2
1	165370159	165414433		RXRG
1	172628154	172636014	1	FASLG
1	179068462	179198819	-1	ABL2
1	206643791	206670223	1	IKBKE
1	218519577	218617961	1	TGFB2
1	223282748	223316624	-1	TLR5
1	226548392	226595780	-1	PARP1
1	243651535	244014381	-1	AKT3

Supplementary Table S2: List of Pediatric Cancer Druggable (continue)

chr	start	end	strand	gene
2	16080686	16087129	1	MYCN
2	25455845	25565459	-1	DNMT3A
2	29415640	30144432	-1	ALK
2	39208537	39351486	-1	SOS1
2	45878484	46415129	1	PRKCE
2	47572297	47614740	1	EPCAM
2	61704984	61765761	-1	XPO1
2	65537985	65659771	-1	SPRED2
2	69092613	69098649	-1	BMP10
2	112656056	112787138	1	MERTK
2	113587328	113594480	-1	IL1B
2	121493199	121750229	1	GLI2
2	136871919	136875735	-1	CXCR4
2	190920423	190927455	-1	MSTN
2	202899310	202903160	1	FZD7
2	204732509	204738683	1	CTLA4
2	204801471	204826300	1	ICOS
2	208627310	208634287	-1	FZD5
2	209100951	209130798	-1	IDH1
2	212240446	213403565	-1	ERBB4
2	222282747	222438922	-1	EPHA4
2	239969864	240323348	-1	HDAC4
2	242792033	242801060	-1	PDCD1
3	12328867	12475855	1	PPARG
3	12625100	12705725	-1	RAF1
3	13521224	13547916	1	HDAC11
3	30647994	30735634	1	TGFBR2
3	32993066	32997841	1	CCR4
3	38179969	38184513	1	MYD88
3	41236328	41301587	1	CTNNB1
3	46395225	46402419	1	CCR2
3	46411633	46417697	1	CCR5
3	49924435	49941299	-1	MST1R
3	53190025	53226733	1	PRKCD
3	55499743	55523973	-1	WNT5A
3	66429221	66551687	-1	LRIG1
3	89156674	89531284	1	EPHA3
3	107762145	107809872	-1	CD47
3	113995760	114029135	1	TIGIT
3	132036211	132087142	1	ACPP
3	134316643	134979309	1	EPHB1

Supplementary Table S2: List of Pediatric Cancer Druggable (continue)

chr	start	end	strand	gene
3	138066539	138124375	1	MRAS
3	138372860	138553780	-1	PIK3CB
3	142168077	142297668	-1	ATR
3	178865902	178957881	1	PIK3CA
3	187386694	187388187	-1	SST
4	843064	926161	-1	GAK
4	1795034	1810599	1	FGFR3
4	15779898	15854853	1	CD38
4	25656923	25680370	1	SLC34A2
4	55095264	55164414	1	PDGFRA
4	55524085	55606881	1	KIT
4	55944644	55991756	-1	KDR
4	84213614	84256306	-1	HPSE
4	99792835	99851788	-1	EIF4E
4	123372625	123377880	-1	IL2
4	123747863	123819391	1	FGF2
4	128802016	128820350	1	PLK4
4	153242410	153457253	-1	FBXW7
4	157681606	157892546	-1	PDGFC
4	177604689	177713881	-1	VEGFC
5	1253262	1295184	-1	TERT
5	35852797	35879705	1	IL7R
5	38845960	38945698	1	OSMR
5	67511548	67597649	1	PIK3R1
5	68530668	68573250	1	CDK7
5	86563705	86687748	1	RASA1
5	131409483	131411859	1	CSF2
5	133530025	133561833		PP2A
5	139226364	139422884	-1	NRG2
5	141000443	141016437	-1	HDAC3
5	141971743	142077617	-1	FGF1
5	149432854	149492935	-1	CSF1R
5	149493400	149535435	-1	PDGFRB
5	149781200	149792492	-1	CD74
5	156512843	156569880	-1	HAVCR2
5	156569944	156682201	1	ITK
5	176513887	176525145	1	FGFR4
5	180028506	180076624	-1	FLT4
6	30844198	30867933	1	DDR1
6	31236526	31239907	-1	HLA-C
6	32162620	32191844	-1	NOTCH4

Supplementary Table S2: List of Pediatric Cancer Druggable (continue)

chr	start	end	strand	gene
6	32936437	32949282	1	BRD2
6	33161365	33168630	-1	RXRB
6	35995488	36079013	1	MAPK14
6	36644305	36655116	1	CDKN1A
6	37137979	37143202	1	PIM1
6	41902671	42018095	-1	CCND3
6	43737921	43754224	1	VEGFA
6	44214824	44221620	1	HSP90AB1
6	82879700	82957471	-1	IBTK
6	86159809	86205500	1	NT5E
6	111981535	112194655	-1	FYN
6	114254192	114332472	-1	HDAC2
6	117609463	117747018	-1	ROS1
6	127439749	127518910	1	RSPO3
6	151977826	152450754	1	ESR1
6	166822852	167319939	-1	RPS6KA2
6	170591294	170599561	-1	DLL1
7	536895	559933	-1	PDGFA
7	18126572	19042039	1	HDAC9
7	22765503	22771621	1	IL6
7	23275586	23314727	1	GPNMB
7	41724712	41742706	-1	INHBA
7	42000548	42277469	-1	GLI3
7	55086714	55324313	1	EGFR
7	75931861	75933612	1	HSPB1
7	81328322	81399754	-1	HGF
7	89783689	89794143	1	STEAP1
7	90893783	90898123	1	FZD1
7	92234235	92465908	-1	CDK6
7	100400187	100425121	-1	EPHB4
7	106505723	106547590	1	PIK3CG
7	116312444	116438440	1	MET
7	128828713	128853386	1	SMO
7	140419127	140624564	-1	BRAF
7	148504475	148581413	-1	EZH2
7	150750899	150755617	-1	CDK5
7	151163098	151217206	-1	RHEB
7	155592680	155604967	-1	SHH
8	6357172	6420930	-1	ANGPT2
8	11351510	11422113	1	BLK
8	22877646	22926692	-1	TNFRSF10B

Supplementary Table S2: List of Pediatric Cancer Druggable (continue)

chr	start	end	strand	gene
8	23047965	23082639	-1	TNFRSF10A
8	31496902	32622548	1	NRG1
8	38268656	38326352	-1	FGFR1
8	39759794	39785963	1	IDO1
8	48685669	48872743	-1	PRKDC
8	56792372	56923940	1	LYN
8	95891998	95908906	-1	CCNE2
8	108261721	108510283	-1	ANGPT1
8	108911544	109095913	-1	RSPO2
8	128747680	128753674	1	MYC
8	141667999	142012315	-1	PTK2
9	4985033	5128183	1	JAK2
9	5450503	5470566	1	CD274
9	21967751	21995300	-1	CDKN2A
9	22002902	22009362	-1	CDKN2B
9	27109139	27230173	1	TEK
9	80331003	80646374	-1	GNAQ
9	87283466	87638505	1	NTRK2
9	91975702	92113045	-1	SEMA4D
9	93564069	93660831	1	SYK
9	98205262	98279339	-1	PTCH1
9	101866320	101916474	1	TGFBR1
9	130547958	130553066	1	CDK9
9	130577291	130617035	-1	ENG
9	133589333	133763062	1	ABL1
9	135766735	135820020	-1	TSC1
9	136895427	136933657	-1	BRD3
9	137208944	137332431	1	RXRA
9	139388896	139440314	-1	NOTCH1
9	139553308	139567130	1	EGFL7
10	6052652	6104288		IL2R
10	6469105	6622263	-1	PRKCQ
10	30722866	30750762	1	MAP3K8
10	35927177	35930362	-1	FZD8
10	43572475	43625799	1	RET
10	48413092	48416853	-1	GDF2
10	54074056	54077802	1	DKK1
10	62538089	62554610	1	CDK1
10	73507316	73533255		VSIR
10	83635070	84746935	1	NRG3
10	89622870	89731687	1	PTEN

Supplementary Table S2: List of Pediatric Cancer Druggable (continue)

chr	start	end	strand	gene
10	90750414	90775542	1	FAS
10	104263744	104393292	1	SUFU
10	123237848	123357972	-1	FGFR2
11	532242	537287	-1	HRAS
11	2150342	2170833	-1	IGF2
11	9595228	9615004	1	WEE1
11	27910385	27912580		HSP90AA2P
11	49168187	49230222	-1	FOLH1
11	60223225	60238233	1	MS4A1
11	64002010	64006259	1	VEGFB
11	66081958	66084515	-1	CD248
11	69455855	69469242	1	CCND1
11	69587797	69590171	-1	FGF4
11	69624992	69633792	-1	FGF3
11	71900602	71907345	1	FOLR1
11	103777914	104035107	-1	PDGFD
11	107992243	108018503	1	ACAT1
11	108093211	108239829	1	ATM
11	112831997	113149158	1	NCAM1
11	118307205	118397539	1	KMT2A
11	119076752	119178859	1	CBL
11	125495036	125546150	1	CHEK1
12	4382938	4414516	1	CCND2
12	6554033	6560884	1	CD27
12	6881678	6887621	1	LAG3
12	12867992	12875305	1	CDKN1B
12	14765576	14849519	-1	GUCY2C
12	25357723	25403870	-1	KRAS
12	48176505	48226915	-1	HDAC7
12	52300692	52317145	1	ACVRL1
12	52345451	52390862	1	ACVR1B
12	56137064	56150911	1	GDF11
12	56360553	56366568	1	CDK2
12	56473641	56497289	1	ERBB3
12	57853918	57866045	1	GLI1
12	58141510	58149796	-1	CDK4
12	64845660	64895888	1	TBK1
12	69201956	69239214	1	MDM2
12	102789645	102874423	-1	IGF1
12	104323885	104347423	1	HSP90B1
12	112856155	112947717	1	PTPN11

Supplementary Table S2: List of Pediatric Cancer Druggable (continue)

chr	start	end	strand	gene
12	130647004	130650285	1	FZD10
13	28577411	28674729	-1	FLT3
13	28874489	29069265	-1	FLT1
13	32889611	32973805	1	BRCA2
13	43136872	43182149	1	TNFSF11
13	86366925	86373623	-1	SLITRK6
13	108903588	108960832	1	TNFSF13B
13	114523522	114567046	-1	GAS6
14	20811741	20826064	1	PARP2
14	23767999	23780968	1	BCL2L2
14	24686058	24701660	-1	NEDD8
14	61654277	62017694	1	PRKCH
14	76424442	76449334	-1	TGFB3
14	102547075	102606036	-1	HSP90AA1
14	105235686	105262088	-1	AKT1
14	105607318	105635161	-1	JAG2
15	38544527	38649450	1	SPRED1
15	40986972	41024354	1	RAD51
15	41221538	41231237	1	DLL4
15	41849873	41871536	1	TYRO3
15	66679155	66784650	1	MAP2K1
15	73976307	74006859	1	CD276
15	76228310	76352136	-1	NRG4
15	88418230	88799999	-1	NTRK3
15	90626277	90645736	-1	IDH2
15	99192200	99507759	1	IGF1R
16	810762	818865	1	MSLN
16	2097466	2138716	1	TSC2
16	23614488	23652631	-1	PALB2
16	23688977	23701688	1	PLK1
16	23847322	24231932	1	PRKCB
16	28943260	28950667	1	CD19
16	30125426	30134827	-1	MAPK3
16	50727514	50766988	1	NOD2
16	58191811	58231824	-1	CSNK2A2
16	71671738	71758604	-1	PHLPP2
17	8108056	8113918	-1	AURKB
17	29421945	29709134	1	NF1
17	32582304	32584222	1	CCL2
17	33426811	33448541	-1	RAD51D
17	33677324	33700720	-1	SLFN11

Supplementary Table S2: List of Pediatric Cancer Druggable (continue)

chr	start	end	strand	gene
17	37617764	37721160	1	CDK12
17	37844167	37886679	1	ERBB2
17	38465444	38513094	1	RARA
17	40465342	40540586	-1	STAT3
17	41196312	41277500	-1	BRCA1
17	42154114	42201070	-1	HDAC5
17	42634925	42636907	1	FZD2
17	56429861	56494956	-1	RNF43
17	56769934	56811703	1	RAD51C
17	57970447	58027925	1	RPS6KB1
17	62006100	62009714	-1	CD79B
17	64298754	64806861	1	PRKCA
17	73314157	73401790	-1	GRB2
17	73996987	74002080	1	CDK3
17	78518619	78940171	1	RPTOR
18	721588	812547	-1	YES1
18	23596578	23671181	-1	SS18
18	60382672	60647666	1	PHLPP1
18	60790579	60987361	-1	BCL2
19	2164148	2232577	1	DOT1L
19	3094408	3124002	1	GNA11
19	4090319	4124126	-1	MAP2K2
19	6583194	6604114	-1	CD70
19	7112266	7294045	-1	INSR
19	8959520	9092018	-1	MUC16
19	10244021	10341962	-1	DNMT1
19	10461209	10491352	-1	TYK2
19	10677138	10679735	-1	CDKN2D
19	11071598	11176071	1	SMARCA4
19	15270444	15311792	-1	NOTCH3
19	15347647	15443356	-1	BRD4
19	17935589	17958880	-1	JAK3
19	30302805	30315215	1	CCNE1
19	35810164	35838258	1	CD22
19	39989535	39999121	1	DLL3
19	40736224	40791443	-1	AKT2
19	41725108	41767671	1	AXL
19	41807492	41859816	-1	TGFB1
19	42212504	42233718	1	CEACAM5
19	49838428	49846592	1	CD37
19	51728320	51747115	1	CD33

Supplementary Table S2: List of Pediatric Cancer Druggable (continue)

chr	start	end	strand	gene
19	54382444	54410906	1	PRKCG
19	55249980	55295776	1	KIR2DL3
19	55281263	55295774	1	KIR2DL1
19	55361898	55378662	1	KIR3DL2
19	57742377	57746916	1	AURKC
20	459116	524465	-1	CSNK2A1
20	10618332	10654694	-1	JAG1
20	30252255	30311792	-1	BCL2L1
20	30639991	30689659	1	HCK
20	31350191	31397162	1	DNMT3B
20	35973088	36034453	1	SRC
20	44637547	44645200	1	MMP9
20	44746911	44758502	1	CD40
20	54944445	54967393	-1	AURKA
21	39751949	40033704	-1	ERG
22	21271714	21308037	1	CRKL
22	22108789	22221970	-1	MAPK1
22	24129150	24176703	1	SMARCB1
22	24813847	24838328	1	ADORA2A
22	29083731	29138410	-1	CHEK2
22	30658818	30662829	-1	OSM
22	39619364	39640756	-1	PDGFB
22	50354161	50357728	1	PIM3
22	50683612	50689834	-1	HDAC10
22	50702142	50709196	-1	MAPK11
X	12885202	12908499	1	TLR7
X	12924739	12941288	1	TLR8
X	15482369	15574652	1	BMX
X	20168029	20285523	-1	RPS6KA3
X	47420516	47431307	1	ARAF
X	48367350	48379202	1	PORCN
X	48659784	48683392	1	HDAC6
X	48770459	48776301	-1	PIM2
X	66764465	66950461	1	AR
X	71549366	71792953	-1	HDAC8
X	83318984	83442933	-1	RPS6KA6
X	100604435	100641183	-1	BTK
X	153845865	153847533	-1	CTAG1B

Aberration	Druggable Gene
Deletion	PPARG, RAF1, HDAC11, TGFBR2, CCR4, MYD88, CTNNB1, CCR2, CCR5, MST1R, PRKCD, WNT5A, LRIG1, TERT, HSPB1, HGF, JAK2, CD274, CDKN2A, CDKN2B, TEK, GNAQ, NTRK2, SEMA4D, SYK, PTCH1, TGFBR1, CDK9, ENG, ABL1, TSC1, BRD3, RXRA, NOTCH1, EGFL7, TSC2, PALB2, PLK1, PRKCB, CD19, MAPK3, NOD2, CSNK2A2, PHLPP2
Gain	AKT3, STEAP1, FZD1, CDK6, EPHB4, PIK3CG, MET, SMO, BRAF, EZH2, CDK5, RHEB, SHH, IL2R, PRKCQ, MAP3K8, FZD8, RET, VSIR, NRG3, PTEN, FAS, SUFU, FGFR2, CCND2, CD27, LAG3, CDKN1B, GUCY2C, KRAS, HDAC7, ACVRL1, ACVR1B, GDF11, CDK2, ERBB3, GLI1, CDK4, TBK1, MDM2, IGF1, HSP90B1, PTPN11, FZD10, BCL2L1, HCK, DNMT3B, SRC, MMP9, CD40, AURKA
Amplification	MCL1, IL6R, MUC1, NTRK1, SLAMF7, NECTIN4, DDR2, DKK1, CDK1
SNVs/INDELs	NOTCH2

Supplementary Table S3: List of druggable gene found exclusively in cfDNA from a patient with Wilm tumor

	Tumor Entity	Sample (n)	Median CPA Score
Healthy.Control	Healthy.Control	10	2.14
Low ctDNA (TF<3%)	Brain tumors	39	4.38
Low ctDNA (TF<3%)	Others	9	3.46
Low ctDNA (TF<3%)	Sarcomas	33	4.23
High ctDNA	Brain tumors	1	6.68
High ctDNA	Others	1	71.6
High ctDNA	Sarcomas	11	29.3

Supplementary Table S4: CPA score per tumor entity and tumor fraction

9.3 Reproducibility

This section contain technical process in order to perfrom the bioinformatics workflow for cfDNA sequencing analysis. This section will cover all process described in the Figure 20 in the method section. The primary aim of this part is to make analysis reproducible as much as possible under the system environment of ODCF cluster. Implementation under other environment/condition must configure path of directory, software/module availability per situation. Every processing script is accessible at the github repository `https://github.com/Pitithat-pu/OE0290_pediatric_workflow`. Feel free to contact pitithat@gmail.com for questions.

9.3.1 Directory structure on the ODCF cluster environment

Project directory of cfDNA samples All sequencing data of cfDNA were transferred to and managed by DKFZ Omics IT and Data Management Core Facility (ODCF) under the project codename OE0290_pediatric_tumor. Their in-house bioinformatics workflows (Method Section 2.2) have constructed most of fundamental directory structure; namely the "project directory". The project directory host raw sequencing files (FASTQ), sequence alignment files (BAM) (except the panel-sequencing). In the future, it'd better checking if they still keep the directory structure as follows:

The overview structure of the project directory (OE0290_pediatric_tumor)

```
/omics/odcf/project/OE0290/pediatric_tumor/
 - exon_sequencing/
   - view-by-pid/
      - 'PID'/
        - 'Sample.ID'/paired/merged-alignment/
        - 'Sample.ID'/paired/run...
 - panel_sequencing/
   - view-by-pid/
      - 'PID'/
        - 'Sample.ID'/paired/run...
 - whole_genome_sequencing/
   - view-by-pid/
      - 'PID'/
        - 'Sample.ID'/paired/merged-alignment/
        - 'Sample.ID'/paired/run...
```

- merged-alignment - contains BAM files and quality control matrices
- run... - contains sequencing FASTQ files (R1,R2) and UMI sequence files (I1)

All sequencing data of tumor, control included in the pediatric cohort were also managed by ODCF under the project codename INFORM. The fundamental directory structure is similar to OE0290_pediatric_tumor. Result of both SNV Calling and INDEL calling workflow (Method Section 2.2) were also located within this structure.

The overview structure of the project directory (INFORM)

```
/omics/odcf/project/inform/sequencing/
 - exon_sequencing/
   - view-by-pid/
      - 'PID'/
        - 'Sample.ID'/paired/merged-alignment/
        - indel_results/
        - snv_results/
 - whole_genome_sequencing/
```

- merged-alignment - contains BAM files and quality control matrices
- indel_results - contains result of ODCF INDEL calling workflow
- snv_results - contains result of ODCF SNV calling workflow

9.3.2 Setting up an analysis directory

We created an directory with full file permission in a separated directory inside the ODCF cluster environment. The analysis of this study are hosted at /omics/odcf/analysis/OE0290_projects/pediatric_tumor/. If you want to host the analysis somewhere else, please adjust the path accordingly. The bash sript "**fill_PID_folders.sh**" will create symbolic links to BAM files and (if applicable) SNV and INDEL results in the project directory of exome-sequencing and low-coverage whole-genome sequencing.

The structure within this directory after running fill_PID_folders.sh

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - exon_sequencing/
   - results_per_pid/
     - 'PID'/
       - alignment/
       - indels/
       - mpileup/
 - whole_genome_sequencing/
   - results_per_pid/
     - 'PID'/
       - alignment/
```

- alignment - contains symbolic links to BAM, BAI and quality control files in the project directory
- indels - contains symbolic links to INDEL calling results
- mpileup - contains symbolic links to SNV calling results

9.3.3 Running AlignmentAndQCWorkflows for Panel Sequencing data

As you may see that ODCF did not process the basic sequence alignment for panel sequencing data. We have to manual run the AlignmentAndQCWorkflows via roddy. Bash script "**PanCanAlignment.sh**" inside the git directory panel_sequencing will run AlignmentAndQCWorkflows.

1. Create a directory RoddyConfig inside panel_sequencing directory; Copy PanCanAlignment.xml from the git repo to the RoddyConfig directory
2. Create a directory target_regions inside panel_sequencing directory; Copy target_regions/panel_target_coverage_plain.bed inside the git repo to target_regions directory
3. Edit PanCanAlignment.sh: Setting the variable PIDs to all PIDs you want to process; TUMOR_SAMPLE_NAME_PREFIXES to the sample prefix (e.g. plasma, csf, serum); TARGET_REGIONS_FILE to path of panel_target_coverage_plain.bed in the previous step.
4. Run PanCanAlignment.sh

The alignment result will be located in the directory "alignment".

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - panel_sequencing/
   - results_per_pid/
     - 'PID'/
       - alignment/
   - RoddyConfig/PanCanAlignment.xml
   - target_regions/panel_target_coverage_plain.bed
```

Remark : This process require several reference files that provide through the path /icgc/ngs_share and ODCF plugin files inside /tbi/software/x86_64/otp/roddy/. Please check availability of all paths inside the script PanCanAlignment.sh. If the script still doesn't work, please contact ODCF IT support; tell them that you want to run the AlignmentAndQCWorkflows.

9.3.4 Pre-processing - UMI workflow (fgbio workflow)

Fgbio toolkit, developed by Fulcrum genomics, provides the UMI processing workflow [126]. As described in Method Section 2.4.1 ,this workflow required sequencing FASTQ files of the paired-end reads (R1 and R2), a FASTQ file of UMI (I1), and a BAM file as inputs of the workflow. These FASTQ files located in the project directory. A set of bash scripts for the whole workflow and a wrap-up script (run_fgbioUMI_withunmapbam.sh) are available in the git repository https://github.com/Pitithat-pu/fgbio_umi.

Setting up and pre-configuring the workflow

1. Download the java library (.jar file) of the workflow from https://fulcrumgenomics.github.io/fgbio/.
2. Clone the git repo https://github.com/Pitithat-pu/fgbio_umi.
3. Inside the cloned directory, edit every file with prefix "fgbio_"; Set variable fgbio_jar to the path of the fgbio-1.x.0.jar
4. Inside the cloned directory, edit every file with prefix "picard_"; Set variable picard_jar to the path of the picard.jar located in the cloned directory

Low-coverage whole-genome sequencing (lcWGS) To run the UMI workflow, setting inside "**run_fgbioUMI_withunmapbam.sh**" have to be edited.

1. Set project_dir to view-by-pid directory inside the project directory
2. Set pids_dir to results_per_pid directory inside the analysis directory
3. Set PIDs to pids you want to perform
4. Set fgbio_workflow_dir to directory where you clone the git repository
5. It is possible to adjust cluster resource configuration per analysis step. Current setting is enough for both panel-sequencing and lcWGS. It may have to be adjust for more memory if the size of raw data increases.

The UMI result will locate in the directory "alignment_umi". The final result will be named with suffix (_realigned.bam), otherwise they are intermediate files.

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - whole_genome_sequencing/
   - results_per_pid/
      - 'PID'/
         - alignment/
         - alignment_umi/... _realigned.bam
```

Gene-panel sequencing (Panel-seq) Similar to lcWGS, the script **"run_fgbioUMI_withunmapbam.sh"** has to be edit accordingly to the panel-sequencing project and analysis directory. The UMI result will also locate in the directory "alignment_umi". The final result will be named with suffix (_realigned.bam), otherwise they are intermediate files.

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - panel_sequencing/
   - results_per_pid/
      - 'PID'/
         - alignment/
         - alignment_umi/... _realigned.bam
```

9.3.5 Extracting sequencing coverage

Low-coverage whole-genome sequencing We use the CollectWgsMetrics function of Picard toolkit to extract genomic sequencing coverage of lcWGS data (Method Section 2.4.2). A bash script "**run_picard_CollectWgsMetrics.sh**" was written to apply CollectWgsMetrics function to BAM files in directory alignment (non-umi) and alignment_umi (umi). We can use the picard library file (picard.jar) provided in our git repository to ensure the compatibility with the script. The output will exist inside directory named "stat_coverage" inside the alignment (or alignment_umi) directory.

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - whole_genome_sequencing/
   - results_per_pid/
     - 'PID'/
       - alignment/
         - stat_coverage/*_CollectWgsMetrics.txt
       - alignment_umi/
         - stat_coverage/*_CollectWgsMetrics.txt
```

The output file CollectWgsMetrics.txt contain a WgsMatrics table. In this table, we use MEAN_COVERAGE for comparing sequencing coverage between samples and PCT_EXC_DUPE is the read duplication rate of the sample.

Whole-exome sequencing We simply use the quality matrix file (*_wroteQcSummary.txt) for extracting on-target coverage (column "coverage QC bases On Target") of WES data. This file should be linked and located inside the alignment directory (Appendix Section 9.3.2).

Panel-sequencing Similar to the process for low-coverage whole-genome sequencing, A bash script "**run_picard_CollectWgsMetrics.sh**" will extract sequencing coverage of BAM file in both directory alignment (non-umi) and alignment_umi (umi). However, we must supply the function with an interval file to calculate the genomic regions that targeted by the designed gene-panel. In our git repository, we provide the interval file of gene-panel used by this study (**panel_target_coverage_plain.interval_list**). We extract the read duplication rate of the sample from column PCT_EXC_DUPE of the result file (*_CollectWgsMetrics.txt).

To compare depth of coverage between panel-seq samples, we extract median on-target depth-of-coverage from a given BAM file and a target-region file in bed format. A bash script "**run_median_ontarget_depth.sh**" apply samtools depth function and additional awk command to get the median read-depth of target-regions. We can find the bed file of target-region in target_regions/panel_target_coverage_plain.bed (Appendix Section 9.3.3). The output will exist inside directory named "stat_coverage" inside the alignment (or alignment_umi) directory.

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - panel_sequencing/
   - results_per_pid/
     - 'PID'/
       - alignment/
         - stat_coverage/*_CollectWgsMetrics.txt
         - stat_coverage/*_mediumdepth.txt
       - alignment_umi/
         - stat_coverage/*_CollectWgsMetrics.txt
         - stat_coverage/*_mediumdepth.txt
```

9.3.6 Estimating DNA oxidation artifact with picard tools

We provide a bash script "**run_CollectOxoGMetric.sh**" for exon_sequencing and panel_sequencing for estimating DNA oxidation artifacts C > A/G > T (Method Section 2.4.3). Given a BAM file, this script will run Picard CollectOxoGMetrics function. We can use the picard library file (picard.jar) provided in our git repository to ensure the compatibility with the script. The script will make a command and submit the command to bsub. The output will be in a new directory named "picard_CollectOxoGMetrics" locating in the PID directory of the given BAM file.

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - panel_sequencing/
   - results_per_pid/
      - 'PID'/
         - picard_CollectOxoGMetrics/*oxoG_metrics.txt
```

9.3.7 Copy-number variant calling - lcWGS

Selecting samples for Panel-of-Normal (PoN) creation We can skip this step if we don't want to recrate PoN. For running ichorCNA with PoN, we can jump to the next section (lcWGS - ichorCNA).

Low-coverage whole-genome sequencing with AccelNGS and Picoplex library

1. We select a group of samples to represent sample population with normally distributed genomic coverage.
2. The instruction of selection will be inside the directory "ichorCNA/NIPTeR_PoN_selection" of the cloned git repository.
3. **OE0290_ped_AccelNGS_coverage_qc.html** and **OE0290_ped_Picoplex_coverage_qc.html** are instructions for sample selection for lcWGS samples sequenced by AccelNGS and Picoplex respectively.
4. Using NIPTeR package to help you select a group of control samples, defined as sample without large copy-number aberration, from the sample population selected previously. Please follow those instruction inside **NIPTeR_OE0290_select_control.html** for AccelNGS samples and **NIPTeR_OE0290_select_control_nonumi.html** for Picoplex samples
5. The result file of the previous step (e.g. NIPT_clean_bamfiles.txt) will tell which samples can be uses as PoN for CNV calling workflow.

Creating and choosing Panel-of-Normal file (.rds) Previously we selected a group of samples for Panal-of-Normal using NIPTeR package. The result file (NIPT_clean_bamfiles.txt) contains samples without large copy-number aberrations. We can follow the instruction suggested by ichorCNA at https://github.com/broadinstitute/ichorCNA/wiki/Create-Panel-of-Normals. In brief, we have to create wig file from those selected BAM files using readCounter (https://github.com/shahcompbio/hmmcopy_utils) and save its full path into a file (wig_files.txt). In this project, we aims to detect CNV using 1MB resolution, so we set –windows 1000000. Finally, the Rscript createPanelOfNormals.R generate the PoN for ichorCNA workflow. Our script "create_PoN.sh" in the directory "ichorCNA" located in the git repository gives an example of how to run this whole step from a group of bam file.

For reproducibility of analysis, we provides three PoN files inside "**ichorCNA_PoN**" directory of the git repository.

1. **PoN_umi_1Mb_97_NIPTeR_median.rds** for analysing Accel-NGS 2S Plus DNA (UMI processed BAM)
2. **PoN_nonumi_1Mb_97_NIPTeR_median.rds** for analysing high-coverage WGS Picoplex 1 ng input xxx-0x-02...mdup.bam or xxx-0x-03...mdup.bam (2LB-098 2LB-087 2LB-065 2LB-062)
3. **PoN_1Mb_Picoplex_median.rds** for analysing lcWGS Picoplex low input xxx-0x-01...mdup.bam

Running ichorCNA We can perform ichorCNA CNV calling and tumor fraction estimation by following the instruction in the git repository of ichorCNA (`https://github.com/broadinstitute/ichorCNA/wiki`). Alternatively, we provide a bash script (**run_ichorCNA_1MB_maxCN4.sh**) for running ichorCNA as described in Method Section 2.4.4. We must adjust those path in the script. In the script, we need to change the path to the installation of ichorCNA (**PATHichorCNA**), path to readCounter binary (**readCounter_bin**) and path to PoN file (**PoN_rds_file**). The script takes two positional parameters: 1) Full path to bamfile, 2) Path to output directory.

Steps to run ichorCNA in brief:

1. Make PoN or use already created PoN as mentioned above.

2. Make sure that readCounter binary and ichorCNA is already installed. Setup paths to readCounter and ichorCNA in the bash script **(run_ichorCNA_1MB_maxCN4.sh)**

3. Select suitable PoN rds file. Set the variable **PoN_rds_file** to the location of the file.

4. Run **run_ichorCNA_1MB_maxCN4.sh.** Giving it two parameters: full path to bamfile and full path to output directory

9.3.8 Copy-number variant calling - WES

This study applied PureCN (version 1.21.3) for performing CNV calling on WES data of cfDNA. The instruction for package installation were provided by the developer at
https://bioconductor.org/packages/release/bioc/vignettes/PureCN/inst/doc/Quick.html.
In addition, we followed their recommendation by installing and using PSCBS for segmentation. For project reproducibility, we provide two wrap-up scripts (run_GC-normalized_coverage.sh and run_PureCN.sh).

The script **run_GC-normalized_coverage.sh** performs GC normalization. In this script, we have to set variable **PureCN_libdir** to the location of PureCN library; **results_per_pid_dir** to results_per_pid analysis directory; **intervals** to the location of capture kit bait interval file (our git repo provides Agilent7withoutUTRs_plain_bait.intervals) and **PIDs** to pid that we want to analyse. The script will generate and submit command to bsub. A GC-normalied coverage file (*_loess.txt.gz) is given at the end of the process.

Once the previous script is finished, the script **run_PureCN.sh** performs PureCN analysis given a coverage file (*_loess.txt.gz). There are several setting to be adjust as follows:

- module_load_cmd = (In ODCF cluster environment) the module load command for using R: must be the same R version you install the PureCN library.
- result_per_pid_dir = path to analysis results_per_pid directory
- PURECN = path to extdata/ of the installed PureCN library
- PureCN_normaldb = Path to normalDB file (PoN of this software):

  We provide normalDB_agilent_v7_hg19.rds in our git repository for Agilent V7 without UTRs capture kit. The selection criteria were in the Method Section 2.4.5.
- mappingbias_file = Path to mapping bias information file:

  We provide mapping_bias_agilent_v7_hg19.rds in our git repository for Agilent V7 without UTRs capture kit.
- internal_file = Path to the capture kit bait interval file: Agilent7withoutUTRs_plain_bait.intervals is available in out git repository.
- snp_blacklist_file = Path to excluded location of repetitive regions:

  We provide SimpleRepeat_hg19_plain.bed for hg19 genome.

The output directory (PureCN) locates in the results_per_pid directory.

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - exon_sequencing/
   - results_per_pid/
      - 'PID'/
         - PureCN/
```

9.3.9 ODCF SNV/IndelCalling workflow for cfDNA WES

We applied ODCF SNV/IndelCalling workflow for identification of somatics mutation. To perform this analysis, we must check the existance of individual-matched plasma-control BAM file in the alignment directory. Within ODCF cluster environment, this workflow can be executed through roddy command. Our git repository provide two necessary files: SNVCalling_WES.xml and applicationProperties.ini. The setting and execution instruction are as follows:

1. Create a directory to host SNVCalling_WES.xml and applicationProperties.ini. Usually, we create a directory name "RoddyConfig" at the same location as the results_per_pid. Copy these two files there.

2. Edit applicationProperties.ini; Set or add path to our "RoddyConfig" to the variable configurationDirectories

3. Edit SNVCalling_WES.xml, Set config value of inputBaseDirectory and outputBaseDirectory to the our analysis directory

4. In the SNVCalling_WES.xml, check the value of configuration (cvalue) "possibleTumorSampleNamePrefixes". The value of this configuration variable must be matching with the prefix of our cell-free DNA sample. Normally, our cell-free DNA would be named aka plasma-01-01, plasma-01-02, plasma-02-01 or etc.

5. Execute roddy command for snvCalling; replace 'PID' and '/path/to/RoddyConfig/' with sample pid, and full path of directory in 1)

   /icgc/ngs_share/ngsPipelines/RoddyStable/roddy.sh run WES_control_pediatric@snvCalling 'PID' --useconfig='/path/to/RoddyConfig/'applicationProperties.ini

6. Execute roddy command for indelCalling; replace 'PID' with sample pid

   /icgc/ngs_share/ngsPipelines/RoddyStable/roddy.sh run WES_control_pediatric@indelCalling 'PID' --useconfig='/path/to/RoddyConfig/'applicationProperties.ini

The output of the workflow will be located in the results_per_pid directory named mpileup and indels.

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - exon_sequencing/
   - results_per_pid/
     - 'PID'/
       - mpileup/
       - indels/
```

If the script still doesn't work, please contact ODCF IT support; tell them that we want to run the SNVCalling or INDELCalling workflow.

9.3.10 Tumor-informed mutation detection

Extracting on-target reads We provide a bash script "**run_extract_on-target_reads.sh**" in our git repository. This script will load and use bedtools (for extracting on-target reads) and samtools (for creating index file) given a target bed file. For panel-seq, the target file is named panel_target_coverage_plain.bed which we has already mentioned in Appendix Section 9.3.3. For whole-exome sequencing data, target files of different version of Agilent SureSelect are available through ICGC/ngs_share directory (commented in the script). This script will find all BAM files per pid and create a bsub command. The command contains the "bedtools intersect" and the "samtools index" command.

The output file will be saved into the alignment file (alignment or alignment_umi)

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - exon_sequencing/
   - results_per_pid/
     - 'PID'/
       - alignment/
         - *.on-target.bam
 - panel_sequencing/
   - results_per_pid/
     - 'PID'/
       - alignment_umi/
         - *.on-target.bam
```

Running addBAMinfo script After we extracted on-target reads, we are ready to interogate tumor mutation on those on-target reads in cfDNA. Our git repository provide several files inside the directory addBAMinfo including:

1. ***_functional_snv.config** and ***_functional_indel.config** contain configuration setting for running the process. We have to set :

   ANALYSIS_DIR = Path to analysis directory

   PIPELINE_DIR = Path to this addBAMinfo directory

   REFERENCE_GENOME = Path to reference genome (.fa); To make sure that the file exist

2. **run_addBAMinfo_per_pids_functional.sh** is the executing script. We have to set :

   result_per_pid_dir = Path to results_per_pid

   PIPELINE_DIR = Path to this addBAMinfo directory; Same as the previous .config file

   PIDs = pids to run the analysis

To execute the process, We simply execute the bash script **"run_addBAMinfo_per_pids_functional.sh"**. The script will run run_addBAMinfo.sh with necessary parameters which later submit the process to the bsub command. The result of process will be inside the results_per_pid directory named as "addAnnotation".

```
/omics/odcf/analysis/OE0290_projects/pediatric_tumor/
 - exon_sequencing/
   - results_per_pid/
      - `PID`/
         - addAnnotation
           - *_compareSOLiD_functional_indels
           - *_compareSOLiD_functional_snvs
```

9.3.11 In-silico size-selection of CfDNA

We provide a bash script "**short_isize_selection.sh**" inside directory whole_genome_sequencing of our git repository to perform the in-silico size-selection. The script accept the path to BAM file as only input parameter. Via samtools and awk command, the script extract sequencing read originated from DNA fragment with size between 50 to 150 bases. The output file will be a BAM file with suffix name "***.shortinsert.bam**". in the same directory as the input BAM.