An Integrative Analysis of ctDNA Fragmentomics and Methylation Across Breast Cancer Subtypes to Identify Subtype-Specific Prognostic Biomarkers

By

Ruhin Rahman Raida

21336025

Nafisa Kaniz

21336012

Tarunima Ferdaus

21236063

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of B.Sc. in Biotechnology

Department of Mathematics and Natural Sciences
Brac University
December 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing a degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ruhin Rahman Raida
21336025

Nafisa Kaniz
21336012

Tarunima Ferdaus
21236063

Approval

The thesis titled “An Integrative Analysis of ctDNA Fragmentomics and Methylation Across Breast Cancer Subtypes to Identify Subtype-Specific Prognostic Biomarkers” submitted by

1. Ruhin Rahman Raida (21336025)
2. Nafisa Kaniz (21336012)
3. Tarunima Ferdaus (21236063)

of Fall’2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Biotechnology on 18th of December, 2025.

Examining Committee:

Supervisor:
(Member)

Tushar Ahmed Shishir
Lecturer
Department of Mathematics and Natural Sciences
BRAC University

Co-Supervisor:
(Member)

Mahmuda Kabir
Lecturer
Department of Mathematics and Natural Sciences
BRAC University

Program Coordinator:
(Member)

Dr. Munima Haque
Associate Professor and Program Director
Biotechnology program,
Department of Mathematics and Natural Sciences,
School of Data and Sciences, BRAC University

Departmental Head:
(Chair)

Dr. Md. Firoze H. Haque
Associate Professor and Chairperson
Department of Mathematics and Natural Sciences

Abstract

In cancer patients, a substantial fraction of cell-free DNA (cfDNA), derived from tumor cells, offers a non-invasive opportunity to interrogate tumor biology. This study aimed to extract, profile and identify unique fragmentation and methylation patterns in four different breast cancer subtypes (HER2+, Luminal B, Triple Negative Breast Cancer, Triple Positive Breast Cancer) compared to healthy controls. Furthermore, it utilizes the fragmentation feature profiles along with epigenetic analysis to evaluate their potential for subtype -specific detection. Public cfDNA sequencing datasets from healthy controls and four types of breast cancer subtypes cases were retrieved from the NCBI SRA repository. Fragmentation features including Fragment Size Distribution (FSD), Fragmentation Size Ratio (FSR), Fragmentation Size Coverage (FSC), BreakPoint Motif (BPM), End Motif (EDM), Copy Number Variation (CNV) and epigenetics profiles like methylation profiles were searched using a python based workflow and downstream analyzed in R. Our findings indicated subtype specific distinctions in TNBC and TPBC, where notable differences in fragmentation and methylation patterns were observed when compared with controls and other subtypes. Among the fragmentomic features evaluated, Fragment Size Ratio showed noticeable difference for TPBC and TNBC but did not demonstrate the same level of distinction for LB or HER2+. In methylation, end motif and breakpoint motif analyses, observable patterns were also more prominent in TNBC and TPBC, while clear differentiation was not consistently evident for LB and HER2+ within this dataset. The integrated analysis of cfDNA hypermethylation and fragmentomic features reveal distinct fragmentation signatures in TPBC and TNBC, emphasizing the potential of cfDNA-based assays for precise, non-invasive molecular subtyping and the development of liquid biopsy strategies for early detection and disease monitoring.

Keywords: cfDNA, Fragmentomics, Methylation, Breast Cancer, Molecular Subtypes

Acknowledgement

We wish to express our deep gratitude to our institution, BRAC University, for enabling us to pursue this research work and for fostering an environment that encourages academic and professional growth. We are especially thankful to our thesis supervisors, Mahmuda Kabir and Tushar Ahmed Shishir, for their invaluable guidance, constructive feedback, and unwavering support, which have been instrumental in shaping the development and completion of this thesis.

We humbly express our gratitude to Almighty Allah, whose mercy, strength, and wisdom enabled us to complete this work. We are deeply thankful to our parents for their constant encouragement, unconditional support, and belief in us throughout every phase of our lives.

Sincerely,

Ruhin Rahman Raida (21336025)

Nafisa Kaniz (21336012)

Tarunima Ferdaus (21236063)

Department of Mathematics and Natural Sciences,

BRAC University

Table of Contents

List of Tables

List of Figures

List of Acronyms

BPM	Breakpoint Motif
BRCA1	Breast Cancer Gene 1
cfDNA	Cell-free DNA
cfDNAFE	Cell-free DNA Feature Extraction
ctDNA	Circulating Tumour DNA
DCIS	DuctalCarcinoma In SItu
DNA	Deoxyribonucleic Acid
DNMT	DNA Methytransfarase
EDM	End Motif
ER	Estrogen Receptor
FSC	Fragment Size Coverage
FSD	Fragment Size Distribution
FSR	Fragment Size Ratio
HDI	Human Development Index
HER2	Human Epidermal Growth Factor Receptor 2
IDC	Invasive Ductal Carcinoma
ILC	Invasive Lobular Carcinoma
LB	Luminal B
LCIS	Lobular Carcinoma In Situ
LINE	Long Interspersed Nuclear Element

MRD	Minimal Residual Disease
NGS	Next-Generation Sequencing
PR	Progesterone Receptor
SAH	S-adenosylhomocysteine
SRA	Sequence Read Archive
TNBC	Triple Negative Breast Cancer
TPBC	Triple Positive Breast Cancer
WGS	Whole Genome Sequencing

Chapter 1

1.1 Introduction

Breast cancer is the most common cancer that women are diagnosed with worldwide, with millions of cases being reported each year [1, 2, 3]. While women are mainly affected by the disease, in the world ranking it holds the second highest position for cancer diagnosis in both male and female populations. In fact, among all cancer-related deaths in women, breast cancer ranks the highest [4]. Furthermore, hundreds of thousands of breast cancer-related deaths are reported globally each year.

Current diagnostic approaches for breast cancer mainly rely on imaging techniques and tissue biopsy. Although widely used, these methods have limitations, including invasiveness, patient discomfort and the risk of sampling errors due to tumor heterogeneity, as a single biopsy may not fully represent the entire tumor. These challenges highlight the need for alternative methods that are less invasive and capable of capturing real-time tumor dynamics more accurately. In response to these challenges in diagnosis and disease monitoring, researchers have increasingly explored non-invasive biomarkers. One such promising source is cell-free DNA (cfDNA), which refers to small DNA fragments released into the bloodstream from normal and tumor cells. When a tumor grows or breaks down, it sheds DNA into blood, which can be detected without any surgical procedure. This forms the basis of liquid biopsy, where blood samples are used instead of tissue biopsy to study cancer-related changes. cfDNA can carry tumor-specific mutations, methylation changes, and other molecular features that reflect the condition of the cancer. Because it is non-invasive, it offers a safer and easier method for repeated testing, early monitoring, and tracking treatment response. Research has shown that cfDNA levels often decrease after surgery or chemotherapy, and rising levels may indicate progression or lack of treatment response. Clinical trials have reported promising results, showing cfDNA as a potential tool for early detection, prognosis, therapy selection, and follow-up. Although more standardization is still needed before routine use, cfDNA analysis

has emerged as a strong candidate for improving cancer diagnosis and patient management. [5, 6].

Among the molecular alterations detectable through cfDNA, DNA methylation has become a major focus in recent research. DNA methylation is one of the main epigenetic processes that directly alters DNA chemically especially in CpG island [7]. CpG Island (CGI) hypermethylation in gene promoters is an early and progressive event in breast cancer. In normal cells, CGIs remain unmethylated to permit gene expression, but in cancer they become aberrantly hypemethylated, silencing critical genes. This epigenetic inactivation contributes directly to carcinogenesis by promoting uncontrolled growth and genetic instability [8,9,10]. cfDNA methylation is essential for the management of breast cancer as it is emerging as a critical non-invasive biomarker [11].

Building on these developments, this study focuses on understanding whether cfDNA can reflect subtype specific characteristics in breast cancer. Since each subtype shows differences in biology and clinical behavior, the work explores if cfDNA methylation patterns along with fragmentomic features such as BreakPoint Motif (BPM), End Motif (EDM), Fragment Size Distribution (FSD), Fragment Size Coverage (FSC), Fragment Size Ratio (FSR), and Copy Number Variation (CNV) show measurable variation among four subtypes. By examining these features from blood derived cfDNA, this study observes whether distinct patterns are present that support better subtype differentiation and contribute to early and non-invasive assessment in clinical settings. Understanding these differences may provide useful information for improving patient classification and disease monitoring without repeated invasive procedures.

1.2 Aim of the Study

1. Extraction of breast cancer cfDNA fragmentation and methylation profiles.
2. Identification of unique fragmentation and methylation patterns in four different subtypes of breast cancer, including HER2+, Luminal B, Triple Negative Breast Cancer, and Triple Positive Breast Cancer.
3. Utilizing fragmentation feature profiles along with epigenetic analysis and motif patterns to distinguish between various subtypes of breast cancer and healthy controls.
4. Evaluate the potential of the extracted features to act as a prognostic biomarker.

.

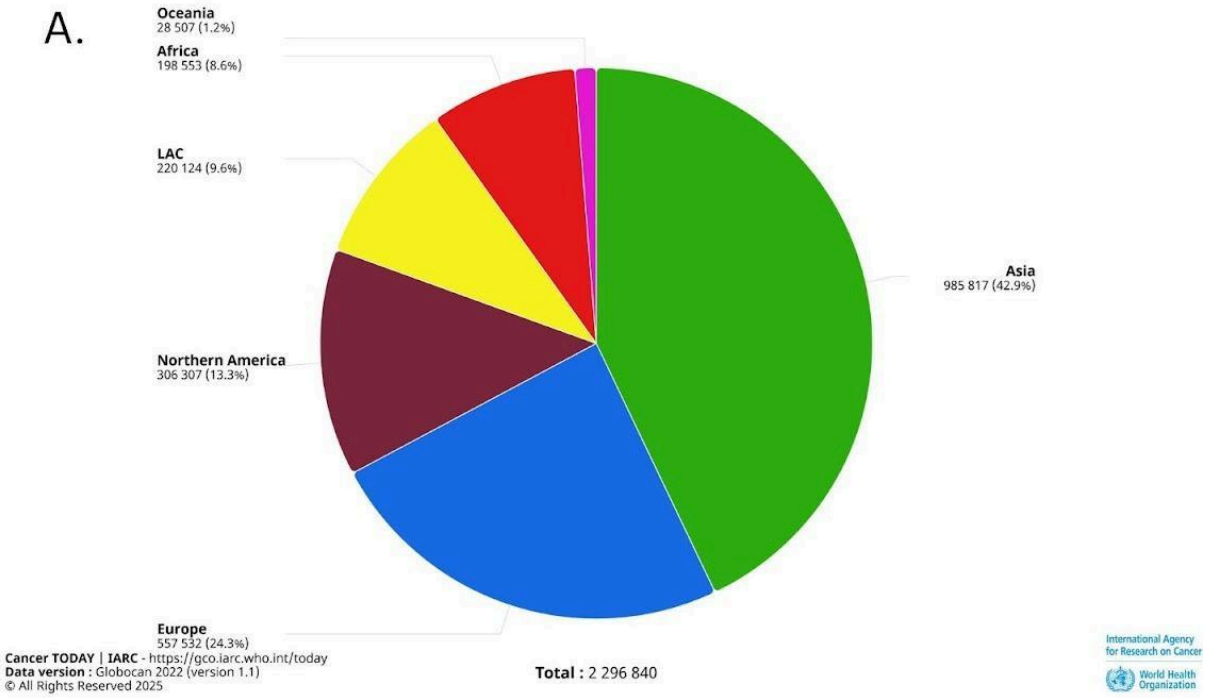
Chapter 2

Literature Review

2.1 Epidemiology

Globally, breast cancer is the most frequently diagnosed malignant tumour among the female population, an estimated 2.3 million incidences being reported among women by the World Health Organization in 2022 [1, 2, 3]. Although primarily affecting women, it ranks second for being the most commonly diagnosed cancer across both sexes and is the leading cause of cancer-related deaths in females [4]. In fact, it was responsible for approximately 670,000 deaths in women worldwide for the year 2022 [2, 3].

While incidences of breast cancer are higher in high-HDI countries, low-HDI countries shoulder the burden of higher mortality rates, possibly due to limited access to early detection and treatment options. A distinct trend can be observed where the mortality: incidence (M:I) ratio increases with decreasing HDI levels [3]. The number of new cases were highest in Australia and New Zealand followed by nations in the Northern ends of America and Europe but lowest in South-central Asia, Middle Africa and Eastern Africa. On the contrary, the highest number of deaths were recorded in some regions of Africa and Oceania whereas the lowest were observed in East Asia and some regions of America. Major causes of this discrepancy in incidence and mortality rates include differences in socio-economic structure and lifestyle choices, as shown in Table 2.1.



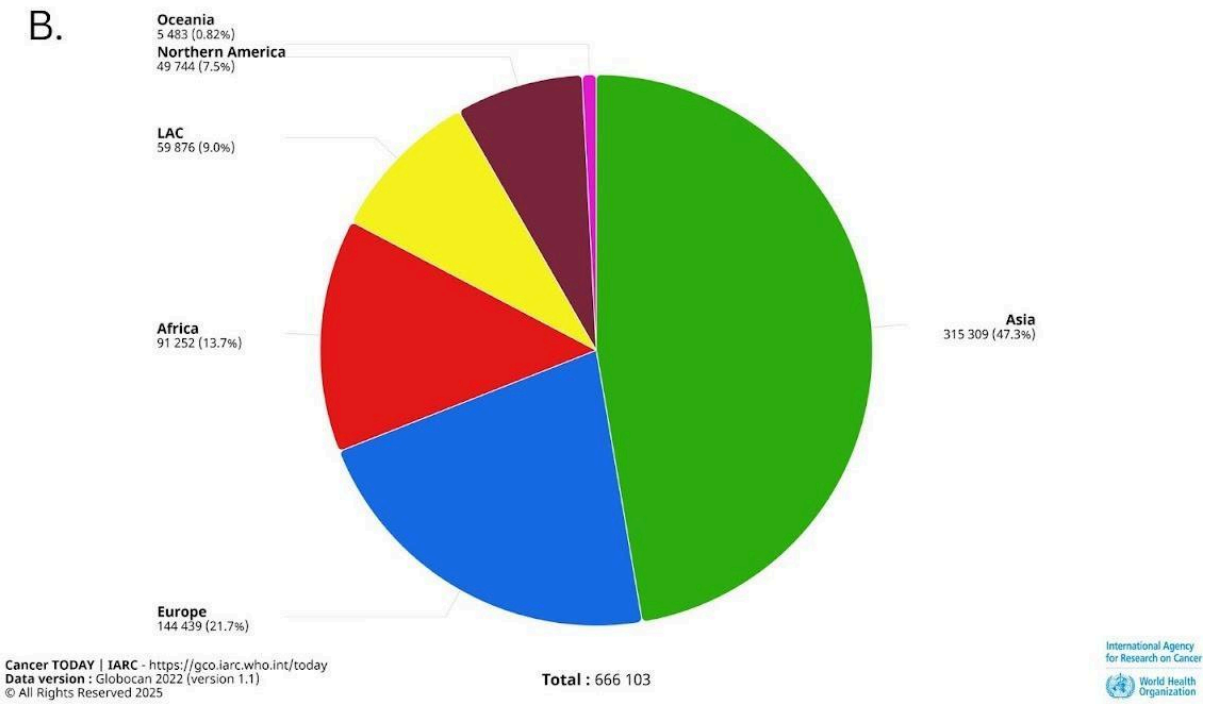


Figure 2.1 : Absolute number of incidence and mortality in females in different continents. (A) Incidence (B) Mortality [5].

KEY CONTRIBUTORS TO INCIDENCE AND MORTALITY	
HIGH INCIDENCE RATES	**HIGH MORTALITY RATES**
Hormonal Contraception	Lower Rates of Early Screening
Higher Alcohol Consumption	Limited Number of Medical Professionals
Little to No Pregnancies	Limited Access to Therapeutic Treatments
Late Pregnancy	Short Supply of Life-Saving Medicine
Less Breastfeeding	Low Awareness
Sedentary Lifestyle	Financial Hardships

Table 2.1: Factors contributing to higher breast cancer incidences in high-HDI countries and higher mortality rates in low-HDI countries.

2.2 Tumour Progression

Breast cancer develops as the accumulation of several processes involving genetics, epigenetics and tumour microenvironment (TME) formation, resulting in neoplasia and eventually, cancer [13, 14]. Genomic changes initiate the carcinogenesis process, influenced by several risk factors such as changes in environment or sex hormones, leading to the formation and proliferation of abnormal cells [15]. As lesions develop further, they accumulate the ability to evade immune cells through activation of immunosuppressor genes [16].

Lesion Type	Frequency (%)	Invasive Nature	Origin	Prognosis	Ref.
DCIS	15-25	No	Milk Ducts	Good outcome with local treatment and high survival chance.	17,18
LCIS	3	No	Milk Glands (Lobules)	Good outcome. Acts only as a risk marker but not malignant itself.	19,20
IDC	80	Yes	Milk Ducts	Varies depending on the molecular subtype	21,22
ILC	10-15	Yes	Milk Glands (Lobules)	Good prognosis with favourable outcomes in a decade of follow-up	23,24

Table 2.2 : Histopathological Classification of Breast Lesions

Classification of cancer cells based on their histopathological and molecular characteristics is an important first step in determining their identity, prognosis, and treatment plans. The change from in situ carcinoma, where abnormal cells are contained within the basement

membrane, to acquiring invasive properties, where they break out and move into the surrounding tissue, marks an increase in cancer aggressiveness and metastatic properties [25,26]. Non-invasive breast cancer is divided into two types which are DCIS (ductal carcinoma in situ) and LCIS (lobular carcinoma in situ). While there are numerous types of invasive breast cancer, the most common types are IDC (invasive ductal carcinoma) and ILC (invasive lobular carcinoma) [27,28].

2.3 Subtypes of Breast Cancer

Based on the expression levels of specific hormone receptors, molecular subtypes of breast cancer can be divided into 5 categories; Luminal A, Luminal B, HER2-positive (Human epidermal growth factor receptor 2 positive), TNBC (Triple Negative Breast Cancer) and TPBC (Triple Positive Breast Cancer a.k.a. Luminal B HER2-positive) [29]. These subtypes are categorized according to the absence or presence of estrogen receptor (ER+/-), progesterone receptor (PR+/-) and human epidermal growth factor receptor (HER2+/-). About 7–75% of invasive breast carcinomas have overexpression of ER while 50% of all ER+ tumour cells show increased expression of PR. Additionally, in about 15-30% of breast cancer, HER-2 gene is overexpressed [30].

Luminal A starts off at the inner lining of the mammary glands. The cells in this type tend to be ER+ and/or PR+ and HER2-. They usually have a low grade and have the best prognosis among the 4 subtypes with less chances of relapse and higher survival rates [29]. Similar to Luminal A, Luminal B originates at the inner lining of the mammary glands. However, it may be PR-negative and either HER2-positive or HER2-negative. Moreover, although ER expression is evident in Luminal B, it is usually low. It has a higher grade of tumour compared to Luminal A and thus a worse prognosis, higher recurrence rate and lower chance of survival. In a study conducted in 2018 it was found that Luminal B subtypes showed heightened risks

of bone metastasis compared to luminal A. Moreover, both types showed higher rates of liver, brain, and lung metastasis [31].

Subtype	Frequency (%)	ER	PR	HER2	ki67	Prognosis	Reference
Luminal A	30-40	+	+	-	Low	Favourable	32-36
Luminal B	20-30	+	+/-	+/-	High	Moderate	
HER2	10-15	-	-	+	High	Poor	
TNBC	10	-	-	-	High	Poor	
TPBC	5-10	+	+	+	Moderate	Moderate	

Table 2.3: Characteristics of Different Subtypes of Breast Cancer

HER2-positive, as its name entails, has an amplified expression of the HER2 gene. Compared to luminal subtypes, they are fast growing and therefore more aggressive with worse prognosis. Lastly, TNBC subtype is negative for all 3 receptors and the tumours are more aggressive than other subtypes. It is mostly common in women below 40 years of age and of African-American descent. Due to its aggressive nature, it has poor prognosis and early relapse after treatment, showing increased rates of progressing into higher stages of cancer. Likewise, HER2+ and TNBC present the worst survival rates among the subtypes with local and regional recurrence being 7.5 and 3.4% for HER2+ and 7.6 and 3.3% for TNBC [37].

2.4 Stages of Breast Cancer

In breast cancer, the stage of the tumour determines the treatment plan and prognosis. Initially, stage is determined for breast cancer diagnosis which can be done through several methods such as physical examination, mammography, ultrasound and tissue biopsy. It is an indicator of the size of the tumour and whether it has spread to other parts of the body from the primary location [72]. The stages are determined through TNM staging, tumour grade and the presence or absence of specific biomarkers.

Stage	Subtage	Tumour Size	Spread to Lymph Node	Metastasis	Invasive
0	-	-	No	No	No
I	IA	-	No	No	Yes
	IB	Less than 2cm (T0/T1/T2/T3)	Yes	No	Yes
II	IIA	Less than 2cm	Yes (1-2 lymph nodes affected)	No	Yes
		2-5 cm	No	No	Yes
	IIB	2-5 cm	Yes (1-3 lymph nodes affected)	No	Yes
		Larger than 5cm	No	No	Yes
III	IIIA	Any Size	Yes (4-9 lymph nodes affected)	No	Yes
		Larger than 5cm	Yes (1-3 lymph nodes affected)	No	Yes
	IIIB	Any Size	Yes (Up to 9 lymph nodes affected)	No	Yes (Swelling of chest walls)
	IIIC	Any Size	Yes (10 or more lymph nodes affected)	No	Yes
IV	-	Any Size	Yes (Spread to other organs)	Yes	Yes

Table 2.4 : Stages of Breast Cancer

The stages of breast cancer are classified as Stages 0, I, II, III and IV. The factors considered in the TNM staging process are T (Tumour Size), N - (Number of nearby lymph nodes where the cancer has spread), M (Metastasis state of the tumour). The grading is determined by the size and shape of the nuclei of cells in and around the tumour and the rate at which the cells are diving. Finally, the presence of the biomarkers ER, PR and HER2 are tested. [38, 39].

2.5 Historical Background and Discovery of cfDNA:

The existence of cell-free DNA(cfDNA) in our blood was first discovered by Mandel and Metais in 1948. Almost two decades later, it was hypothesized by Bendich and his colleagues that tumor-derived cfDNA may have a role in metastasis. Finally, in 1966, Tan and colleagues discovered the first link of circulating cell-free DNA to disease. Eleven years later, in 1977, Leon and colleagues utilized radioimmunochemistry to establish that for at least half of cancer patients the level of cfDNA in their bloodstream was much higher than in normal control subjects. They further noted that patients with metastatic cancer exhibited significantly higher cfDNA in blood. It took additional 12 years for the first experimental proof to support that cfDNA in cancer patients contain tumor DNA based on temperature stability measurements. [40]

2.6 Structural Characteristics of cfDNA

Cell-free DNA in the bloodstream mainly originates from normal cellular turnover, particularly apoptosis and necrosis. In healthy individuals, cfDNA fragments are typically short, double-stranded pieces of about 150–170 bp. This reflects the underlying chromatin organization inside the nucleus, where ~147 bp of DNA wrap around a histone octamer to form a nucleosome, and adjacent nucleosomes are joined by short linker DNA segments [41].

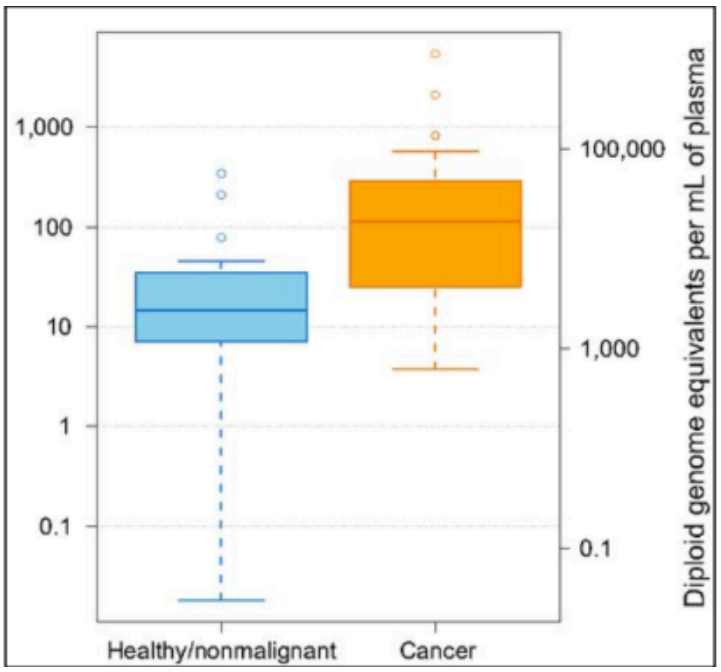


Figure 2.2 : Plasma cfDNA concentrations in healthy individuals and cancer patients, showing markedly higher cfDNA levels in the cancer group

During apoptosis, endonucleases cleave DNA precisely at these linker regions, generating fragments that correspond to one or more nucleosomal units. As a result, cfDNA exhibits characteristic fragment-size peaks at approximately ~167 bp (mononucleosomal), ~320 bp (dinucleosomal), and ~480 bp (trinucleosomal) [42].

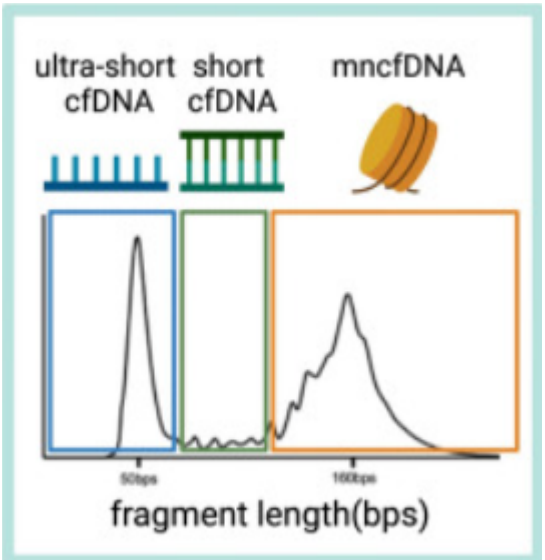


Figure 2.3: Major cfDNA fragment length categories, including ultra-short, short, and mononucleosomal fragments

2.7 ctDNA as a Tumor-Derived Subfraction

In cancer patients, a significant fraction of the total cfDNA originates from tumor cells, known as circulating tumor DNA (ctDNA), making them a subgroup of the total cfDNA. One of the key factors that sets apart the ctDNA from the normal cfDNA is the unique fragmentation pattern, revealing the condition of the biological origin of the ctDNA. This subgroup carries tumor-specific genetic and epigenetic alterations and shows distinct fragmentation behavior. One key difference is that ctDNA fragments tend to be shorter than cfDNA released from healthy cells, likely due to more chaotic and irregular DNA cleavage occurring within the tumor microenvironment [43].

In general cases, cfDNA arising from apoptosis mimics the usual length of DNA inside the healthy cells, consisting of a nucleosome along with some linker DNA to connect with the next nucleosome. These fragments are similar to the normal cells both in size and shape, demonstrating a regular chromatin structure throughout the body, reflecting non-malignancy. In contrast, ctDNA in cancer patients tends to be shorter.

Biological characteristics of cfDNA in healthy vs cancer individuals:

Feature	**Healthy individuals**	**Cancer patients**
Fragmentation Origin	Mainly apoptosis of blood and normal cells	Mix of apoptosis + necrosis; tumor necrosis produces longer, degraded fragments
cfDNA integrity	High integrity	Decreased cfDNA integrity
Fragment Length	~150 bp	Often shorter than 150 bp; irregular and more

		heterogeneous fragments
Genomic Content	No tumor specific genomic alterations	Contains tumor-derived ctDNA
Proportion of low-molecular-weight DNA	Low; most fragments cluster around ~150–180 bp	Very high; up to **90–95%** of cfDNA can be low-molecular-weight (<150 bp)

Table 2.5: Biological characteristics of cfDNA in healthy vs cancer individuals

2.8 Clinical Applications of cfDNA

Optimizing treatment for cancer requires molecular profiling of tumors. In recent years, next-generation sequencing (NGS) techniques have been used to determine treatments based on the exact genomic profile of a particular tumor. Existing clinical approaches generally involve sequencing previous primary tumor tissues collected through surgery or biopsy. However, this is an invasive procedure and might not provide sufficient samples for several metastatic diseases and all organs. Additionally, single-site tissue samples do not take genetic heterogeneity into account, which is a significant drawback because primary tumors and metastases frequently have different genetic profiles.

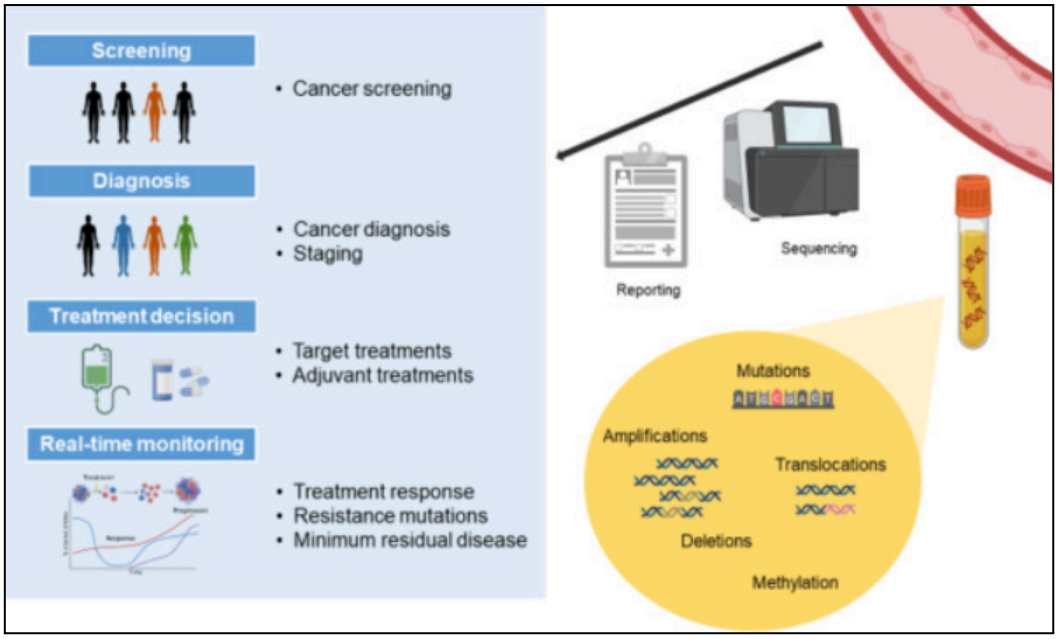


Figure 2.4 : Clinical applications of genome-wide cfDNA fragmentation in cancer patients

Tumor cells also develop genetic modifications related to resistance during therapy because of clonal evolution. To overcome these issues, liquid biopsy, specifically circulating tumor DNA (ctDNA) analysis, has been extensively investigated. It was discovered in the 1990s that "cell-free DNA" (cfDNA) circulating in the plasma of patients with malignant tumors included tumor-derived genetic changes. Circulating nucleic acids can be detected in the blood, urine, and cerebrospinal fluid, and their diagnostic assessment in liquid specimens is commonly referred to as liquid biopsy. Furthermore, the cfDNA-based liquid biopsy assay has been utilized in clinical practice to detect genetic modifications for targeted therapy [5] and for possible clinical uses in cancer screening, diagnosis, detection of minimal residual disease (MRD), and real-time tracking of tumor evolution and resistance [44].

2.9 Epigenetic Regulation Through DNA Methylation in CpG-rich Regions

Alterations in the chromosomes in the gene expression rather than in the DNA sequence that focus on the heritable and stable changes are known as epigenetics [45]. Virtually all cells in

an organism contain the same genetic information, but all cells are not expressed simultaneously. In a wider perspective, different gene expression profiles in different cells and organs within multicellular organisms are mediated by epigenetic processes [46.]

DNA methylation is one of the main epigenetic processes that directly alters DNA chemically. A substance called a methyl group is added to DNA in order for DNA methylation to occur. Demethylation is another method of removing this chemical from DNA. Genes are often turned "off" by methylation and "on" by demethylation. One method the body uses to regulate gene expression is DNA methylation. Methylation and demethylation aid in determining the amount of protein produced, but they do not alter the DNA code, or the base sequence [47]. Most DNA methylation is essential for normal development, and it plays a very important role in a number of key processes including genomic imprinting, X-chromosome inactivation, and suppression of repetitive element transcription and transposition and, when dysregulated, contributes to diseases like cancer [48,49]

A CpG island is a stretch of DNA with a high frequency of the "CpG" dinucleotide (a cytosine followed by a guanine), typically found in the regulatory regions of genes, such as promoter. [50]. As methylcytosine is mutagenic, vertebrate genomes are mostly methylated at the dinucleotide CpG, making them CpG-deficien.[51]. However, CpG islands (CGIs), which are typically 1000 base pairs (bp) long and exhibit an enhanced G+C base composition, limited CpG depletion, and frequent lack of DNA methylation, punctuate the globally methylated, CpG-poor genomic landscape. Despite the diversity of their individual nucleotide sequences, CGIs have been separated as a rather uniform section of the genome because of their fundamental characteristics (52,53,54).

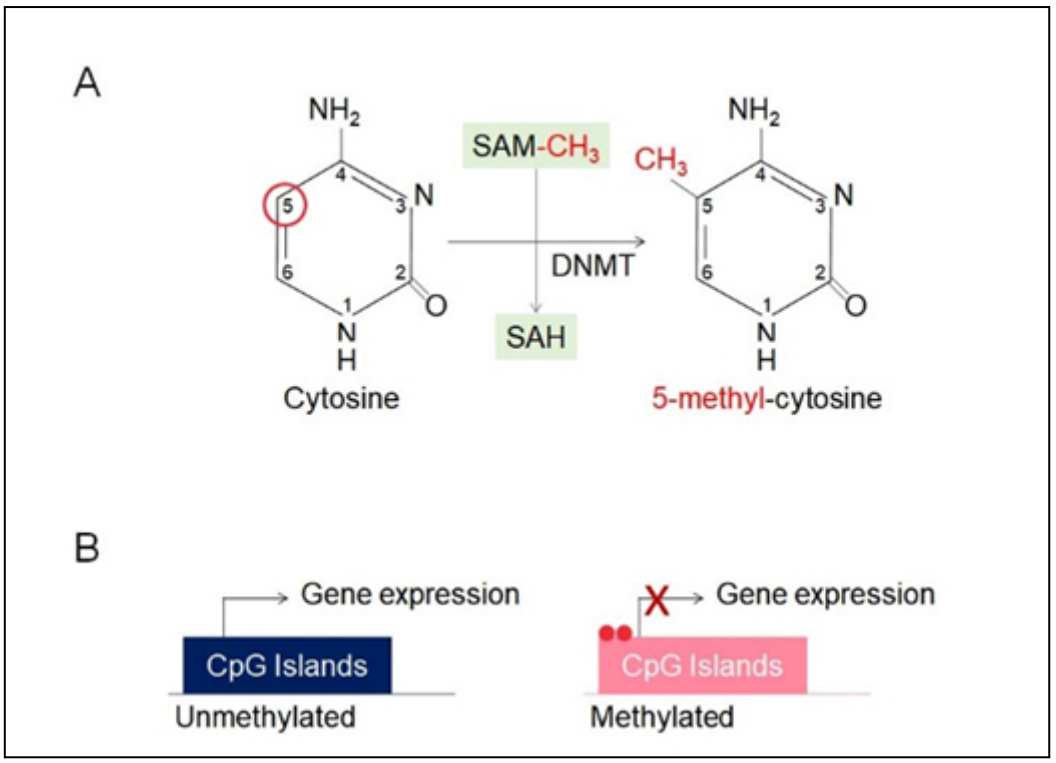


Figure 2.5 : DNA methylation and CpG island regulation. (A) DNA methylation involves the transfer of a methyl group (-CH_3) from S-adenosylmethionine (SAM) to the 5th carbon of cytosine within CpG dinucleotides, catalyzed by DNA methyltransferases (DNMTs), producing 5-methylcytosine (5-mC) and S-adenosylhomocysteine (SAH). (B) CpG island methylation regulates gene expression by silencing transcription. Unmethylated CpG islands allow active transcription, while methylated CpG islands block transcription by preventing access of transcriptional machinery. [55]

2.10 Significance of CpG islands in cancer

Methylation changes in CpG islands (CGI) affect gene expression. In addition to such local changes, global hypomethylation and chromosomal instabilities have been found in cancers [7]. Altered methylation is thought to be an early event in breast cancer and the number of genes affected increases with progression. CpG islands (CGIs) are typically located in the promoter regions of genes and are usually unmethylated in normal cells to allow gene

expression. However, in cancer, these islands frequently undergo aberrant hypermethylation, particularly in tumor suppressor gene promoters, leading to transcriptional silencing without altering the underlying DNA sequence. This epigenetic inactivation represents a key mechanism in carcinogenesis, contributing to uncontrolled cellular proliferation, loss of apoptosis, and genomic instability. Several studies have demonstrated the significance of CpG island hypermethylation in breast cancer [8,9,10]

2.11 Aberrant Methylation in Breast Cancer

Epigenetic deregulation is a hallmark of breast cancer and among these alterations, abnormal DNA methylation, both global hypomethylation and promoter-specific hypermethylation plays a central role in tumour development, progression and molecular subtype differentiation. Indeed it is well established that two kinds of changes in the DNA methylation pattern occur in breast cancer, regional hypermethylation of certain genes [57] and global hypomethylation [58]

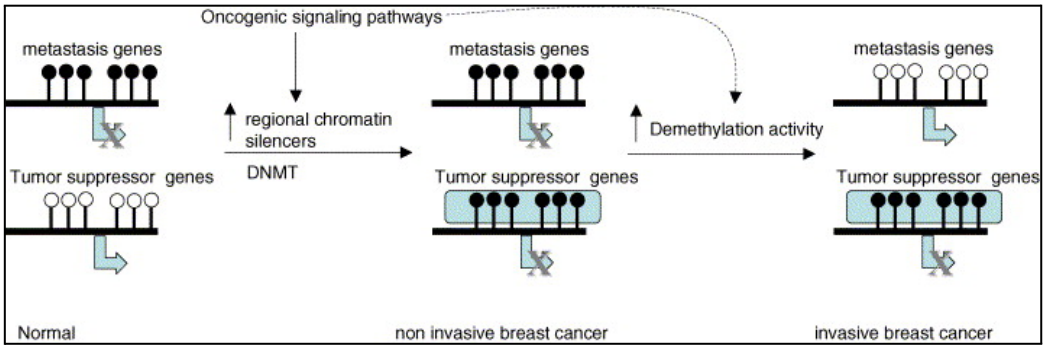


Figure 2.6 : Hypermethylation and hypomethylation in breast cancer. In normal epithelial breast cells tumor suppressor genes are active (transcription indicated by horizontal arrow) and unmethylated (open circles indicate unmethylated CGs). Metastatic genes are methylated and inactive (closed circles). Activation of oncogenic pathways leads to interaction of specific repressors with certain tumor suppressors leading to chromatin inactivation and DNA methylation (closed circles). At a more advanced stage the same or additional oncogenic

pathways lead to induction of demethylase activity resulting in global demethylation as well as demethylation and activation of genes required for metastasis leading to a highly metastatic phenotype. Tumor suppressor genes are protected from demethylation by regional repressors maintaining an inactive chromatin structure [59].

Feature	DNA Hypomethylation	DNA Hypermethylation
Definition	A significant decrease in the level of methyl groups CH3 attached to DNA, relative to a normal cell.	An abnormal increase in the level of methyl groups CH3 attached to specific DNA regions.
Genomic Location	Primarily occurs at global regions of the genome, such as highly repetitive sequences (e.g., LINE-1 and Alu elements).	Primarily occurs at locus-specific regions, specifically at CpG islands located in gene promoter regions.
Effect on Gene	Leads to the re-activation/overexpression of genes and DNA sequences that should be silent.	Leads to gene silencing (turning off) and the suppression of transcription.
Result in Cancer	Contributes to genomic instability, chromosomal abnormalities, and the activation of oncogenes or transposable elements.	Silences tumor suppressor genes (e.g., p16, BRCA1), removing key brakes on cell growth and DNA repair.

Table 2.6 : Brief difference between DNA hypomethylation and DNA hypermethylation

2.12 Global Methylation Changes in Breast Cancer

Global methylation changes refer to the overall, genome-wide level of DNA methylation, an epigenetic process that adds methyl groups CH3 to DNA bases, typically cytosines in CpG dinucleotides. These changes are crucial because they affect genomic stability and are a hallmark of various diseases. For instance, in aging and carcinogenesis, a consistent and widespread change is global hypomethylation (an overall decrease in methylation), which destabilizes the genome by activating normally silent repetitive elements (like LINE-1), often occurring alongside specific promoter hypermethylation (increased methylation) that silences critical tumor suppressor genes and drives malignancy [60].

2.13 Subtype-specific Methylation

Different types of subtypes such as HER2+, luminal A, Luminal B, TPBC, TNBC exhibit a wide range of patterns in methylation. Studies have identified DNA methylation signatures are linked to different breast cancer subtypes such as luminal A and B tumors demonstrating increased promoter CpG island hypermethylation [61]. When compared to other forms of breast cancer, epigenetic changes in TNBC show unique methylation patterns that can inhibit the expression of tumor suppressor genes and encourage the development and spread of cancer [61,62]. In contrast to other breast cancer subclasses where promoter hypermethylation events were less common, methylation analysis of primary breast cancers revealed widespread promoter hyper-methylation of epigenetic biomarker genes among triple-negative breast cancers. Triple-negative breast tumors frequently exhibit aberrant DNA hypermethylation [63] Overall, methylation-based subtyping could serve as a valuable prognostic and therapeutic tool in breast cancer management.

2.14 Role of cfDNA Methylation

cfDNA methylation is essential for the management of breast cancer as it is emerging as a critical non-invasive biomarker. It provides significant potential in liquid biopsy as it can be easily obtained from a minimal blood draw and this also mitigates the possibility of radiation exposure [64]. Moreover, aberrant DNA methylation patterns often occur early in tumorigenesis which makes them excellent candidates for early cancer detection and most treatable stages [64]. In case of prognosis and recurrence monitoring, the methylation status of certain genes in cfDNA has been associated with poor overall survival and may help predict the likelihood of cancer metastasis or recurring after the treatment [65]. Different methylation signatures are depicted among different breast cancer subtypes and identifying these specific patterns in cfDNA can help stratify patients and guide the selection of the most effective therapeutic treatments [64]. Hence DNA methylation proves to be one of the most effective and promising tools for improving the early detection system and treatment for breast cancer patients.

Chapter 3

Rationale of the Study

Current breast cancer diagnosis relies mostly on invasive tissue biopsies, which have few limitations such as patient discomfort and sampling errors due to tumor heterogeneity. This makes non-invasive alternatives essential. In particular, liquid biopsy has shown promising results in addressing these issues by analyzing cell-free DNA (cfDNA). Tumor derived cfDNA carries a broad spectrum of tumor specific alterations such as fragmentation patterns, copy number variations and epigenetic modifications such as DNA methylation. Among these aberrant CpG island hypermethylation is a hallmark of breast tumorigenesis and holds strong potential as a sensitive biomarker. In parallel, fragmentomic features such as breakpoint motif, end motif, fragment size distribution, fragment size coverage, fragment size ratio provides additional layers of biological information reflecting chromatin structure, nucleosome positioning and genomic instability. Despite rapid advances, there is still limited understanding of how cfDNA methylation and fragmentation vary across different molecular subtypes of breast cancer. Finding subtype-specific cfDNA signatures could greatly enhance early identification, risk stratification and individualized monitoring because each subtype - HER2-positive, Luminal B, Triple Negative, and Triple Positive has unique biological traits and clinical outcomes. A non-invasive method capable of detecting these subtypes would reduce reliance on repeated biopsies, improve patient comfort and support real time clinical decision making.

This study is therefore grounded in the necessity to investigate cfDNA as a complete prognostic biomarker source by integrating both methylation and fragmentation profiles in order to ascertain whether these characteristics can distinguish between breast cancer subtypes. Determining such subtype-specific cfDNA patterns may improve individualized treatment of breast cancer patients and lead to more accurate, minimally invasive diagnostic techniques.

Chapter 4

Methodology

4.1 Workflow Overview

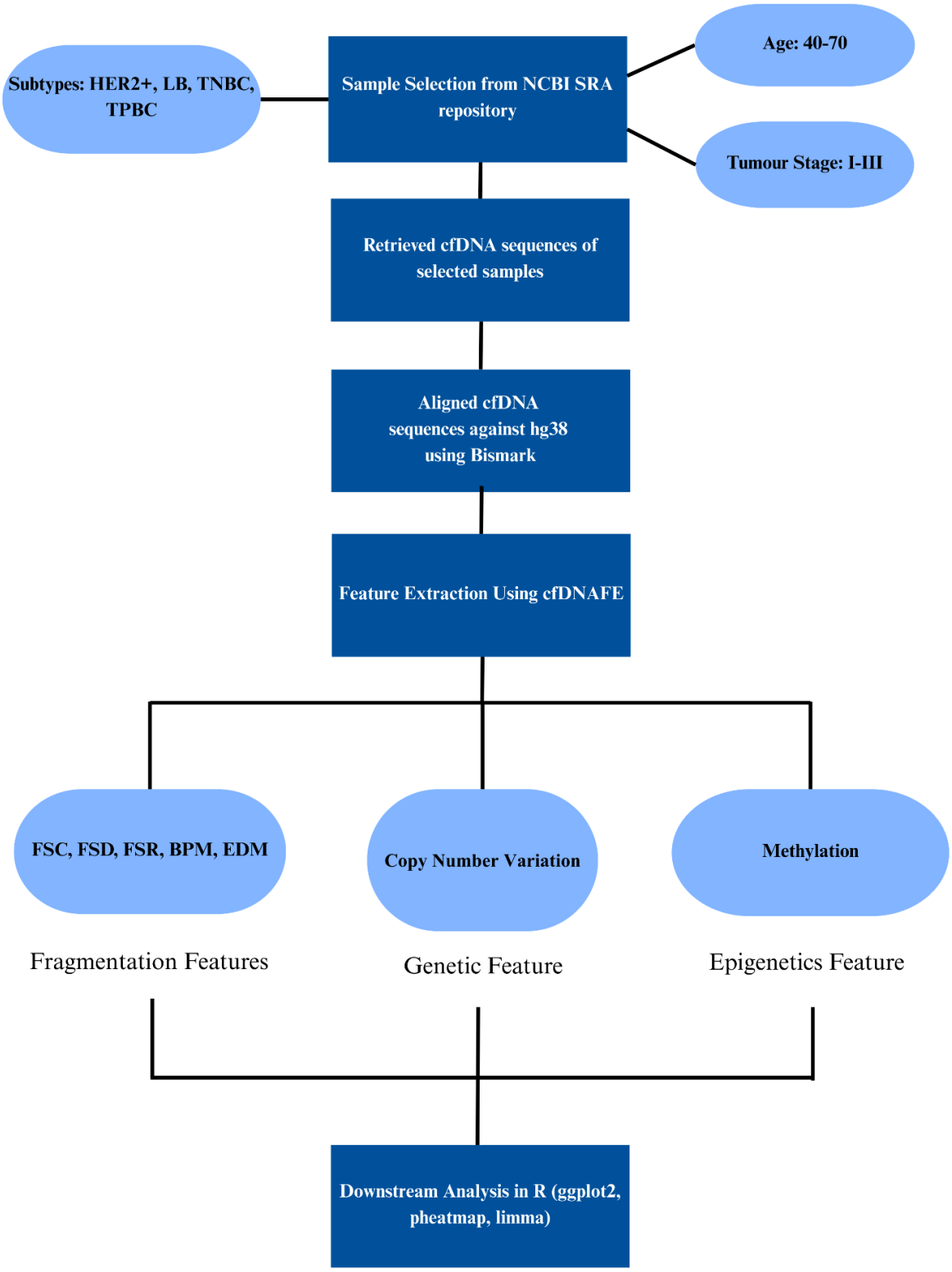


Figure 4.1: Computational and Analytical Workflow of the Study

4.2 Sample Selection

A total of 80 cfDNA whole-genome bisulfite sequencing samples were used in this study. Among them, 40 samples were obtained from healthy individuals, while the remaining 40 were breast cancer cases. The case samples were selected based on clinical criteria, including an age range of 40–60 years and a confirmed cancer stage of II–III. These 40 breast cancer samples were further divided into four molecular subtypes: Luminal B (n = 10), HER2+ (n = 10), TNBC (n = 10), and TPBC (n = 10).

4.3 Data Acquisition

Raw sequencing data (SRA format) of cfDNA were obtained from the Sequence Read Archive (SRA) repository in NCBI using SRA Toolkit and corresponding FASTQ files were generated. Genome alignment was carried out between cfDNA sequences and reference genome (hg38) retrieved from the Ensembl database.

4.4 Processing and Alignment

Following data acquisition, the FASTQ files were aligned to the hg38 reference genome using Bismark, a bisulfite-aware aligner designed for whole-genome bisulfite sequencing data. The bisulfite-treated cfDNA reads underwent alignment, deduplication, and indexing to ensure high-quality mapping. Processed alignment files were prepared according to the input specifications required for cfDNAFE, the toolkit used for multi-omics fragmentation feature extraction.

4.5 Fragmentation Analysis

Further multi-omics feature extraction was conducted using cfDNAFE, a python-based toolkit. These features include FSD (Fragment Size Distribution), FSR (Fragment Size Ratio), FSC

(Fragment Size Coverage), CNV (Copy Number Variation), BPM (Breakpoint Motif), EDM (End Motif) and Methylation.

4.5.1 Fragment Size Ratio (FSR)

cfDNA fragments are divided into three categories based on their size: short (65-150bp), intermediate (151-260bp), and long (261-400bp). The genome is divided into 5MB bins, and the ratio of each type of fragment, short, intermediate, and long, present in each bin is calculated. The corresponding ratios are then plotted on graphs that display comparisons between genomic regions and fragment lengths.

4.5.2 Fragment Size Coverage (FSC)

FSC quantifies the abundance of short, intermediate, long, and total DNA fragments within contiguous 5MB genomic bins. Prior to normalization, fragment counts within each bin are corrected for GC content biases. The final output is a Z-score plot, which indicates the deviation of the fragment number in a specific bin from the genomic mean, providing a standardized measure of fragment abundance relative to the rest of the genome.

4.5.3 Fragment Size Distribution (FSD)

FSD provides a fragment distribution profile across 41 chromosome arms and each chromosome arm is divided into 5bp segments. The number of cfDNA fragments present in each segment is divided by the total number of fragments in the chromosome arm to calculate the ratio, which is then plotted in the fragment size distribution graph. The size of all the fragments that are used in the calculation is between 65bp to 400bp.

4.5.4 Copy Number Variation (CNV)

CNV was assessed by estimating fragment count differences across the genome using a 1 megabase (1 MB) binning method. The aligned cfDNA reads were grouped into consecutive 1 MB genomic windows, and the number of fragments in each bin was compared to the baseline average to determine whether the region showed relative gain or loss. Bins with higher than expected fragment counts suggest possible amplification, whereas bins with reduced counts may indicate deletion-type signals. This fragment depth based method enables the detection of broad copy number changes in plasma-derived cfDNA at a chromosomal scale.

4.6 Motif Profiling

Motif profiling in this study consisted of two complementary analyses: Breakpoint Motif (BPM) profiling and End Motif (EDM) profiling.

4.6.1 BreakPoint Motif (BPM)

BPM refers to the short sequence patterns surrounding the exact location where the DNA strand breaks. When cfDNA is released into the bloodstream, the genomic DNA is cleaved by different enzymes or degradation processes, leaving characteristic sequence motifs at and around the breakpoint. BPM profiles examine the nucleotides immediately upstream and downstream of the cleavage site, allowing the detection of sequence preferences at breakpoints. These preferences can differ depending on the biological source of cfDNA, the type of tissue, and the dominant cell death mechanism. Certain cancers are known to have distinct breakpoint motif patterns, which may reflect altered nuclease activity, chromatin accessibility, or tumor-specific degradation pathways.

4.6.2 End Motif (EDM)

EDM refers to the short sequence of bases found at the exposed end of each cfDNA fragment. Instead of examining the region around the break, EDM looks specifically at the first few nucleotides that appear at the fragment's terminal end after cleavage. These end motifs are produced by the way DNA is cut, and certain patterns tend to appear more frequently in specific biological conditions. Differences in these end motif profiles between healthy and cancer samples may reflect changes in nuclease activity or chromatin processing in disease. Variations in end motif patterns can indicate which tissue the cfDNA came from and may help distinguish tumor-derived cfDNA.

4.7 Methylation Profiling

The methylation analysis was conducted using a fragment based approach from whole genome bisulfite sequencing data, applying the UXM methodology. UXM provides extraction of a methylation profile that contains three types of fragments with minimum 4 CpGs. This classification also includes U reads that are less than or equal to 25% methylated CpGs. In a M read, it accounts for more than or equal to 75% methylated CpGs. In a X read, the percentage of methylated CpGs is above 25% and less than 75%. In this study, only the methylated (M) fragments were used for analysis. This focused selection isolates fragments characteristic of high methylation density. These M-read proportions were calculated across a pre-defined set of 1246 cell-type specific markers.

4.8 Statistical Analysis

Statistical analyses were performed to compare multi-omics cfDNA features across healthy controls and the four breast cancer subtypes. Data were normalized using Z-scores and

ratio-based metrics, and p-value–based testing was applied to identify the top ten or thirty motifs with significant differences.

4.9 Visualization

Finally, downstream analysis and visualization was performed through the generation of graphs and heatmaps using the R packages ggplot2 and pheatmap. Fragmentation-derived features and methylation-based outputs were visualized using comparative plots, genome-wide profiles, and clustered heatmaps to highlight differences between healthy controls and the four breast cancer subtypes.

Chapter 5

Results

5.1 Analysis of fragmentation feature profiles

5.1.1 Fragment Size Ratio (FSR)

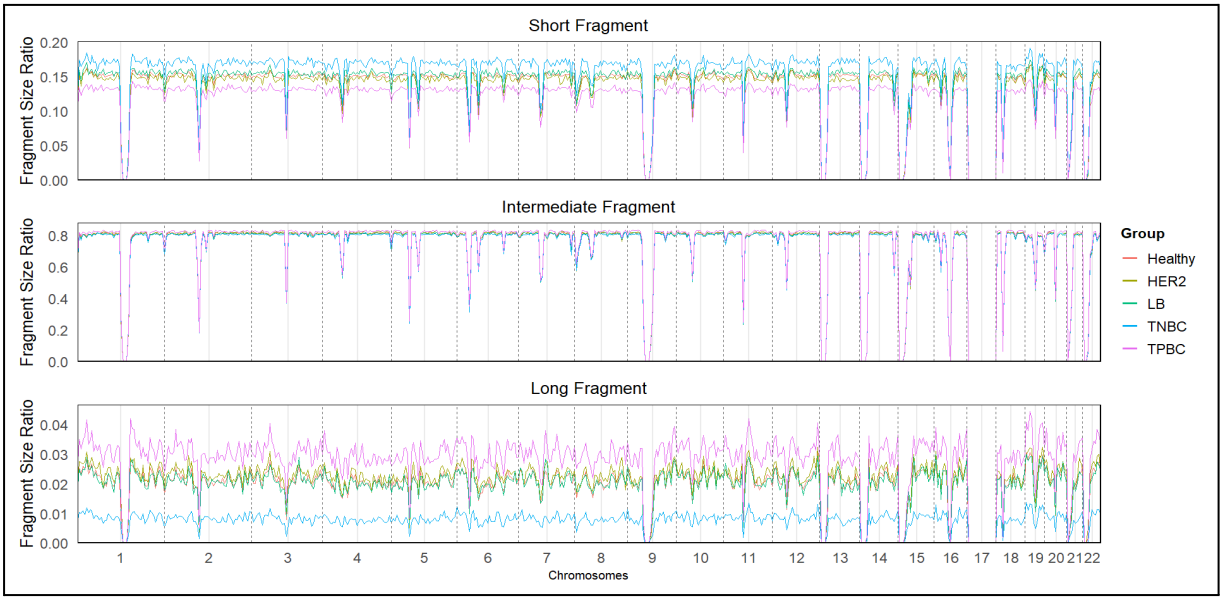


Figure 5.1 : Fragment Size Ratio Profiles of Breast Cancer Subtypes and Control

The FSR profiles across the autosomal genome showed distinct fragmentation patterns, particularly for TNBC and TPBC, which demonstrated opposite trends compared to each other and to the other groups. In contrast, healthy controls, LB and HER2+ displayed a narrower and more comparable range, with short fragments remaining around 0.15, intermediate fragments near 0.8 and long fragments around 0.02–0.03 across most autosomes. In the short fragment panel, TNBC exhibited the highest ratio (> 0.15), whereas TPBC showed the lowest. HER2+, Luminal B, and healthy controls displayed intermediate values relative to those 2 cancer subtypes. The intermediate fragments formed most of the cfDNA, with values close to 0.80 across the genome, consistent with the usual ~167 bp nucleosome-protected length. For long fragments, TPBC displayed the highest ratio (0.03–0.05), while TNBC demonstrated the lowest (around 0.01), with the remaining groups positioned between these two.

Overall, TNBC and TPBC presented opposite patterns in both short and long fragment ratios, indicating subtype-specific fragmentation characteristics [66]

5.1.2 Fragment Size Coverage (FSC)

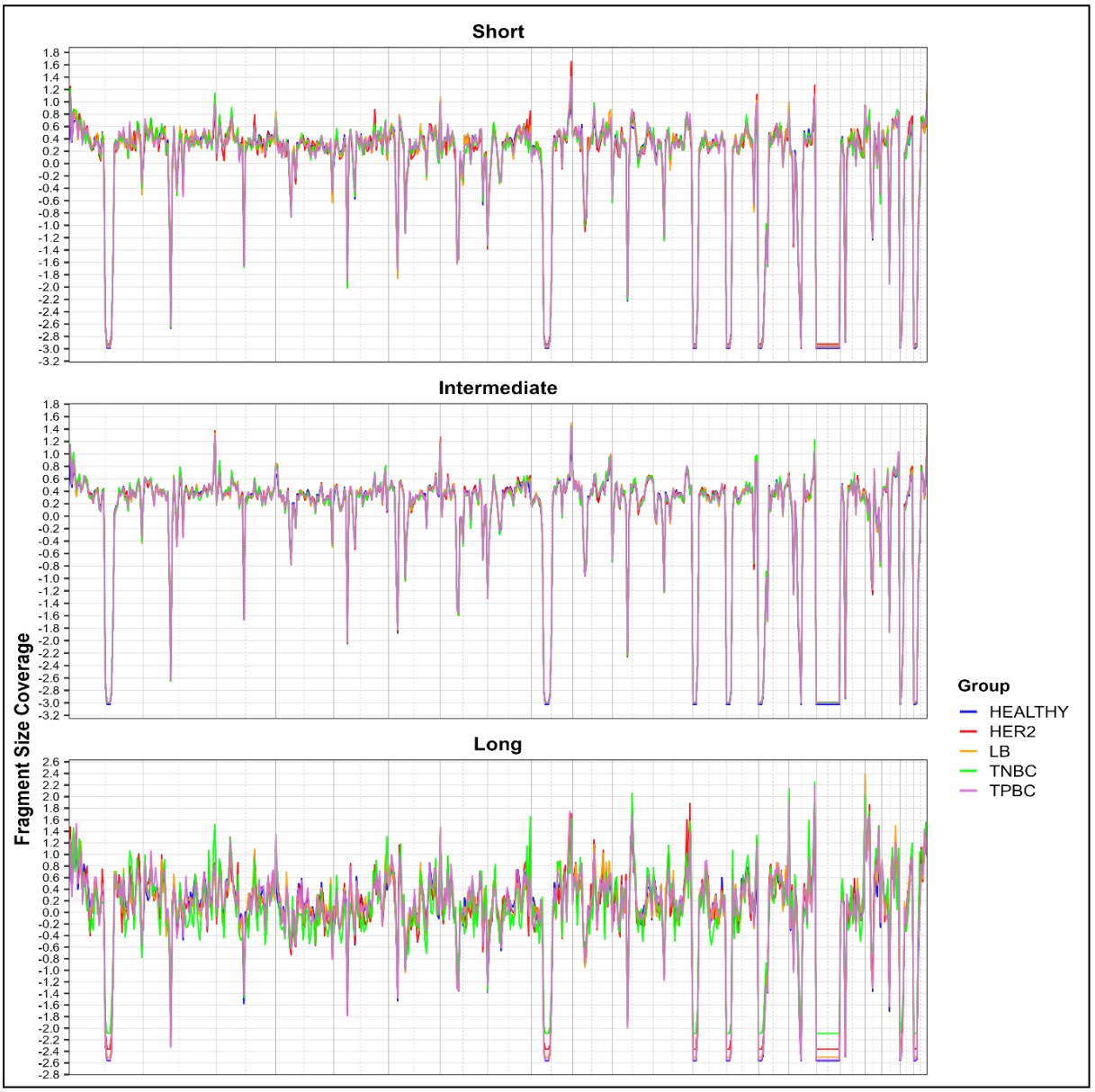


Figure 5.2 : Fragment Size Coverage Profiles of Breast Cancer Subtypes and Control

Fragment size coverage (FSC) was measured for short, intermediate, long, and total cfDNA fragments across chromosomes 1–22 using 5 Mb genomic bins. The overall pattern of coverage was largely similar in healthy controls and in the four breast cancer subtypes (HER2+, Luminal B, TNBC and TPBC). The same sharp drops in coverage appeared at

specific chromosomal points in every group, which are known regions containing highly repetitive DNA such as telomeres and centromeres where reads cannot be uniquely aligned. Intermediate fragments made up the biggest share of total coverage, which fits the expected 167 bp nucleosomal structure of cfDNA. Long fragments showed more fluctuation but did not show a distinct trend for any subtype. Based on these findings, FSC did not show a noticeable difference between the breast cancer subtypes at this genomic scale.

5.1.3 Fragment Size Distribution (FSD)

Similar peaks are visible in each chromosome arm except chromosome number 17, where no fragmentation event is observed. The peaks in healthy controls, TNBC, HER2+, and LB are of similar height, with frequencies ranging from 0.130 to 0.140 observed in the same region for each chromosomal segment. The bulk of the fragments are derived from one size, notably around the beginning of each segment. Small, negligible points are also observed for HER2+, TPBC, and healthy control in the middle of each chromosomal segment, which corresponds to the range of long fragment length (261-400bp).

The overall FSD profile is similar across all groups, including the four subtypes and healthy controls. However, TNBC exhibits the highest fragment distribution among all other subtypes for 17 of the 41 chromosome arms analysed. The highest frequency, at just above 0.14, is also observed from TNBC samples. On the other hand, TPBC has the lowest fragment distribution along its chromosomes, with frequencies primarily ranging from 0.120 to 0.125.

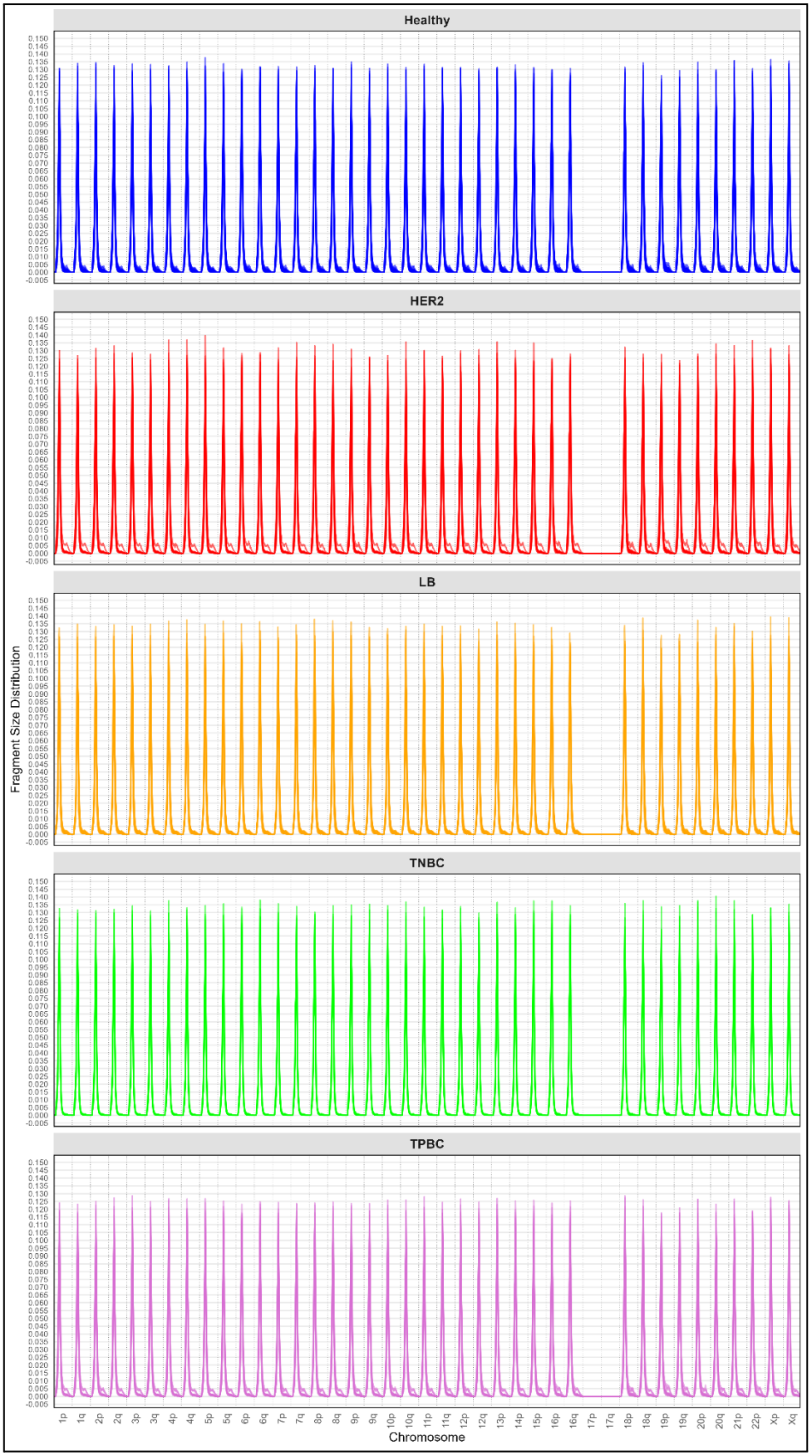


Figure 5.3 : Fragment Size Distribution Profiles of Breast Cancer Subtypes and Control

5.1.4 Copy Number Variation (CNV)

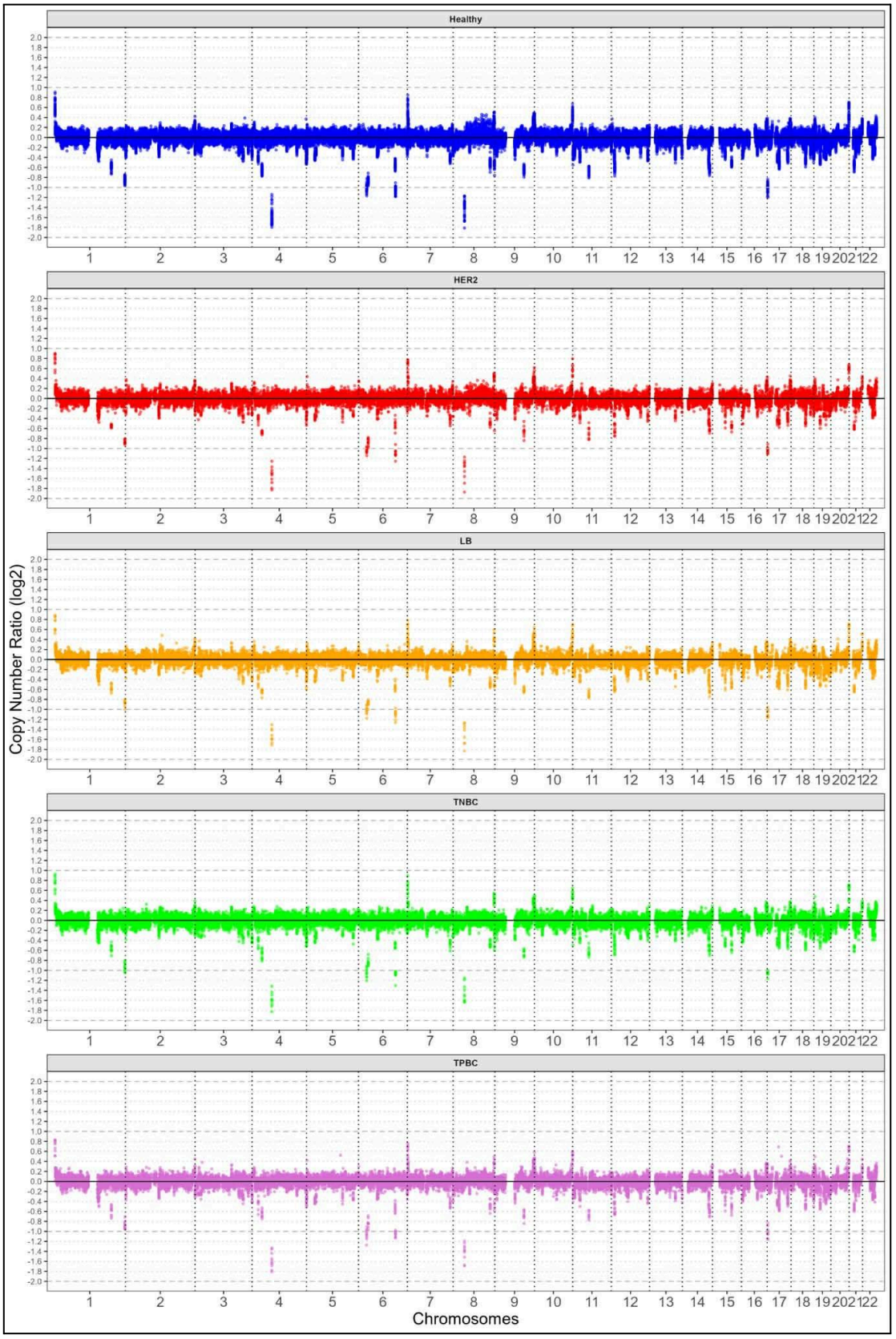


Figure 5.4 : Copy Number Variation Profiles of Breast Cancer Subtypes and Control

The CNV profiles across the autosomal genome were largely stable and remained close to the baseline for all groups, including healthy controls and the four breast cancer subtypes. HER2+ and TPBC showed minor fluctuations around the baseline without a consistent upward or downward shift across chromosomes. The Luminal B group displayed a slight downward trend, suggesting a small increase in deletion-type signals, though not uniform across the genome. TNBC exhibited moderate fluctuations with a clearer downward shift in certain chromosomal regions; however, the pattern was not continuous or subtype-defining. No group demonstrated prominent amplification or deletion events that persisted across multiple contiguous chromosomal segments. Overall, none of the subtypes displayed a distinctive CNV signature that clearly separated them from each other or from healthy controls.

5.2 Analysis of Motif Patterns

5.2.1 BreakPoint Motif (BPM)

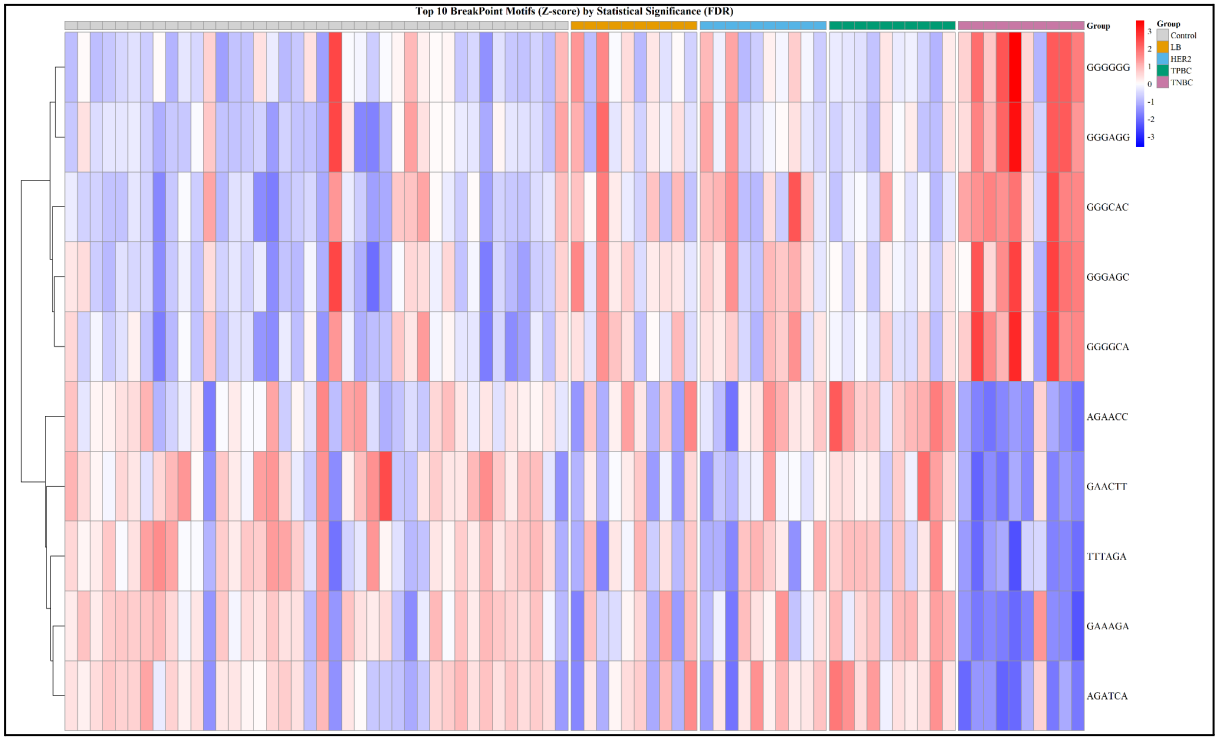


Figure 5.5 : **The Heat Map of 6-mer Motif Profiles of 40 Healthy cfDNA Samples and Breast Cancer Subtypes (LB, HER2+, TPBC, TNBC) Samples**

Breakpoint motif profiling was performed to characterize sequence-specific DNA cleavage preferences across breast cancer subtypes. Using the limma package, we identified the top ten most significantly dysregulated breakpoint motifs among the five groups. Z-score heatmap visualization revealed a subtype-specific fragmentation pattern. Notably, TNBC samples demonstrate an enrichment of G-rich motifs GGGGGG, GGGAGG, GGGGCA) in the top cluster and and strong depletion of A-rich motifs (AGAACC, GAACTT, GAAAGA) in the bottom cluster. To compare the control samples exhibits the opposite pattern in both clusters. However, TPBC, LB and HER2+ subtypes showed more balanced enrichment and depletion patterns, similar to the control group.

5.2.2 End Motif (EDM)

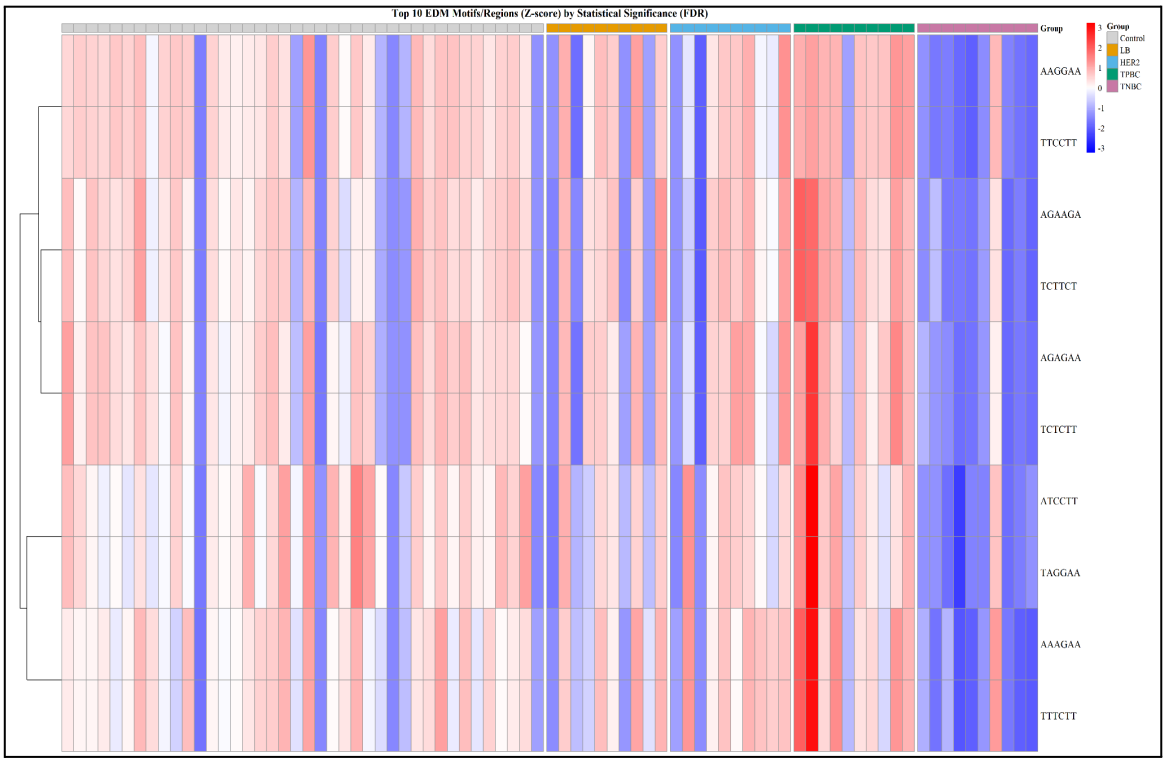


Figure 5.6 : **The Heat Map of 6-mer End Motif Profiles of 40 Healthy cfDNA Samples and Breast Cancer Subtypes (LB, HER2+, TPBC, TNBC) Samples**

Each cfDNA sample initially contained a comprehensive profile of thousands of possible 6-mer breakpoint motif frequencies. Our analysis used statistical filtering to identify the top ten most subtype differentiating motifs, which are the only one represented in the final heatmap visualization.

The two of the resulting subtypes revealed highly subtype-specific EDM signatures. The resulting heatmap revealed distinct subtype-specific fragmentation signatures. Control samples show balanced motif enrichment and depletion. In contrast, the TNBC sample shows strong and consistent depletion of multiple motifs, forming a blue pattern across most rows. Conversely, TPBC samples demonstrate an enriched pattern across the motifs. However, HER2+ and LB demonstrate a mixed pattern with moderate enrichment and depletion.

5.3 Analysis of Epigenetic Feature Profiles

5.3.1 Methylation

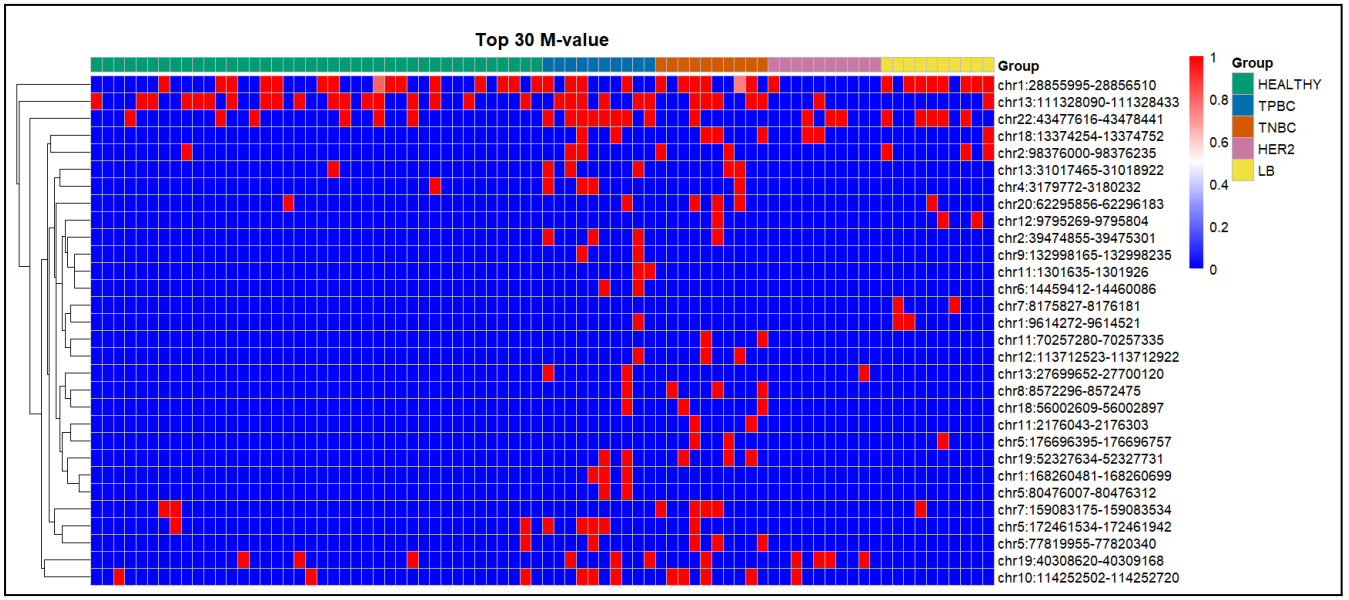


Figure 5.7 : Observed Methylation Pattern for Top Most Significant Markers (≥ 75% Methylated CpG island)

Significant differences in blue and red regions can be observed from the methylation heatmap comparing four subtypes of breast cancer and healthy controls. In healthy samples, marker regions are overwhelmingly blue, whereas a much higher amount of red is visible in the cancer segment. Out of all the subtypes, TNBC and TPBC show a larger number of markers with a high M-value of 0.8-1.0 (shown as red squares). HER2+ exhibits the lowest number of markers with high M-value among all the subtypes, largely mimicking the healthy controls' pattern. Interestingly, the first row exhibits a high M-value consistently across healthy control and the subtypes TNBC, TPBC, and LB. However, for HER2+, it mostly shows a low M-value.

Chapter 6

Discussion

Among all the subtypes studied, it was possible to clearly distinguish and differentiate between TPBC and TNBC from other groups using the feature profiles FSR, FSD, BPM, EDM and Methylation. While CNV and FSC did not prove to be good parameters for subtype identification, increasing the sample size may deliver more conclusive results, not only from these features but also for better differentiation between all four subtypes analysed.

The Fragment Size Ratio (FSR) profiles demonstrated clear subtype-specific distinctions, particularly for TNBC and TPBC, which displayed opposite trends across short and long fragment categories. The x axis here represents consecutive 5 Mb genomic bins across chromosomes 1 to 22 and the y axis demonstrates the proportion of cfDNA fragments belonging to each size category. From the graph, TNBC showed a higher proportion of short fragments, whereas TPBC exhibited a higher ratio of long fragments. Short cfDNA fragments are typically linked to increased nuclease activity and apoptotic fragmentation. Therefore, the elevated short-fragment ratio in TNBC may indicate increased cellular turnover and more active fragmentation dynamics. Conversely, the comparatively higher proportion of long fragments in TPBC suggests fewer fragmentation events, which may be influenced by differences in tumour biology and cfDNA release mechanisms. Healthy controls, HER2+, and Luminal B presented a narrower and more consistent range around 0.15 for short fragments, 0.8 for intermediate fragments, and 0.02–0.03 for long fragments. Most circulating DNA fragments belonged to the intermediate range, consistent with the 167 bp nucleosome-protected size typically produced during normal apoptotic processes. This pattern indicates that the majority of cfDNA across all groups originates through canonical apoptotic release. In the graph, sharp declines were observed near most of the centromeric and telomeric regions. These genomic regions consist of highly repetitive DNA, which limits accurate mapping because identical sequences occur across multiple sites. As a result, sequencing reads cannot be confidently assigned to a single genomic position. Consequently, these regions show

reduced fragment counts not due to biological absence of cfDNA, but due to alignment constraints. The contrasting trends between TNBC and TPBC suggest that FSR provides valuable insights into subtype-related fragmentation behaviour. With TNBC showing distinctly higher short-fragment representation and TPBC showing higher long-fragment ratios, FSR demonstrated strong potential for distinguishing between these two breast cancer subtypes. With an increased sample size, it may be possible to identify clearer patterns among the remaining subtypes, which could strengthen the utility of FSR as a supportive non-invasive measure for subtype differentiation.

The Fragment Size Coverage (FSC) analysis did not show distinct patterns that differentiate the four breast cancer subtypes from each other or from healthy controls. Across the autosomal chromosomes, the coverage values for short, intermediate, and long fragments followed a very similar trend in all groups, indicating that coverage distribution remains generally consistent regardless of subtype. The intermediate fragment category contributed the highest overall coverage, which is consistent with the known nucleosomal size of ~167 bp generated during controlled apoptotic fragmentation. This pattern appearing across both cancer subtypes and healthy individuals suggests that the dominant process producing cfDNA fragments is shared, even though the biological context may differ. Sharp drops in coverage were visible at specific chromosomal locations in all groups and were noticeable in short, intermediate, and long fragment categories. These points correspond to regions rich in repetitive DNA, including centromeres and telomeres, where sequencing reads cannot be confidently aligned to a single genomic position. As a result, the coverage decline most likely reflects a technical limitation of short-read alignment rather than a biological fragmentation difference. Long fragment coverage showed more noticeable fluctuation from chromosome to chromosome; however, the change did not follow a consistent pattern that could be linked to a specific subtype. Although long fragment coverage fluctuated across the chromosomes in all

subtypes, these changes did not represent a subtype-specific difference. With the current dataset, FSC did not demonstrate strong potential for separating breast cancer subtypes at this level of analysis. A larger dataset may help determine whether small differences exist that are not visible with the present sample number. Future work with more samples may show its significance when assessed alongside other cfDNA features such as fragment size ratio, breakpoint motifs or methylation-based measures.

Fragment distribution is observed to be the highest for a single specific fragment size in all the graphs. The peaks correspond to a region approximately between short and intermediate fragment size ranges. Given the nature of cfDNA from healthy cells, it can be safely assumed that the fragments are normal cfDNA with a size of about 167bp. As TPBC shows an overall lower fragment distribution compared to other groups, it can be easily identified and separated for subtype detection. The occurrence of a lower fraction is likely due to fragmentation events being comparatively less frequent in TPBC cancer cells compared to other types of neoplastic and non-neoplastic cells. Additionally, a small but significant fraction of long fragments visible within HER2+, TPBC, and control samples makes HER2+ indistinguishable from other groups analysed. The CNV analysis did not identify consistent chromosomal gains or losses that could be linked specifically to any of the breast cancer subtypes included in this study. Although TNBC showed slightly stronger downward shifts, and Luminal B displayed a mild overall decrease, these changes were scattered and did not form a stable or continuous pattern along the genome. The largely baseline-aligned CNV profiles suggest that copy number changes in plasma cfDNA were not strong enough within this cohort to distinguish subtypes reliably. The absence of clear separation may be influenced by the small sample size, which limits the ability to detect subtle CNV differences. Unlike tissue-based CNV profiling, cfDNA contains a mixture of tumor-derived and normal DNA, which may reduce the visibility of detectable changes. Based on these findings, CNV alone does not appear to provide

subtype-specific information in this dataset. However, future work with more samples may clarify whether CNV gains or losses contribute useful information when combined with other cfDNA features.

Break point motif profiling demonstrates subtype specific fragmentation signatures that anticipate biological insight into the underlying cleavage mechanisms in breast cancer. In TNBC a strong enrichment of G-rich motifs and depletion of A-rich motifs are specifically observed showing that in TNBC, DNA frequently breaks next to G-rich sequences whereas it avoids breaking at A-rich sequences. This indicates elevated oxidative stress and altered nuclease activity [67]. In contrast, in TPBC, HER2+, LB subtypes a more balanced enrichment and depletion pattern was noticed that closely resembled healthy controls. This distinction highlights that tumor derived cfDNA fragmentation is not uniform across breast cancer subtypes rather , it demonstrates a molecular landscape unique to each subtype. Overall, break point motif analysis adds an additional dimension to cfDNA fragmentation profiling by representing subtype-specific cleavage preferences, with TNBC standing out as the most epigenetically and structurally disrupted groups. These findings support the potential of breakpoint motifs as non-invasive biomarkers for subtype differentiation and disease aggressiveness.

The end motif fragmentation signatures in this study provides an important understanding of chromatin structure and nuclease activity across breast cancer subtypes. The strong and widespread depletion of end-motifs in TNBC indicates highly disordered and unpredictable DNA fragmentation, which is consistent with the aggressive biology and genomic instability characteristic of this subtype [68]. On the contrary, in TPBC the enriched end motif patterns reflect a more organized fragmentation profile suggesting lower genomic instability, preserved nuclease fidelity and more consistent cleavage behavior. This aligns with the molecular characteristics of TPBC [69]. Meanwhile in LB and HER2+ the mixed and variable patterns

suggest that these subtypes retain some sequence specific fragmentation signals while also exhibiting cancer-associated disruptions. Overall, these findings demonstrate that cfDNA fragmentation is not uniform across breast cancer subtypes. Instead, it reflects the distinct biological landscapes of each subtype and highlights the potential of end-motif profiling as a non-invasive tool for capturing distinct tumor behaviors.

The top 30 markers are largely unmethylated in healthy samples, as evident from the low M-scores throughout the segment. Compared to the control, cancer samples exhibit elevated methylation regions. TNBC and TPBC contain the most markers with hypermethylation, followed by LB. Among all the subtypes, methylation of CpG islands occurred the least in HER2+ samples. For the location chr1:28855995-28856510, all groups show hypermethylation except HER2+, which mostly contains an average rate of CpG methylation. Overall, due to the high variations shown by TPBC and TNBC compared to control samples, epigenetic patterns prove to be an excellent candidate for identifying these two subtypes during breast cancer diagnostic applications. Furthermore, due to its differential nature from the rest of the subtypes, chr1:28855995-28856510 proves to be a reliable biomarker to distinguish between HER2+ and other subtypes of breast cancer, making it a favorable candidate for HER2+ detection from liquid biopsy.

One limitation of this work is the size of the cohort, as a greater number of samples may help determine whether similar patterns exist in subtypes that did not show separation here. The analysis was carried out computationally using secondary data and without laboratory generated samples, which limits the extent to which these observations can be confirmed at this stage.

Looking forward, cfDNA holds strong future potential as a minimally invasive biomarker for early detection, subtype classification and treatment monitoring in breast cancer. Because

cfDNA can be obtained from a simple blood sample, it may reduce dependence on repeated tissue biopsies and allow more frequent assessment of disease changes over time. Expanding this work with larger cohorts, inclusion of clinical and treatment response data and direct laboratory validation could support the development of cfDNA based approaches that contribute more confidently to patient management [70,71].

Chapter 7

Conclusion

This study explored whether cfDNA can reflect subtype specific characteristics in breast cancer by examining fragmentation-based measures such as BreakPoint Motif (BPM), End Motif (EDM), Fragment Size Distribution, Fragment Size Coverage, Fragment Size Ratio and Copy Number Variation together with methylation related features across four subtypes and healthy controls. Our findings indicated subtype specific distinctions in TNBC and TPBC, where notable differences in fragmentation and methylation patterns were observed when compared with controls and other subtypes. Among the fragmentomic features evaluated, Fragment Size Ratio showed noticeable difference for TPBC and TNBC but did not demonstrate the same level of distinction for LB or HER2+. In methylation, end motif and breakpoint motif analyses, observable patterns were also more prominent in TNBC and TPBC, while clear differentiation was not consistently evident for LB and HER2+ within this dataset.

Overall, this study supports the prospect that cfDNA contains measurable signals related to subtype specific characteristics and highlights the importance of further research to translate these findings into practical tools that support early diagnosis and better informed clinical decisions in breast cancer care.

References

1. Smolarz, B.; Nowak, A.Z.; Romanowicz, H. Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). Cancers 2022, 14, 2569. https:// doi.org/10.3390/cancers14102569
2. WHO. Breast cancer [Internet]. World Health Organization. 2025. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
3. Kim J, Harper A, McCormack V, Sung H, Houssami N, Morgan E, et al. Global patterns and trends in breast cancer incidence and mortality across 185 countries. Nature Medicine [Internet]. 2025 Feb 24;31(4).
4. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians [Internet]. 2024 Apr 4;74(3):229–63. Available from: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21834
5. Eibl RH, Schneemann M. Cell-free DNA as a biomarker in cancer. Extracellular Vesicles and Circulating Nucleic Acids. 2022;3(3):178–98.
6. Cisneros-Villanueva M, Hidalgo-Pérez L, Rios-Romero M, Cedro-Tanda A, Ruiz-Villavicencio CA, Page K, et al. Cell-free DNA analysis in current cancer clinical trials: a review. British Journal of Cancer [Internet]. 2022 Feb 1;126(3):391–400. Available from: https://www.nature.com/articles/s41416-021-01696-0
7. Lin CH, Hsieh SY, Sheen I-Shyan, Lee WC, Chen TC, Shyu WC, et al. Genome-wide Hypomethylation in Hepatocellular Carcinogenesis1. Cancer Research [Internet]. 2001 May;61(10):4238–43. Available from: https://aacrjournals.org/cancerres/article/61/10/4238/507446/Genome-wide-Hypomethylation-in-Hepatocellular
8. Lehmann U, Länger F, Feist H, Glöckner S, Hasemeier B, Kreipe H. Quantitative Assessment of Promoter Hypermethylation during Breast Cancer Development. The American Journal of Pathology [Internet]. 2002 Feb 1 [cited 2022 Jul 18];160(2):605–12. Available from: https://www.sciencedirect.com/science/article/pii/S0002944010648808
9. Jackson K, Yu MC, Arakawa K, Fiala E, Youn B, Fiegl H, et al. DNA hypomethylation is prevalent even in low-grade breast cancers. Cancer Biology & Therapy. 2004 Dec;3(12):1225–31.

10. Subramaniam MM, Chan JY, Soong R, Ito K, Ito Y, Yeoh KG, et al. RUNX3 inactivation by frequent promoter hypermethylation and protein mislocalization constitute an early event in breast cancer progression. Breast Cancer Research and Treatment. 2008 Feb 6;113(1):113–21.
11. 1.Hum M, Lee ASG. DNA methylation in breast cancer: early detection and biomarker discovery through current and emerging approaches. Journal of Translational Medicine. 2025 Apr 23;23(1).
12. 1.Cancer Today [Internet]. Iarc.fr. 2025 [cited 2025 Dec 10]. Available from: https://gco.iarc.fr/today/en/dataviz/pie?mode=population&group_populations=0&sexes=2&cancers=20&types=1
13. Beckmann, M. W., Niederacher, D., Schnürch, H. G., Gusterson, B. A. & Bender, H. G. Multistep carcinogenesis of breast cancer and tumour heterogeneity. J. Mol. Med. 75, 429–439 (1997).
14. Bejarano, L., Jordao, M.J.C., and Joyce, J.A. (2021). Therapeutic targeting of the tumor microenvironment. Cancer Discov. 11, 933–959. https://doi.org/10.1158/2159-8290.CD-20-1808.
15. Fletcher, O. & Houlston, R. S. Architecture of inherited susceptibility to common cancer. Nat. Rev. Cancer 10, 353–361 (2010).
16. Posner MC, Wolmark N. Non-invasive breast carcinoma. Breast Cancer Research and Treatment. 1992 Oct;21(3):155–64.
17. Delaloge S, Seema Ahsan Khan, Wesseling J, Whelan T. Ductal carcinoma in situ of the breast: finding the balance between overtreatment and undertreatment. Lancet. 2024 May 1;
18. Martins, F., Sofiya, L., Sykiotis, G.P. et al. Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance. Nat Rev Clin Oncol 16, 563–580 (2019). https://doi.org/10.1038/s41571-019-0218-0
19. Timbres J, Kohut K, Caneppele M, Troy M, Schmidt MK, Roylance R, et al. DCIS and LCIS: Are the Risk Factors for Developing In Situ Breast Cancer Different? Cancers [Internet]. 2023 Jan 1;15(17):4397. Available from: https://www.mdpi.com/2072-6694/15/17/4397
20. King TA, Pilewskie M, Muhsen S, Patil S, Mautner SK, Park A, Oskar S, Guerini-Rocco E, Boafo C, Gooch JC, De Brot M, Reis-Filho JS, Morrogh M, Andrade VP, Sakr RA, Morrow M. Lobular Carcinoma in Situ: A 29-Year Longitudinal Experience Evaluating Clinicopathologic Features and Breast Cancer

Risk. J Clin Oncol. 2015 Nov 20;33(33):3945-52. doi: 10.1200/JCO.2015.61.4743. Epub 2015 Sep 14. PMID: 26371145; PMCID: PMC4934644.

21. Miyazaki Y, Shimizu J, Kubo Y, Tabata N, Aso T. Quantitative classification of invasive and noninvasive breast cancer using dynamic magnetic resonance imaging of the mammary gland. Journal of Clinical Imaging Science. 2022 Aug 5;12(45):45.
22. Wang, J. et al. Progression from ductal carcinoma in situ to invasive breast cancer: molecular features and clinical significance. Signal Transduct. Target. Ther. 9, 83 (2024).
23. 1.Posner MC, Wolmark N. Non-invasive breast carcinoma. Breast Cancer Research and Treatment. 1992 Oct;21(3):155–64.
24. Luveta J, Parks RM, Heery DM, Cheung KL, Johnston SJ. Invasive Lobular Breast Cancer as a Distinct Disease: Implications for Therapeutic Strategy. Oncology and Therapy. 2019 Dec 24;8(1):1–11.
25. American Cancer Society. Breast cancer facts & figures 2019–2020. Atlanta (GA): American Cancer Society; 2019
26. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747-52. doi: 10.1038/35021093. PMID: 10963602.
27. Luveta J, Parks RM, Heery DM, Cheung KL, Johnston SJ. Invasive Lobular Breast Cancer as a Distinct Disease: Implications for Therapeutic Strategy. Oncology and Therapy. 2019 Dec 24;8(1):1–11.
28. Chamalidou C, Fohlin H, Albertsson P, Arnesson LG ., Einbeigi Z, Holmberg E, et al. Survival patterns of invasive lobular and invasive ductal breast cancer in a large population-based cohort with two decades of follow up. The Breast [Internet]. 2021 Oct [cited 2022 Sep 15];59:294–300. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8361199/
29. Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA. Subtypes of Breast Cancer [Internet]. Mayrovitz HN, editor. PubMed. Brisbane (AU): Exon Publications; 2022. Available from: https://www.ncbi.nlm.nih.gov/books/NBK583808/

30. Iqbal N, Iqbal N. Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications. Molecular Biology International. 2014 Sep 7;2014(1):1–9.
31. Xiao W, Zheng S, Yang A, Zhang X, Zou Y, Tang H, et al. Breast cancer subtypes and the risk of distant metastasis at initial diagnosis: a population-based study. Cancer Management and Research [Internet]. 2018 Nov 5 [cited 2021 May 23];10:5329–38. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6225920/
32. Goyette ERJ, Muller KE. Molecular subtypes [Internet]. PathologyOutlines.com; [cited 2025 Dec 10]. Available from: https://www.pathologyoutlines.com/topic/breastmolecularsubtypes.html
33. Carvalho E, Canberk S, Schmitt F, Vale N. Molecular Subtypes and Mechanisms of Breast Cancer: Precision Medicine Approaches for Targeted Therapies. Cancers [Internet]. 2025 Mar 25;17(7):1102. Available from: https://www.mdpi.com/2072-6694/17/7/1102#B5-cancers-17-01102
34. Table - PMC [Internet]. Nih.gov. 2025 [cited 2025 Dec 10]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3144634/table/tbl2/
35. Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA. Table 1, Characteristics of subtypes of breast cancer [Internet]. www.ncbi.nlm.nih.gov. 2022. Available from: https://www.ncbi.nlm.nih.gov/books/NBK583808/table/Ch3-t0001/?report=objectonly
36. Xie J, Yang Z, Li Z, Zhang T, Chen H, Chen X, et al. Triple-positive breast cancer: navigating heterogeneity and advancing multimodal therapies for improving patient outcomes. Cancer Cell International. 2025 Mar 5;25(1).
37. Ignatov A, Eggemann H, Burger E, Ignatov T. Patterns of breast cancer relapse in accordance to biological subtype. Journal of Cancer Research and Clinical Oncology. 2018 Apr 19;144(7):1347–55.
38. Breast Cancer Stages | Understanding Breast Cancer Staging [Internet]. Cancer.gov. 2025. Available from: https://www.cancer.gov/types/breast/stages
39. Breast Cancer Staging [Internet]. ACS. Available from: https://www.facs.org/for-patients/the-day-of-your-surgery/breast-cancer-surgery/breast-cancer-types/breast-cancer-staging/
40. Volik S, Alcaide M, Morin RD, Collins C. Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies. Molecular Cancer Research. 2016 Jul 15;14(10):898–908.

41. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 1997 Sep;389(6648):251–60.
42. Jiang P, Lo YMD. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. Trends in Genetics. 2016 Jun;32(6):360–71.
43. Jiang P, Lo YMD. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. Trends in Genetics. 2016 Jun;32(6):360–71.
44. Yazaki S, Tokura M, Aiba H, Kojima Y, Shiraishi K. Clinical applications of cell-free DNA-based liquid biopsy analysis. Translational Oncology [Internet]. 2025 Sep 6;61:102519. Available from: https://www.sciencedirect.com/science/article/pii/S1936523325002505
45. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes & development [Internet]. 2009;23(7):781–3. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19339683
46. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology [Internet]. 2012 Jul 11;38(1):23–38. Available from: https://www.nature.com/articles/npp2012112
47. CDC. Genetics Basics [Internet]. Genomics and Your Health. 2024. Available from: https://www.cdc.gov/genomics-and-health/about/index.html
48. Gopalakrishnan S, Van Emburgh BO, Robertson KD. DNA methylation in development and human disease. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2008 Dec;647(1-2):30–8.
49. Jin B, Tao Q, Peng J, Soo HM, Wu W, Ying J, et al. DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. Human Molecular Genetics. 2007 Nov 27;17(5):690–709.
50. Jung M, Pfeifer GP. CpG Islands [Internet]. Maloy S, Hughes K, editors. ScienceDirect. San Diego: Academic Press; 2013. p. 205–7. Available from: https://www.sciencedirect.com/science/article/abs/pii/B9780123749840003491
51. Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research [Internet]. 1980;8(7):1499–504. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC324012/
52. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell [Internet].

1985 Jan [cited 2019 Dec 19];40(1):91–9. Available from: https://www.cell.com/cell/fulltext/0092-8674(85)90312-5

53. Cross SH, Charlton JA, Nan X, Bird AP. Purification of CpG islands using a methylated DNA binding column. Nature Genetics. 1994 Mar;6(3):236–44.
54. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, et al. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. Reik W, editor. PLoS Genetics. 2010 Sep 23;6(9):e1001134.
55. Figure - PMC [Internet]. Nih.gov. 2025 [cited 2025 Dec 10]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC12491199/figure/F3/
56. Lin CH, Hsieh SY, Sheen I-Shyan, Lee WC, Chen TC, Shyu WC, et al. Genome-wide Hypomethylation in Hepatocellular Carcinogenesis1. Cancer Research [Internet]. 2001 May [cited 2025 Dec 10];61(10):4238–43. Available from: https://aacrjournals.org/cancerres/article/61/10/4238/507446/Genome-wide-Hypomethylation-in-Hepatocellular
57. Baylin SB. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. Human Molecular Genetics. 2001 Apr 1;10(7):687–92.
58. Narayan A, Ji W, Zhang XY, Marrogi A, Graff JR, Baylin SB, et al. Hypomethylation of pericentromeric DNA in breast adenocarcinomas. International Journal of Cancer. 1998 Sep 11;77(6):833–8.
59. Szyf M, Pakneshan P, Rabbani SA. DNA methylation and breast cancer. Biochemical Pharmacology [Internet]. 2004 Sep 15;68(6):1187–97. Available from: https://pubmed.ncbi.nlm.nih.gov/15313416/
60. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and Aging. Cancer Research. 2016 Jun 2;76(12):3446–50.
61. Stefansson OA, Moran S, Gomez A, Sayols S, Arribas-Jorba C, Sandoval J, et al. A DNA methylation-based definition of biologically distinct breast cancer subtypes. Molecular Oncology. 2014 Nov 5;9(3):555–68.
62. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature [Internet]. 2012 Sep 23;490(7418):61–70. Available from: https://www.nature.com/articles/nature11412
63. Roll J, Rivenbark AG, Sandhu R, Parker JS, Jones WD, Carey LA, et al. Dysregulation of the epigenome in triple-negative breast cancers: Basal-like and claudin-low breast cancers express aberrant DNA hypermethylation. 2013 Dec 1;95(3):276–87.

64. Hum M, Lee ASG. DNA methylation in breast cancer: early detection and biomarker discovery through current and emerging approaches. Journal of Translational Medicine. 2025 Apr 23;23(1).
65. Gao L, Zhang L, Liu J, Wang L, Fu Q, Liu S, et al. Discovery and validation of cell-free DNA methylation markers for specific diagnosis, differentiation from benign tumors, and prognosis of breast cancer. Breast Cancer Research. 2025 Jun 16;27(1).
66. Massey DJ, Koren A. Telomere-to-telomere human DNA replication timing profiles. Scientific Reports. 2022 Jun 10;12(1).
67. Kostyuk S, Smirnova T, Kameneva L, Porokhovnik L, Speranskij A, Ershova E, et al. GC-Rich Extracellular DNA Induces Oxidative Stress, Double-Strand DNA Breaks, and DNA Damage Response in Human Adipose-Derived Mesenchymal Stem Cells. Oxidative Medicine and Cellular Longevity. 2015;2015:1–15.
68. Asleh K, Riaz N, Nielsen TO. Heterogeneity of triple negative breast cancer: Current advances in subtyping and treatment implications. Journal of Experimental & Clinical Cancer Research. 2022 Sep 1;41(1).
69. Vici P, Pizzuti L, Natoli C, Gamucci T, Di Lauro L, Barba M, et al. Triple positive breast cancer: A distinct subtype? Cancer Treatment Reviews. 2015 Feb;41(2):69–76.
70. Ma L, Guo H, Zhao Y, Liu Z, Wang C, Bu J, et al. Liquid biopsy in cancer current: status, challenges and future prospects. Signal Transduction and Targeted Therapy. 2024 Dec 1;9(1).
71. Dang DK, Park BH. Circulating tumor DNA: current challenges for clinical utility. Journal of Clinical Investigation. 2022 Jun 15;132(12).
72. 1.Breast Cancer Stages | Understanding Breast Cancer Staging [Internet]. Cancer.gov. 2025. Available from: https://www.cancer.gov/types/breast/stages

Supplement

Supplement Table A : List of cfDNA samples analyzed in this study, including SRA accession numbers, sample group, subtype, age range, cancer stage.

SRA Accession	Sample Group	Subtype	Age Range	Cancer Stage
SRR23280258	Healthy control	NA	42	NA
SRR23279984	Healthy control	NA	40	NA
SRR23279966	Healthy control	NA	44	NA
SRR23280232	Healthy control	NA	42	NA
SRR23279959	Healthy control	NA	40	NA
SRR23279816	Healthy control	NA	47	NA
SRR23280079	Healthy control	NA	41	NA
SRR23279796	Healthy control	NA	41	NA
SRR23279795	Healthy control	NA	45	NA
SRR23279794	Healthy control	NA	45	NA
SRR23279793	Healthy control	NA	52	NA
SRR23279791	Healthy control	NA	60	NA
SRR23280184	Healthy control	NA	54	NA
SRR23280178	Healthy control	NA	56	NA
SRR23280177	Healthy control	NA	58	NA
SRR23279799	Healthy control	NA	53	NA
SRR23280245	Healthy control	NA	52	NA
SRR23280012	Healthy control	NA	56	NA
SRR23280001	Healthy control	NA	56	NA
SRR23280020	Healthy control	NA	54	NA
SRR23280007	Healthy control	NA	50	NA
SRR23280006	Healthy control	NA	54	NA
SRR23280005	Healthy control	NA	40	NA
SRR23279800	Healthy control	NA	58	NA
SRR23280003	Healthy control	NA	41	NA
SRR23279992	Healthy control	NA	53	NA
SRR23279991	Healthy control	NA	54	NA
SRR23279989	Healthy control	NA	52	NA
SRR23279998	Healthy control	NA	50	NA
SRR23279996	Healthy control	NA	50	NA
SRR23279994	Healthy control	NA	54	NA
SRR23279839	Healthy control	NA	41	NA
SRR23279838	Healthy control	NA	44	NA
SRR23279833	Healthy control	NA	46	NA

SRA Accession	Sample Group	Subtype	Age Range	Cancer Stage
SRR23280231	Healthy control	NA	46	NA
SRR23279783	Healthy control	NA	42	NA
SRR23280060	Healthy control	NA	43	NA
SRR23280228	Healthy control	NA	42	NA
SRR23280227	Healthy control	NA	40	NA
SRR23280226	Healthy control	NA	50	NA
SRR23280119	Case	HER2	48	II
SRR23280195	Case	HER2	48	II
SRR23280193	Case	HER2	55	II
SRR23280077	Case	HER2	47	II
SRR23280076	Case	HER2	44	II
SRR23280074	Case	HER2	41	II
SRR23280066	Case	HER2	61	II
SRR23279926	Case	HER2	52	II
SRR23280051	Case	HER2	50	II
SRR23280050	Case	HER2	41	II
SRR23280067	Case	TNBC	56	I
SRR23279896	Case	TNBC	60	I
SRR23279885	Case	TNBC	39	I
SRR23280201	Case	TNBC	50	II
SRR23279916	Case	TNBC	37	II
SRR23280186	Case	TNBC	68	II
SRR23280150	Case	TNBC	45	II
SRR23280175	Case	TNBC	43	III
SRR23280127	Case	TNBC	44	III
SRR23280052	Case	TNBC	47	IIIA
SRR23280205	Case	Luminal B - HER2	55	II
SRR23280203	Case	Luminal B - HER2	47	II

SRA Accession	Sample Group	Subtype	Age Range	Cancer Stage
SRR23280196	Case	Luminal B - HER2	54	II
SRR23280063	Case	Luminal B - HER2	38	II
SRR23279785	Case	Luminal B - HER2	62	II
SRR23279778	Case	Luminal B - HER2	56	II
SRR23280055	Case	Luminal B - HER2	67	II
SRR23280046	Case	Luminal B - HER2	41	II
SRR23280240	Case	Luminal B - HER2	57	II
SRR23279969	Case	Luminal B - HER2	62	II
SRR23279761	Case	Luminal B	58	II
SRR23280138	Case	Luminal B	51	II
SRR23280109	Case	Luminal B	45	II
SRR23280204	Case	Luminal B	57	II
SRR23280199	Case	Luminal B	64	II
SRR23280197	Case	Luminal B	39	II
SRR23280191	Case	Luminal B	77	II
SRR23280075	Case	Luminal B	48	II
SRR23279931	Case	Luminal B	66	II
SRR23279930	Case	Luminal B	44	II

Supplementary Tools / Software Availability

The following tools were used for data acquisition, processing, and analysis in this study: The cfDNAFE toolkit used for multi-omics cfDNA feature extraction is publicly available on GitHub: https://github.com/Cuiwanxin1998/cfDNAFE

Other tools used in this study include:

- SRA Toolkit – for downloading and converting raw cfDNA sequencing data from the NCBI Sequence Read Archive (SRA). Available at: https://github.com/ncbi/sra-tools
- Bismark – for alignment of bisulfite-treated cfDNA reads to the reference genome. Available at: https://www.bioinformatics.babraham.ac.uk/projects/bismark/
- cfDNAFE – for multi-omics feature extraction from cfDNA. Available at: https://github.com/Cuiwanxin1998/cfDNAFE
- R packages for downstream analysis:

  ggplot2: https://ggplot2.tidyverse.org

  pheatmap: https://CRAN.R-project.org/package=pheatmap

  limma: https://bioconductor.org/packages/release/bioc/html/limma.html