Detecting Cancer from Cell-free DNA Sequencing Fragmentation Patterns

Lin Chen

St Cross College
University of Oxford

A thesis submitted for the degree of
Master of Science (by Research) in Clinical Medicine

Michaelmas 2023

Abstract

Cell-free DNA(cfDNA) represents all non-encapsulated DNA circulating freely in the human circulatory system, which usually exists in the form of wrapping around nucleosomes in the circulating system. Nucleosomes are the basic components of chromatin and play an important role in gene regulation. Through cfDNA sequencing, we can get a glimpse of the position of nucleosomes in the human body and infer gene expression levels. Some of the cfDNA in tumour patients comes from tumour cells, which is so called circulation tumour DNA(ctDNA). So searching for these ctDNA can be used as a non-invasive way to detect tumours. By searching for tumour-specific expression genes, we can locate the contribution of ctDNA in cfDNA to find a non-invasive way to diagnose and detect tumour. In this thesis, we focused on tumour-specific expression genes. We observed the correlation between gene expression and cfDNA sequencing pattern, including seq coverage and fragmentation patterns. Therefore we located the contribution of ctDNA in cfDNA to find a non-invasive way to diagnose and detect tumour patients in real time.

Detecting Cancer from Cell-free DNA Sequencing Fragmentation Patterns



Lin Chen

St Cross College

University of Oxford

A thesis submitted for the degree of

Master of Science (by Research) in Clinical Medicine

Michaelmas 2023

Acknowledgements

I am profoundly thankful to my supervisor Prof. Benjamin Schuster-Böckler. It was a very arduous journey for both of us, but ultimately it paid off. I would like thank Ben for always encouraging me, supporting me, and helping me throughout these years. I would like to thank the lab members for providing a relaxed and enjoyable environment, which made my life in Oxford much easier and happier.

I would like to express my sincere gratitude to my examiners, Prof. Dan Woodcock and Prof. Florent Moullière, for their constructive feedback and suggestions on the revision of this article. Their input was invaluable in refining my work.

I would also like to thank Prof. Sarah de Val, Prof. Colin Goding, Prof. Robert Gilbert and St Cross College for their support during this period. The strength they provided me both academically and in life enabled me to overcome many difficulties.

I would like to express my gratitude to Prof Chunxiao Song for his help during my first year and for providing many academic and life suggestions. I would also like to thank my friends for supporting me when I was feeling down. Thanks to the girls of the dance club and my ballet coach Yuka Kodama, and thanks to Dr. Leaver and the Oxford Willow team for providing me medical help. I Thank Dr Jingfei Cheng, Dr Ying Bi, Dr Yuxin Mi, Dr Yun Feng, from whom I gleaned a wealth of knowledge that has helped me maintain a balance between my academic pursuits and personal life.

Thanks to Mr. Xinyi Shao for travelling thousands of miles to provide me with emotional support when I was seriously ill. I am very grateful for the time we spent facing the disease together and support each other. I am really excited that we finally defeated it.

I am deeply grateful to Mr. Jiahao Xin for or the unwavering companionship and tender care he provided throughout our shared journey. His presence was a source of strength and comfort. His assistance, both practical and emotional, was invaluable, and I am deeply appreciative of the support he offered during our cohabitation. The delightful moments we shared in Oxford, making the experience truly memorable.

Lastly and most importantly, I extend my heartfelt thanks to my amazing family for not giving up on me when I was seriously ill and for helping me seek medical assistance. I am especially grateful to my father, Mr Donghui Chen, for picking me up and taking me back home when I was struggling with severe depression. Thanks

to my mother, Mrs Xiayi Lin, for taking care of me during pandemic in Oxford. Thanks for their cats, Meiqiu and Rose for companion and lots of joy. A special mention goes to my precious boy kitten Nasa Chen, as an emotional companion who is always by my side during my Oxford journey, he cured my loneliness and depression with his fluffy fur and adorable nature. He is not proficient in human language, but I believe he understands that I love him.

Abstract

Cell-free DNA(cfDNA) represents all non-encapsulated DNA circulating freely in the human circulatory system, which usually exists in the form of wrapping around nucleosomes in the circulating system. Nucleosomes are the basic components of chromatin and play an important role in gene regulation. Through cfDNA sequencing, we can get a glimpse of the position of nucleosomes in the human body and infer gene expression levels. Some of the cfDNA in tumour patients comes from tumour cells, which is so called circulation tumour DNA(ctDNA). So searching for these ctDNA can be used as a non-invasive way to detect tumours. By searching for tumour-specific expression genes, we can locate the contribution of ctDNA in cfDNA to find a non-invasive way to diagnose and detect tumour. In this thesis, we focused on tumour-specific expression genes. We observed the correlation between gene expression and cfDNA sequencing pattern, including seq coverage and fragmentation patterns. Therefore we located the contribution of ctDNA in cfDNA to find a non-invasive way to diagnose and detect tumour patients in real time.

Contents

1 Introduction **1**

1.1 Biology of Cell-free DNA and Circulating Tumour DNA 1

1.1.1 Cell-free DNA 1

1.1.1.1 Overview 1

1.1.1.2 History of Cell-free DNA 2

1.1.1.3 The Biological Understanding of Cell-free DNA . . 3

1.1.2 Circulating Tumour DNA 5

1.1.3 Application of Cell-free DNA Analysis in Clinical Tumour Detection 6

1.2 Nucleosomes and Gene Regulation 8

1.2.1 Overview 8

1.2.2 The History of Nucleosomes 9

1.2.3 Nucleosomes are Directly Involved in Gene Regulation . . . 9

1.2.3.1 The Structure of Nucleosomes 10

1.2.3.2 Mechanism of Dynamic Nucleosome Positioning . . 12

1.2.3.3 Nucleosome Positioning Involves in Gene Regulation 15

1.3 Nucleosomes Occupancy in Circulating Free DNA Regulation 17

1.4 Aims of This Thesis 18

2 Data Obtaining and Primary Data Processing **21**

2.1 Data Cohorts and Samples 21

2.1.1 Snyder Cohorts 22

2.1.2 The Genotype-Tissue Expression Data 22

2.1.3 The Blood Cells Expression Data 23

2.1.4 The Oesophageal Adenocarcinoma Cohort 23

2.2 Data Processing 24

2.2.1 Alignment and Formatting 24

2.2.2 Read Depth and Endpoints 25

2.2.3 Coverage and Coverage Range within the Same Region . . . 25

2.2.4 Fragments Length and Endpoints 25

2.2.5 Gene Expression Analysis 26

2.2.6 Copy Number Variation Analysis 27

2.2.7 Window Protection Score 27

3 Cell-free DNA Sequencing Reveals the Occupancy of Nucleosomes In Vivo 29
3.1 Introduction . . . 29
3.2 Results . . . 30
3.2.1 Deep Sequenced Healthy Donor Cell-free DNA Data . . . 30
3.2.1.1 Cell-free DNA Sequencing Insert Size Distribution of Deep Sequenced Non-tumour Samples . . . 30
3.2.1.2 Read Coverage Reflects the Occupancy Preference of Nucleosomes in Gene Functional Area . . . 32
3.2.1.3 The Endpoints of the Mapped Reads Presume Nucleosome Core Particle (NCP) Boundaries . . . 32
3.2.1.4 Sequencing Features Related to Gene Expression Level in Healthy Blood Cells . . . 32
3.2.1.5 Reads Spanning Over Single or Double Nucleosomes 34
3.2.2 TAPS-trial Data . . . 36
3.3 Conclusion and Discussion . . . 39

4 Tumour Patients' Cell-free DNA Sequencing Reveals Tumour Frequency in Plasma 41
4.1 Introduction . . . 41
4.2 Results . . . 42
4.2.1 Stratifying the Genes by Their Tumour-Specific Expression . 42
4.2.1.1 Sequencing Features Correlating with Tumour-Specific Expressed Genes . . . 44
4.2.1.2 Coverage Range Correlates with Tissue RNA Sequencing Data . . . 44
4.2.1.3 Coverage Around Transcription Start Sites Correlates with Gene Expression Level . . . 47
4.2.1.4 Correlation of the Tumour Frequency and the Sequencing Data . . . 47
4.2.1.5 Different Expression Level in EAC for Different Read Length for All Tumour Samples . . . 50
4.2.2 Signal Strength Correlates with Tumour Frequency . . . 50
4.2.3 Coverage and Breakpoints Combination . . . 53
4.3 Conclusion and Discussion . . . 54

5 Conclusion 57

References 59

1 Introduction

1.1 Biology of Cell-free DNA and Circulating Tumour DNA

1.1.1 Cell-free DNA

1.1.1.1 Overview

A cell's genetic information is stored in deoxyribonucleic acid (DNA) molecules. The molecule is made up of monomers called nucleotides. In eukaryotic cells, DNA is usually encapsulated in the nucleus. However, there exist circulating-free DNA fragments in the circulatory system.

Cell-free DNA (cfDNA) represents all non-encapsulated DNA circulating freely in the human circulatory system (Volik et al. 2016). Bloodstream is not the only body fluid that can be used for liquid biopsy; cerebrospinal fluid (CSF) (Rhodes, Honsinger, and Sorenson 1994), pleural fluid (Sriram et al. 2012), and asautocites (Yan et al. 2021) are all potential sources from which tumour-derived material (including cfDNA) can be obtained.

Cell-free DNA consists of double-stranded DNA fragments with various lengths. In blood, cfDNA mostly exists in the form of complexes bound to DNA and proteins, with only a few free DNA fragments. It is present in both plasma and serum. In healthy individuals, the concentration of cfDNA in plasma ranges from 1

to 10 ng·ml^{-1}. Under physiological conditions, cfDNA in blood mainly comes from apoptotic cells of the hematopoietic lineage (Sun et al. 2015).

Under certain diseases and special conditions, other cells will also release cfDNA into the blood, including circulating tumour DNA (ctDNA) derived from tumour cells and cfDNA derived from fetuses. Elevated cfDNA may also originate from tissue damage, with a large portion of the increase likely originating from specific tissue sources (Filippone and Farber 2021).

It is generally believed that DNA fragments released by cell apoptosis or necrosis are the main source of cfDNA (Jahr et al. 2001). However, the pathophysiological processes of different diseases are distinct, and the pathogenic mechanisms and disease microenvironments also vary. Therefore, the process of producing cfDNA caused by different diseases is extremely complex. A single mechanism cannot fully explain the source of cfDNA production, and these two production mechanisms often coexist.

Various disease processes can produce cfDNA, so its blood concentration is closely related to the disease process and is expected to be used as a clinical biomarker to diagnose apoptosis, tissue damage, and inflammatory processes. In clinical samples of inflammation, injury, and cancer tissue, the concentration level of cfDNA in diseased individuals is significantly different from that in healthy individuals. Elevated cfDNA levels are observed in patients with various types of cancer, with a significant increase for metastatic disease. Conversely, as a response to treatment is observed, cfDNA levels decrease (Leon et al. 1977). In patients with colorectal cancer, higher DNA levels in plasma were significantly associated with shorter survival (Schwarzenbach et al. 2008).

1.1.1.2 History of Cell-free DNA

In 1948, scientists first reported the presence of nucleic acids in human blood (Mandel 1948). In 1966, significantly elevated levels of cfDNA in plasma were observed in blood samples from patients with systemic lupus erythematosus (Tan et al. 1966). Soon after, scientists confirmed that cfDNA levels were elevated in the blood of more than half of cancer patients (Leon et al. 1977).

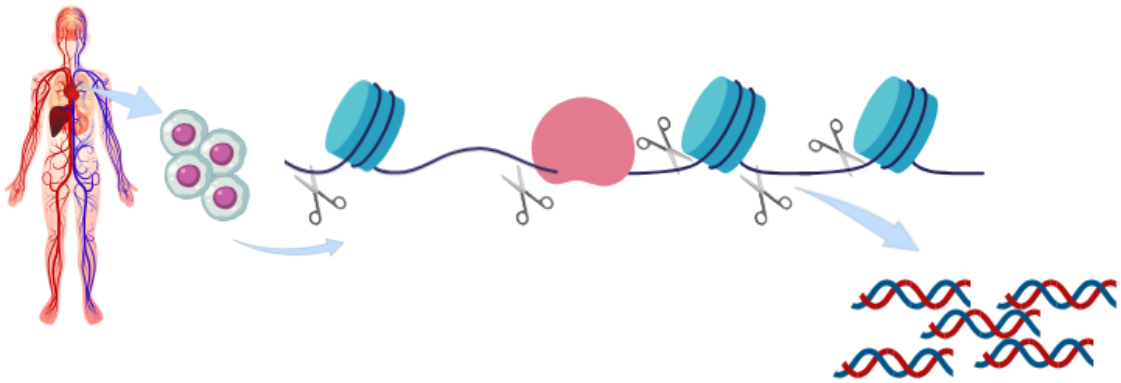


Figure 1.1: The source of cfDNA in the blood. Most cfDNA in the circulatory system originates from the apoptosis of hematopoietic cells. Cells undergoing apoptosis release genetic material into the circulatory system, and these DNA molecules are cleaved and exist in the blood in the form of fragments.

1.1.1.3 The Biological Understanding of Cell-free DNA

Understanding the biological properties of cfDNA is fundamental to its clinical use.

Previous studies related to cfDNA length distribution hint at its origin.

A main peak at 167 bp and a series of small peaks with a periodicity of 10 bp and a size of approximately 143 bp were observed, along with fragments that were integral multiples of the length of the oligonucleosomal DNA. This pattern suggests that cfDNA molecules are likely released by enzymatic cleavage during apoptosis, and the DNA is packaged with nucleosomes into the circulation (Jahr et al. 2001).

Apoptosis, a form of programmed cell death, is a significant contributor to the production of cfDNA. Unlike necrosis, which is primarily induced by hypoxia, apoptosis is an actively orchestrated cellular process triggered by a variety of physiological or pathological factors. It is distinguishable from necrosis through observation, as it involves characteristic cell shrinkage and the degradation of DNA by endonucleases into orderly 180 to 200 base pair fragments (as illustrated in Fig. 1.1). Upon cell death, there is a marked contrast between apoptosis and necrosis. In necrosis, cells swell, their membranes rupture, and their permeability changes, leading to the scattering of organelles into the intercellular space. These remnants must be cleared by macrophages to prevent inflammation of the surrounding tissue. Apoptosis is a more frequently occurring mode of cell death compared to necrosis

and is regulated by a balance of anti-apoptotic and pro-apoptotic factors. Excessive apoptosis or an impediment in the clearance of apoptotic cells can result in the release of free nucleic acids when the cell membrane ruptures. This can activate the body's innate immune system, triggering an inflammatory response. Thus, the controlled process of apoptosis plays a crucial role in maintaining tissue homeostasis and responding to pathological conditions(Bock and Riley 2023).

In addition, long fragments (more than 10 kb) were also observed, which may be related to cell necrosis (Jiang and Lo 2016). Cell necrosis begins in trauma and sepsis. In this process, cells swell and rapidly rupture, releasing a large amount of intracellular components such as free nucleic acids. This process may be the source of fragments with extended lengths of 10 kb in cfDNA.

Once released from the cell, cfDNA is easily damaged and degraded, thereby being cleared from the plasma by nucleases, with a short half-life of between 16 minutes and 1 hour (Yu et al. 2013). Most cfDNA has typical fragment sizes reflecting nucleosomal particles (~165 bp) as well as multiple nucleosomes. This may be because the DNA wrapped around the histone core is protected from digestion by apoptotic nucleases and therefore has a longer half-life than naked DNA in the blood (Ulz et al. 2016).

Accelerated apoptosis and incomplete clearance of nucleic acids leaking into the extracellular environment significantly contribute to the pathological elevation of cfDNA (Duvvuri and Lood 2019). In addition to concentration, these cfDNA fragments carry characteristic biomarkers of the cell of origin that can also be used for clinical diagnosis. Free fetal cfDNA in maternal plasma contains fetal-specific methylation levels and can be used as non-invasive prenatal diagnostic tools; tumour-derived nucleic acids have tumour tissue-specific methylation status and can be used for cancer screening and localisation; cfDNA after organ transplantation can reflect the status of the donor.

Plasma free nucleic acids are closely related to the subcellular organelle structure, chromatin structure, gene expression, and nuclease content of the source tissue. Analysis of nucleic acid fragment length, sequence, and topology can provide more

diagnostic and therapeutic information for inflammation and disease. Therefore, cfDNA can serve as a rapid real-time monitor to track genetic and epigenetic changes through non-invasive assessment (Sun et al. 2015).

CfDNA in the blood of colorectal cancer patients can be used for diagnosis and response to therapy. The blood cfDNA concentrations of 55 patients and 20 healthy donors were measured, and 14 follow-up samples are available. All patients had different cfDNA concentrations, ranging from 22 $ng{\cdot}ml^{-1}$ to 3922 $ng{\cdot}ml^{-1}$. In contrast, healthy individuals have stable concentrations. During the treatment process, the changes in cfDNA are not continuous but fluctuate, regardless of the chemotherapy drug (Leon et al. 1977).

Although high levels (>1000 $ng{\cdot}mg^{-1}$) are highly correlated with short survival. Elevated cfDNA levels in the patient's blood may indicate cellular death, making high levels of cfDNA a manifestation of a physiological or pathological process. Experimental results show that cfDNA concentration can be used for early diagnosis (Schwarzenbach et al. 2008).

1.1.2 Circulating Tumour DNA

Cancer develops through successive changes in genes that control cell growth. Many studies have shown that in solid tumours such as colon or breast cancer, an average of about 80 genes contain subtle mutations that are present in nearly every tumour cell but not in normal cells. Therefore, these somatic mutations have the potential to serve as highly specific biomarkers (Diehl et al. 2008). When this mutation is detected in body fluids, although no detectable tumour cells are found, it still shows the presence of circulating tumour-related DNA (Lecomte et al. 2002).

Circulating tumour DNA (ctDNA) is tumour-derived DNA fragments in the blood (Wan et al. 2017). In cancer, especially in advanced stages, liquid biopsies contain circulating tumour cells, as well as cell-free DNA released from circulating or primary tumour cells. Because ctDNA can reflect the entire tumour genome, it has received attention for its potential clinical utility.

The exact mechanism of ctDNA release is unknown. In addition to cell-free DNA, which may come from cell apoptosis or cell necrosis, ctDNA may also come from the active release of active tumour cells. The fragmentation and concentration of ctDNA are positively related to tumour weight. Non-tumour cells undergo less cleavage when releasing DNA into the bloodstream. The size profile of ctDNA is different from that of cell-free DNA. In ctDNA, ctDNA smaller than 80bp occupies a higher part (Mouliere et al. 2011).

Thus, in the early stages of cancer development, cells are shed from the tumour tissue into the circulation. In early tumours, nucleic acids are released into the blood or lymph through the processes of apoptosis and necrosis of primary tumour cells or acellular tumour cells. Although nucleic acid circulates in the form of nucleosomes, it retains at least some nuclear chromatin characteristics (Alix-Panabières, Schwarzenbach, and Pantel 2012).

In cancer, especially in advanced stages, elevated cell-free DNA levels can be observed due to high levels of necrosis and apoptosis in tumours. This tumour-derived cell-free DNA is often referred to as circulating tumour DNA, or ctDNA (Akca et al. 2013).

These ctDNA carry highly specific biomarkers. By analysing these tumour fragments and evaluating the sensitivity of molecular biology techniques, tumour mutations can be identified and their footprints can be tracked. Specific genomic or epigenomic markers of tumour cells can be tracked, thus obtaining prognostic and diagnostic information for cancer management (Stewart and Tsui 2018).

1.1.3 Application of Cell-free DNA Analysis in Clinical Tumour Detection

Molecular analysis of cell-free DNA has led to the important finding that plasma DNA from cancer patients contains tumour-associated mutations that could be used for cancer diagnosis and follow-up. The ability to extract circulating tumour DNA (ctDNA) from human plasma has led to huge advances in non-invasive cancer detection. Most notably, it led to what is now called liquid biopsy.

Liquid biopsies provide information about tumour-specific biomarkers and cancer cells in the blood, as a means of diagnosing the type and stage of cancer. This type of biopsy is non-invasive and allows for routine clinical screening, which is important for identifying cancer recurrence after initial treatment. Biomarker-based tumour diagnosis is typically obtained from tissue samples through invasive surgery. However, in some cases, a tissue biopsy cannot be performed. Therefore, a non-invasive means of obtaining a biopsy is urgently needed (Jiang et al. 2023).

Liquid Biopsy is now widely used in the management of cancer patients (Alix-Panabieres, Marchetti, and Lang 2023). The main components of liquid biopsy are cell-free DNA, Circulating Tumour Cells (CTCs), and exosomes. Cell-free DNA and ctDNA have excellent tumour screening potential.

Prognostic information in genomic profiling does not require tumour tissue biopsy. A lot of evidence shows the perioperative dynamics of liquid biopsy. To provide a potential understanding of the complex relationship between surgery and therapy outcome, careful examination of liquid biopsy has been determined and provided important diagnostic or therapeutic information in the perioperative window (Jordan et al. 2022).

Potentially tumour-specific, through quantitative detection or special sequencing, it has been applied in many diseases, so cancer also has potential (Yan et al. 2021). The detection of cell-free DNA in plasma or serum can be regarded as a "liquid biopsy", which is beneficial to carry out various related diagnoses and avoids invasive tissue biopsy. Using blood as a test sample makes it possible to collect test samples repeatedly, which is conducive to tracking and monitoring the development of the disease and the treatment process of the tumour.

Many publications have demonstrated that the non-invasive nature of cfDNA analysis makes it particularly valuable for patients who may not be suitable for traditional biopsy methods. Diehl (Diehl et al. 2008) provided insights into the clinical implications of mutations in cfDNA from patients with metastatic colorectal cancer. Murtaza (Murtaza et al. 2013) identified the potential of cfDNA analysis in the management of cancer treatment. Newman (Newman et al. 2014) investigated

the tracking of therapy resistance and tumour evolution using cfDNA in non-small-cell lung cancer. Zardavas (Zardavas et al. 2015) discussed the clinical utility of liquid biopsies in breast cancer. Chan (Chan et al. 2017) provided an overview of the clinical impact of liquid biopsies in advanced cancers. Cai (Cai et al. 2018) demonstrated that ctDNA can be used as a biomarker to track tumour dynamics in real time. Lung (Lung et al. 2020) explored the role of cfDNA in early detection and monitoring of cancer.

1.2 Nucleosomes and Gene Regulation

1.2.1 Overview

Nucleosomes are the basic repeating units that make up eukaryotic chromatin.

The diameter of the human cell nucleus is only $5\mu m$, and the length of the longest chromosome is only $10\mu m$, but the length of the DNA molecule is $7.2cm$. In such a small space, DNA with a total length of nearly 200 centimeters can be accommodated. This also means that, in order to assemble into the nucleus, the DNA must be compacted in an extremely efficient manner. It is important to understand how DNA forms chromosomes and fits into the tiny nucleus of a cell. In addition to winding, eukaryotic chromatin also adopts a series of complex structures, so that double-stranded DNA molecules eventually form compact chromatin, or even chromosomes. Each human cell contains approximately 30 million nucleosomes. The length of DNA coiled outside the core of the histone octamer is not always uniform. The general variation range is between 180 bp and 200 bp. Of these 200 bp, 146 bp coil directly outside the histone octamer core. This DNA is not easily digested by nucleases, and the rest of the DNA is used to ligate the next nucleosome. Nucleosomes are shaped like a flat plate or cylinder. Thus, the length of DNA is compressed to 7 times the length of chromatin fibers. Chromatin is composed of a series of nucleosomes. DNA has a compact packing ratio of about 40 when a series of nucleosomes are helically arranged to form beaded filaments. When the filaments themselves are further compressed into euchromatin, the DNA has a compression packing ratio of about 1000. Each chromosome is thus a long

chain of nucleosomes, which, when viewed with an electron microscope, takes on the appearance of a string of beads. During mitosis, chromatin is further compressed into chromosomes. The compression accumulation rate is as high as 8400, which is only one ten thousandth of the length in the stretched state, and can be packed into the nucleus (Alberts et al. 2002).

1.2.2 The History of Nucleosomes

In 1956, Wilkins(Wilkins 1956), who provided X-ray diffraction evidence for the double-helix model, and Vittorio Luzzati (Luzzati and Nicolaieff 1963) conducted X-ray diffraction studies on chromatin and found that chromatin has a repeating structure with an interval of $10nm$. Based on this, Luzzati speculated that the combination of histones and DNA forces the DNA to fold or wind into a repeating structure with a period of $10nm$.

In 1974, Don and Ada Olins first observed nucleosomes as granules under the electron microscope (Olins and Olins 1974), and Roger Kornberg proposed their existence and structure as histone octamers (Kornberg 1977), (Kornberg and Lorch 1992). The role of nucleosomes has been demonstrated *in vitro* as a transcriptional regulator in 1987 (Han and Grunstein 1988) and *in vivo* in 1988 (Clark-Adams et al. 1988).

1.2.3 Nucleosomes are Directly Involved in Gene Regulation

Nucleosome position is involved in gene regulatory processes in many cells, and is crucial for cell differentiation as well as the response of tissues and organs to microenvironmental changes associated with disease (Teif and Clarkson 2019).

The position of the nucleosome is not random and exhibits specific cellular characteristics. This characteristic reflects the fact that nucleosomes are involved in regulating biochemical reactions in cells by limiting the contact between regulatory factors and nucleic acid molecules through steric hindrance effects. Short strands of circulating free DNA are produced by the digestion of chromatin by apoptotic nucleases, which tend to cleave dead cell DNA between nucleosomes. Therefore,

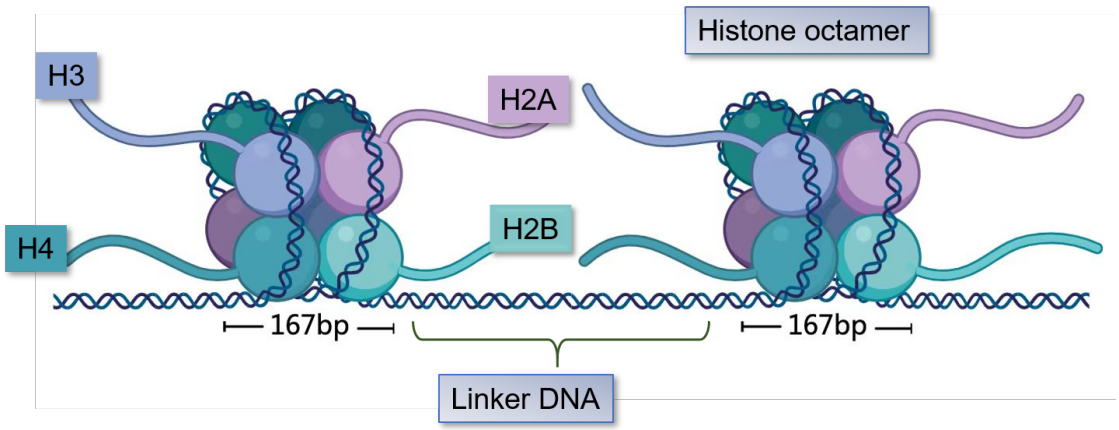


Figure 1.2: Nucleosomes are the fundamental units of eukaryotic chromatin, consists of a histone octamer with DNA wrapped around it. The nucleosome core particles are connected by linker DNA, and histone H1 binds at the entry of the linker DNA.

circulating free DNA in body fluids can reflect the location of the nucleosome in the nucleus of apoptotic cells(Shtumpf et al. 2022).

1.2.3.1 The Structure of Nucleosomes

As the main packaging unit of DNA in the nucleus, nucleosomes significantly impact the accessibility of DNA.

X-ray diffraction experiments were performed on the core particles of chromatin nucleosomes to obtain their crystal structures (Luger et al. 1997).

The nucleosome consists of approximately 1.75 left-handed superhelical turns with a 200 bp fragment of DNA wrapped around a histone octamer. Each histone octamer is composed of two sets of H2A and H2B dimers and H3 and H4 dimers. DNA molecules wrap around the periphery of the histone octamer to form the core particle of the nucleosome (Fig. 1.2). In addition, the nucleosome core particles are connected through a linker region connecting approximately 50 bp of DNA. Histone H1 binds at the opening of the linker DNA duplex coiled on the octamer and binds to the DNA molecule linking two adjacent nucleosomes. Histone H1 consists of a group of closely related proteins that are only half as numerous as core histones and therefore can be easily extracted from chromatin. When all H1 is removed, the nucleosome structure remains unaffected, indicating that H1 is located outside the protein core (Olins and Olins 1978; Kornberg 1977).

The interface between DNA and histones is extensive.

The histone folding domains directly describe the organisation of 121 base pairs of DNA. Each histone fold pair, whether half of an H3-H4 tetramer or half of an H2A-H2B dimer, is associated with 27-28 base pairs of DNA, with a 4 base pair linker between these units. When a DNA molecule interacts with a histone octamer, binding occurs primarily on the DNA phosphodiester backbone. The DNA strand faces the protein within 2.5 turns of the double helix. Each helical turn of the chain binds only two consecutive segments of phosphate groups to a histone fold pair. The length of bound DNA is extended in both directions by additional backbone segments containing a loop that connects the salt to the phosphate group. Extended DNA sites overlap from one histone fold pair to the next. The amino-terminal tails of histones pass through the turns of the DNA supercoil to contact neighbouring particles. The lack of coherence between multiple histone/DNA binding sites causes DNA to deviate from the ideal superhelical geometry. Sequence-dependent anisotropic flexibility, as well as the intrinsic curvature of DNA (resulting in the rotational setting of DNA segments on the histone core), are important determinants of the allowed octamer positions(Pennings et al. 1989).

Hydrophobic interactions and salt bonds anchor DNA and proteins within nucleosomes(Davey et al. 2002). All core histones are rich in lysine and arginine and are highly positively charged due to the predominance of bases, whose positive charge effectively neutralises the negatively charged DNA phosphate backbone. The long, disordered polymer tail domain contributes to the thermal stability of the nucleosome core due to the abundance of positive charges. Hydrophobic interactions thus explain the ability of virtually any DNA sequence to bind to the histone octamer core(Ausio, Dong, and Van Holde 1989).

Nucleosomes avoid promoter and termination regions. After the transcription start site (TSS), the arrangement of the nucleosome has a canonical distance. The TSS is located at the nucleosome boundary and must compete with the +1 nucleosome before the transcription machinery can be initiated. In metazoans, the TSS is located in the nucleosome-free region, suggesting that RNA polymerase II

competes with the first nucleosome after initiation. Periodic sequences of genes, such as AA or TT, will accelerate nucleosome formation, while consecutive adenine-thymine pairs will create nucleosome-free regions. These sequences are scattered but form a landscape of nucleosome dynamics. At the 5' end of genes, nucleosomes will be enriched, and during post-transcriptional modification, nucleosomes will be more dynamic. The TSS of many genes is directly associated with the first nucleosome downstream of the promoter nucleosome-free region(Jiang and Pugh 2009).

Linker DNA outside the nucleosomal histone octamer core moves back and forth between the two stacks of the nucleosome core, forming a cleaved two-headed helix and does not follow a path compatible with a one-headed solenoid helix. The length of the linker DNA is most likely relaxed by the elongation of the DNA contained in the nucleosome core. A continuous fibre model of the tetranucleosomes interface between sequentially stacked nucleosomes along the starting point of a single helix is polymorphic(Schalch et al. 2005).

Histone H1 alters the conformational landscape of the nucleosome by linking two stretches of linker DNA together and reducing its flexibility. The C-terminal domain (CTD) of H1 localises primarily to a single linker, whereas the H1 globular domain contacts the nucleosomal duplex and both linkers, binding more tightly to the CTD-distal linker. These results indicate that nucleosomes possess strong asymmetry and that this conformation may influence the assembly and structure of higher-order structures(Bednar et al. 2017).

1.2.3.2 Mechanism of Dynamic Nucleosome Positioning

Recombination, replication, mitotic condensation, and transcription involve chromatin substrates and are therefore affected by their structure (Alberts et al. 2003). As the primary unit of chromatin, nucleosomes are associated with histones involved in packaging DNA and with many proteins required for gene expression, DNA replication, and DNA repair processes.

In principle, every DNA sequence can fold into a nucleosome, but the array of nucleosomes in cells is irregular. In eukaryotes, histone genes and nucleosome

structures are extremely conserved. In vitro experiments show that nucleosomes are remodelled on linear or circular DNA fragments hundreds of base pairs in length through certain chromatin remodelling complexes. As cation concentration increases, nucleosome arrays reversibly self-assemble into oligomeric structures *in vitro* rather than forming 30 nm fibres in the absence of any other chromatin components (Maeshima et al. 2016).

During the process of compressing the DNA double helix, A-T rich sequences are more easily compressed than G-C rich sequences, so histones tend to be positioned on DNA with the lowest potential energy, maximising the A-T content inside the DNA coil. Therefore, DNA fragments containing short A-T rich sequences separated by an integer number of DNA turns are thermodynamically more likely to bend around nucleosomes than DNA fragments lacking this feature. Furthermore, because the DNA in the nucleosome is kinked at multiple locations, the ability of a given nucleotide sequence to accommodate this deformation also affects the position of the DNA on the nucleosome. The intrinsic affinity of histones to DNA regions ensures nucleosome positioning. This may explain the very strong localisation of nucleosomes to DNA sequences. However, there is no strong preference for nucleosome binding among DNA sequences abundantly distributed in chromatin *in vivo*. The position occupied by nucleosomes on DNA is relatively random and can occupy any one of multiple positions in the DNA sequence (Shtumpf et al. 2022).

In vivo, nucleosomes are not just static modules of chromatin structure; neither are nucleosome positions constant. On the contrary, nucleosomes are dynamic within the nucleus. Initial nucleosome assembly proceeds by salt dialysis in a multistage process determined by the initial recruitment of histone octamers. However, in various cellular activities involving nucleosomes, there are other tightly bound proteins on the DNA, and histones are constantly disassembled and reassembled onto the DNA template in a site-specific manner. During these repeated activities, histones are sometimes replaced by variants. These histones carry a large number of post-translational modifications, and these post-translational covalent modifications affect histone-DNA interactions, histone-histone interactions,

and interactions between histones and regulatory factors (Bai and Morozov 2010). DNA-binding proteins also affect the assembly of nucleosomes adjacent to them to a certain extent. Some proteins occupy the binding sites between nucleosomes and DNA, forcing nucleosomes to leave, while others help recruit histones (Singh and Mueller-Planitz 2021).

Chromatin remodelling is an enzyme-assisted process that promotes nucleosomal DNA entry by remodelling the structure, composition, and positioning of nucleosomes. Through the interaction between DNA molecules through histone octamers, the genetic code contained in DNA is tightly packaged in the nucleus to form repeating units of nucleosomes. These repeating units are further bundled together to form a dense chromatin structure. Moving nucleosomes changes the boundary conditions to make room for other nucleosomes, even if they are not the nearest neighbours (Teif and Rippe 2009).

In many cases, the repositioning of just one or two nucleosomes provides a critical step in activating gene expression from a specific promoter. This condensed structure blocks many DNA regulatory regions, preventing them from interacting with transcription machinery proteins and regulating gene expression. To overcome this problem and allow dynamic access to condensed DNA, a process called chromatin remodelling alters nucleosome structure to expose or hide regions of DNA for transcriptional regulation (Radman-Livaja and Rando 2010).

The ATP-consuming nucleosome remodeler is involved in the structural reorganization of nucleosomes, including nucleosome sliding and DNA site exposure (Becker and Hörz 2002). ATP-related nucleosome position and sliding play an important role in regulating the relative position of nucleosomes to each other and the mutual position of nucleosomes to DNA (Becker and Workman 2013). In addition to their common involvement in transcriptional regulation, these ATP-dependent chromatin remodelling complexes also have unique functions in other cellular processes. Some nucleosome remodelers can sense the length of the linker DNA outside the histone octamer and thereby adjust the histone octamer to equalize the length of the DNA at both ends of the nucleosome. Due to the input of energy, this process usually

takes precedence over interactions between DNA sequences and histones through hydrophobic interactions (Yang et al. 2006).

The primary role of some remodelling complexes is to allow access to nucleosomal DNA, while others are primarily involved in re-forming nucleosomes when access to DNA is no longer needed (De Dieuleveult et al. 2016). While some remodelling complexes organise nucleosomes into appropriate bundles and create equal spacing between nucleosomes, there are also remodelling complexes that disrupt nucleosomes. Remodelling complexes have been shown to play a central role in chromatin assembly and maintenance of advanced chromatin structure after DNA replication, participate in DNA double-strand break repair and nucleotide excision repair, and play a crucial role in DNA damage response. Remodelers also mediate transcriptional repression in the nucleus and maintain embryonic stem cell pluripotency.

1.2.3.3 Nucleosome Positioning Involves in Gene Regulation

Nucleosomes are thought to carry epigenetic information through covalent modifications of their histone cores.

The position of nucleosomes in the genome is not random; therefore, knowing the position of each nucleosome can determine its role in gene regulation. The exact location where a nucleosome binds to DNA depends on multiple factors, including the DNA sequence and the presence and nature of other proteins that bind to the DNA. This makes the arrangement of nucleosomes on DNA highly dynamic, changing at any time according to the needs of the cell (Radman-Livaja and Rando 2010).

To achieve a high degree of compaction, most genomic DNA in eukaryotes is integrated into nucleosomes; however, regulatory factors and transcriptional machinery must have access to chromatin to extract genetic information. This conflict can be partially resolved by specific arrangements of nucleosome positions on the genome. In all eukaryotic species, promoters and other regulatory sequences are more depleted of nucleosomes, whereas transcribed regions tend to be occupied by high-density arrays of well-positioned nucleosomes. This nucleosome positioning pattern and its dynamic regulation help transcription factors reach their target

sites and play a crucial role in determining transcription levels, cell-to-cell variation, and activation or repression dynamics (Bai and Morozov 2010).

Therefore, nucleosome positioning is directly involved in gene regulation by determining the access of transcription factors (TFs) to the promoters and the RNA polymerases to the gene body. Occupancy by nucleosomes in functional regions such as promoters or enhancers is correlated with gene silencing (Lorch, LaPointe, and Kornberg 1987).

Nucleosome sliding is one of the possible mechanisms for large-scale tissue-specific expression of genes. When tissue-specific genes are expressed in a specific tissue, the nucleosomes at their transcription start sites are missing, while for the same group of genes that are not expressed in other tissues, their transcription start sites are occupied by nucleosomes. Nucleosome positioning maps of several organisms indicate that the TSS is marked by a nucleosome-depleted region flanked by strongly positioned nucleosomes. Furthermore, the nucleosome is positioned at the TSS of silenced genes, while it is displaced to expose the TSS in highly expressed genes, which indicate that gene expression levels in vivo are determined by the accessibility of the TSS(Bargaje et al. 2012).

Chromatin remodelling plays a significant role in the regulation of gene expression by providing the transcription machinery with dynamic access to the densely packed genome(De Dieuleveult et al. 2016). Furthermore, nucleosome movement by chromatin remodellers is critical for several important biological processes, including chromosome assembly and segregation, DNA replication and repair, embryonic development and pluripotency, and cell cycle progression. Dysregulation of chromatin remodelling results in loss of transcriptional regulation of these critical checkpoints required for normal cell function, leading to various disease syndromes, including cancer.

Increasing evidence shows that epigenetic changes, which describe perturbations outside the genome, are clearly related to tumour development. Dynamic chromatin remodelling of the basic nucleosome structure, including dynamic interactions between ATP-dependent complexes, covalent histone modifications, utilisation of

histone variants, and DNA methylation, is fundamental for many cellular processes, including transcriptional regulation and DNA damage repair. Dysregulation of these processes has been implicated in the development of cancer(Wang, Allis, and Chi 2007).

1.3 Nucleosomes Occupancy in Circulating Free DNA Regulation

Extensive research has been conducted to understand the intricate relationship between nucleosome positioning and gene regulation, with published studies utilizing micrococcal nuclease (MNase) assays to generate mononucleosome-bound DNA fragments. These studies have successfully mapped the global nucleosome occupancy, providing valuable insights into the specific nucleosome patterns that occur at promoters(Pedersen et al. 2014).

Promoters of actively expressed protein-coding genes are characterized by a nucleosome-depleted region (NDR) at the TSS, which is crucial for the binding of transcription factors and the initiation of transcription. Additionally, these promoters are associated with a strongly positioned +1 nucleosome and a downstream nucleosome phasing, which together contribute to the regulation of gene expression. The 150 base pairs upstream of the TSS region are particularly important as they enhance the accessibility of DNA to the DNA-dependent RNA polymerase. This polymerase is responsible for transcribing protein-coding genes, and the well-positioned nucleosomes upstream and downstream of the TSS play a pivotal role in this process. Moreover, a reduction in total mapped reads coverage is observed around actively expressed gene bodies, with this reduction extending up to 5 kb. This suggests a general decrease in nucleosome occupancy in these regions, which may be a strategic adaptation by the cell to facilitate gene expression(Schones et al. 2008; Yuan et al. 2005; Barski et al. 2007).

1.4 Aims of This Thesis

As described above, cfDNA has revolutionized the clinical landscape, particularly in non-invasive prenatal testing, cancer diagnosis, and treatment monitoring. In the context of cancer, cfDNA can be shed by both healthy and tumour cells, making it a valuable biomarker for detecting cancer-related genetic alterations. Its clinical applications include the identification of actionable mutations for personalized therapy, monitoring tumour burden and treatment response, as well as detecting minimal residual disease and recurrence post-treatment.

Nucleosome profiling is a powerful technique for understanding the three-dimensional organization of the genome and how it influences gene expression. It involves the analysis of nucleosome positioning, occupancy, and modification across the genome. This method has broad applications in genomics, including the identification of functional genomic elements, such as promoters, enhancers, and transcription factor binding sites. Nucleosome profiling has also been instrumental in studying the epigenetic changes associated with development and disease states. For instance, it can reveal the chromatin architecture changes during cellular differentiation or in response to environmental stimuli. Furthermore, in the context of liquid biopsies, nucleosome-based profiling has emerged as a promising approach for detecting and characterizing circulating nucleic acids in body fluids, offering insights into the tumour genome and its dynamic changes over time

The primary objective of this thesis is to delve into the wealth of information contained within cfDNA sequencing data. We aim to utilize this data to determine the distribution of nucleosomes within apoptotic cells, thereby inferring the expression patterns of genes regulated by these nucleosomes. Our approach involves a comprehensive analysis of the spatial arrangement of nucleosomes and their occupancy at specific genomic regions.

By identifying tumour-specific expression genes, we hope to pinpoint the contribution of circulating tumour DNA within the broader context of circulating free DNA. This could potentially lead to the development of a non-invasive diagnostic tool capable of monitoring tumour patients in real time. Such a

tool would revolutionize the early detection and ongoing management of cancer, providing medical professionals with a powerful new resource in the fight against this devastating disease.

2

Data Obtaining and Primary Data Processing

2.1 Data Cohorts and Samples

The datasets used in this manuscript are divided into two groups, “Healthy”(or “Non-tumour”) and “Tumour”, based on the pathological status of the sources. The cell-free DNA (cfDNA) sequencing and RNA sequencing datasets for the Healthy group come from different datasets, which will be described in detail below, while the cfDNA and RNA sequencing for the Tumour group both come from the same clinical trial(Fig. 2.1).

After preprocessing and filtering, both cfDNA datasets underwent fragmentomics and coverage analysis, while the RNA sequencing dataset was used to obtain the expression levels of genes in different tissues and cells, thereby stratifying and classifying the genes.

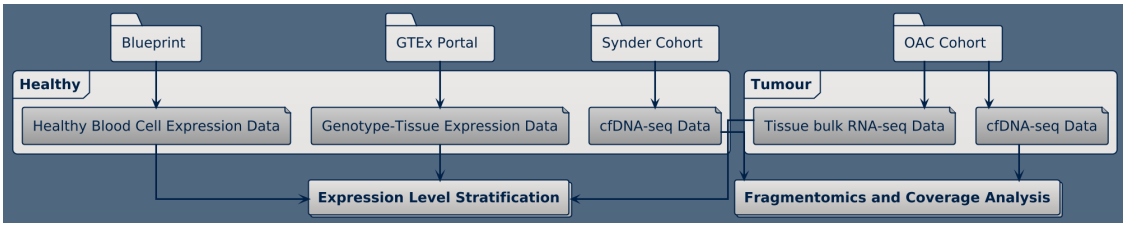


Figure 2.1: The sources of datasets and their analysis processes

To ensure transparency and reproducibility, we have selected the following data cohorts for our analysis:

2.1.1 Snyder Cohorts

The Synder dataset (Snyder et al. 2016) is a deeply sequenced cfDNA dataset that utilizes a large volume of human peripheral blood plasma contributed by an unknown number of healthy individuals. The sequencing libraries was prepared by performing end-repair and adapter ligation on cfDNA fragments purified from the plasma. Subsequently, these libraries were sequenced to achieve a coverage depth of 96×, which corresponds to 1.5 billion (G) sequenced fragments.

The Snyder dataset was selected as it was the most deeply sequenced data available at the time of our study. Its was used for similar purposes made it a suitable positive control and an invaluable resource for our training dataset. The depth of sequencing in the Snyder dataset allows for a more comprehensive analysis and serves as a benchmark for comparing and validating our findings.

For the Snyder Cohorts, we obtained the datasets from the Gene Expression Omnibus (GEO) Accession: GSM1833219. These datasets serve as the healthy control group, providing a baseline for comparison with the disease cohorts.

2.1.2 The Genotype-Tissue Expression Data

The Genotype-Tissue Expression (GTEx) project (Consortium 2015) is a comprehensive research initiative aimed at exploring the genetic basis of gene expression in humans. It endeavors to establish a comprehensive public resource that connects genetic variation to gene expression, providing valuable insights into the molecular mechanisms that underpin human health and disease. The GTEx project has involved the analysis of gene expression in a wide array of human tissues, offering a detailed view of how genetic variations influence gene expression across different tissues. By examining postmortem human tissue samples from hundreds of individuals, GTEx has been able to identify significant correlations between genetic variants

and gene expression levels. This wealth of data has been made publicly available to facilitate research into the genetic components of various diseases and conditions.

The Genotype-Tissue Expression data used for the analyses described in this manuscript were obtained from the GTEx Portal (dbGaP accession number phs000424.v8).

2.1.3 The Blood Cells Expression Data

The Blueprint database (Fernandez et al. 2016) is a valuable resource for researchers studying epigenomics, which is the study of epigenetic modifications to DNA and associated proteins. It is particularly focused on providing data related to the epigenomes of hematopoietic cell types. The Blueprint Data Analysis Portal (BDAP) offers a user-friendly interface for the comparative analysis of epigenomes, allowing researchers to explore genomic regions, genes, and pathways in the context of hematopoietic lineage differentiation The Blueprint database includes a comprehensive collection of epigenetic features and provides normalized expression values for a significant number of RNA sequencing samples from pure populations of stroma and immune cells.

Therefore, We fetched the RNA sequencing data from the Blueprint project, selecting specific cell types to pool together as healthy plasma RNA sequencing data. This data serves as a reference for gene expression levels in healthy blood cells.

The blood cells RNA sequencing data was also adopted from Blueprint, and we selected 5 cell types which contribute most to cfDNA in plasma and pooled them together as healthy blood cells RNA sequencing data. These 5 types of cells are: Granulocyte monocyte progenitor cell, Megakaryocyte-erythroid progenitor cells, monocyte common lymphoid progenitor,CD14-positive and CD16-negative classical monocyte .

2.1.4 The Oesophageal Adenocarcinoma Cohort

The oesophageal adenocarcinoma (OAC) cohort is derived from the LUD2015-005 clinical trial, which was a phase I/II study that tested durvalumab as a first-line treatment for patients with inoperable oesophageal or gastro-oesophageal cancer.

The trial included patients who had not previously received systemic anti-cancer therapy for their advanced disease and had a good performance status. After an initial period of treatment with these drugs, all participants then received chemotherapy along with continued durvalumab until the end of their chemotherapy cycles. The trial was designed to assess the safety of the treatments and gather preliminary data on their effectiveness, rather than to provide conclusive results.

During this period, participants will provide plasma and whole endoscopic biopsies at various stages of the tumour for the construction of cfDNA libraries and bulk tissue RNA sequencing libraries. Therefore, the OAC dataset offers an individual's corresponding cfDNA and tissue expression data at different stages of the disease.

This cohort represents a critical dataset for our analysis due to its relevance to our research questions and the quality of the data provided. It was utilized in our study because we aimed to determine whether tumour-specific gene expression, potentially prognostic for response to therapy, could be detected using the fragmentomic approach.

We acknowledge that there are numerous publicly available datasets, and our choice was influenced by the specific research objectives of this study. A comparison of our chosen datasets with others was conducted based on factors such as sample size, disease representation, and data quality.

2.2 Data Processing

2.2.1 Alignment and Formatting

Fragments shorter than or equal to the read length were consensus-called and adaptor-trimmed. The remaining consensus single-end reads (SR) and individual pair-end (PE) reads were aligned to the human reference genome (GRCh38, 1000 Genomes phase 2 technical reference) using the ALN algorithm in BWA v0.7.17-r1188. Aligned SR and PE data were stored in BAM format using Samtools 1.14 (utilizing htslib 1.14). BAM files for each sample were merged across lanes and sequencing runs to ensure comprehensive data integration.

2.2.2 Read Depth and Endpoints

A fragment's coverage is defined as all positions between the two (inferred) fragment ends, inclusive of endpoints. We extracted per base depth using Samtools depth and reads endpoints using Samtools mpileup. At this stage, we only retained reads mapped in proper pair to ensure accuracy in our subsequent analyses.

2.2.3 Coverage and Coverage Range within the Same Region

We selected TSS±1kb as the observation area. Coverage is defined as the sum of per base depth within this region.

$$\text{coverage} = \sum_{\text{start site}}^{\text{end site}} \text{depth} \tag{2.1}$$

Coverage range is defined as the range of per base depth in this region.

$$\text{coverage_range} = \text{Max(depth)} - \text{Min(depth)} \tag{2.2}$$

We performed Spearman correlation tests on the seq data of each gene and its expression level. Data points where coverage or range equals zero are omitted.

2.2.4 Fragments Length and Endpoints

The fragment length is defined as the signed observed Template Length (TLEN). For primary reads where all reads in the template are mapped to the same reference sequence, the absolute value of TLEN equals the distance between the mapped end and the mapped start of the template, inclusively (i.e., end − start + 1).

The TLEN field is positive for the leftmost segment of the template and negative for the rightmost. For reads with negative TLEN, the start is at read start − (mapped read length) + TLEN. For reads with positive TLEN, the start is at read start.

2.2.5 Gene Expression Analysis

We obtained 22,836 Ensembl genes identified as protein-coding genes and transcript starting sites, adjusting for strand and direction of transcription. For genes with multiple Transcription Starting Sites (TSSs), we selected the first TSS of the gene body as the representative starting point.

We extracted per gene TPM across the samples and calculated their mean as the gene expression level in healthy blood cells. We sorted genes by their expression TPM order, stratified the genes into 4 categories by their TPM. The top expressed genes are the first 25% genes, the high category is 25%-50%, median category are 50%-75%, and the last 25% genes as low expressed.

We used solid tumour tissue RNA sequencing data from OAC cohort to calculate the per gene TPM and calculate their mean as the gene expression levels in tumour.

We define genes with TPM equal to 0 across all samples as unexpressed in the group, and genes with mean TPM larger than 10 as expressed genes in the group. If a gene shows a significant difference between the samples within the same group (Student's t-test p value > 0.1), we define it as a groupwise abnormally expressed gene. After removing genes that were abnormally expressed in samples within their group, we classified the remaining genes into three expression patterns: all expressed, all unexpressed, and tumour-specific expressed.

To group genes into tissue-specific groups, we compared differences in expression of the same gene between two groups of tissues. To ensure that genes are stably expressed in different cells in each group, we eliminated abnormal expression in a single cell type within the same group. By comparing the expression differences of genes in healthy blood cell groups and tumour groups, we divided all genes into four sets: expressed in both groups of tissues (all expressed), only expressed in tumours (tumour only), not expressed in both groups of tissues (unexpressed), and others. Here, since we focused on the tumour-only gene set.

2.2.6 Copy Number Variation Analysis

The prediction of the tumour fraction in the tumour sample cfDNA dataset was performed by the Copy Number Variation (CNV) method (Adalsteinsson et al. 2017). The analysis with ichorCNA(0.3.2) was performed using ultra-low-pass whole genome sequencing (ULP-WGS) data, with window sizes set to either 100kb or 1MB to capture the genomic landscape of copy number alterations (CNAs). Prior to running ichorCNA, a GC-content bias correction using HMMcopy has been conducted, which computed read coverage and normalized the data to account for sequencing depth variations. ichorCNA was then applied to the corrected ULP-WGS data to estimate the tumour fraction, defined as the proportion of cfDNA that is tumour-derived.

2.2.7 Window Protection Score

The windowed protection score (WPS) of a window of size k was defined as the number of molecules spanning the window minus those with an endpoint within the window(Snyder et al. 2016). We assign the determined WPS to the center of the window with a width of 120bp.

3

Cell-free DNA Sequencing Reveals the Occupancy of Nucleosomes In Vivo

3.1 Introduction

Nucleosomes play a crucial role in gene regulation by controlling the access of transcription factors to promoters and RNA polymerases to the gene body. Their occupancy in functional regions such as promoters or enhancers is closely linked to gene silencing. Capitalising on this, we utilised cfDNA sequencing to detect nucleosome footprints and glean insights into the transcriptome of human tissue cells. We hypothesised that the genomic distribution of cfDNA could mirror nucleosome positions, thereby reflecting alterations in the dynamic epigenetic microenvironment (Fig. 3.1).

To empirically test this hypothesis, we employed a deep-sequenced cfDNA dataset from healthy individuals, which exhibited a high correlation with haematopoietic blood cells. Our approach involved the following steps: Firstly, we analysed the size distribution of cfDNA fragments, identifying a predominant peak at 167bp, aligning with the size of a histone core This suggested that cfDNA fragments are protected by nucleosomes Secondly, we examined the read coverage distribution around transcription start sites (TSS), noting a significant drop in the -150bp to +50bp region, indicative of a nucleosome-depleted region (NDR). Thirdly, we

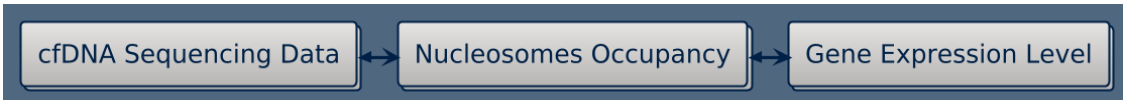


Figure 3.1: cfDNA sequencing data can reflect nucleosome occupancy, which in turn is closely related to gene expression level, thus there is a connection between the three.

analysed the endpoints of mapped reads, observing a clustering near nucleosome core particle boundaries and a reduction within the NDR. Fourthly, we stratified genes by their mean expression levels across different blood cells and analysed sequencing features such as read depth and insert size. Fifthly, we extracted reads of specific lengths and assessed their contributions to the coverage distribution, finding that reads corresponding to dinucleosome sizes provided a clearer signal at the NDR. Lastly, we validated our model on samples from individual tumour and non-tumour patients, obtaining consistent results.

These findings substantiate our hypothesis that cfDNA distribution is reflective of nucleosome positioning and related to gene expression levels. The consistency of our results across different sample preparations confirms the reproducibility of our observations. Moreover, the cfDNA signature of tumour samples aligns with solid tumour gene expression, highlighting the potential of cfDNA as an indicator of tissue-specific gene expression.

3.2 Results

3.2.1 Deep Sequenced Healthy Donor Cell-free DNA Data

3.2.1.1 Cell-free DNA Sequencing Insert Size Distribution of Deep Sequenced Non-tumour Samples

To explore the current landscape of cfDNA, we first used a previously published cfDNA sequencing dataset. This conventional sequencing library was prepared by end repair and adaptor ligation to cfDNA fragments purified from plasma pooled from an unknown number of healthy individuals with 96-fold coverage(Snyder et al. 2016).

Since cfDNA is considered to be wrapped around the histone core, we first examined the size of the fragment isolated from plasma. The size profile of cfDNA

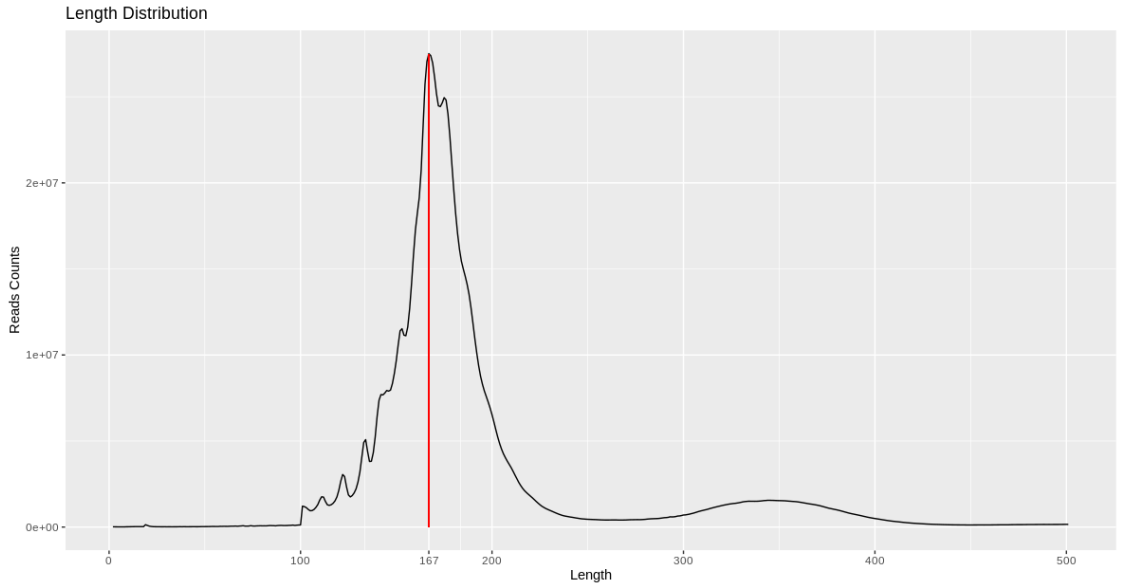


Figure 3.2: Reads length distribution of the healthy samples, only reads mapped in proper pair. A primary peak in fragment length at 167 bp (red line) is consistent with the association with the size of nucleosome core particles. Several groups of small peaks with a period of about 10 bp are observed.

from mixed healthy donors demonstrated a predominant peak around 167bp, which is the same size as a histone core. The peaks are consistent with a nucleosome-protected region. The endpoints of the mapped reads may indicate nucleosome core particle (NCP) boundaries. These size distribution profiles confirm that cfDNA molecules can be strong evidence pointing to the position of nucleosomes. There is a secondary peak at 300–400bp, which may suggest a double nucleosome dimer, consisting of 2 nucleosome core particles and linker DNA. In addition, we observed several groups of small peaks with a period of about 10 bp, which are consistent with the helical pitch of DNA wrapping on the nucleosome core. (Fig. 3.2)

This length distribution strongly supports a previously proposed model: nucleic acid molecules wrapped around nucleosomes are protected, and cleavage often occurs between two nucleosomes. At the same time, there is also DNA helix-related cleavage associated with nucleosome binding.

3.2.1.2 Read Coverage Reflects the Occupancy Preference of Nucleosomes in Gene Functional Area

The coverage distribution of reads exhibits a strong signal within regions where nucleosomes are positioned, as determined by in vitro methods (Gaffney et al. 2012).

This coverage also correlates with genomic features, such as transcription start sites (TSS). We counted reads within 2 kilobases around the TSS of the first transcript of protein-coding genes. Aggregated coverage at ±1 kilobase around TSS reveals inferred positions of nucleosomes upstream and downstream of TSS. Meanwhile, a significant drop within the -150 base pairs to +50 base pairs range has been observed, suggesting a nucleosome depletion region (NDR) at this region. The average insert length (TLEN) reflects reads inferring the given position. A peak at the NDR region is observed, indicating long fragments spanning over the NDR (Fig. 3.3A,B).

3.2.1.3 The Endpoints of the Mapped Reads Presume Nucleosome Core Particle (NCP) Boundaries

We then examined the endpoints of the mapped reads. As expected, the aligned fragment endpoints clustered adjacent to NCP boundaries and were significantly reduced in the NDR area. However, downstream of the NDR region, indicative of an NCP, a sharp peak was observed, demonstrating a well-positioned cleavage site. Concurrently, all three characteristics showed clear periodicity around the size of the NCP downstream of TSS. By combining these features, we conclude that reads mapping this region either span over it or indicate low coverage here. If a read maps here, its length is relatively longer than in other regions (Fig. 3.3C).

3.2.1.4 Sequencing Features Related to Gene Expression Level in Healthy Blood Cells

We stratified genes into four categories based on their mean expression in five different healthy blood cells. On average, 32.0% (±1.1% mean SD) of cfDNA originated from granulocytes, 29.7% (±0.8%) from erythrocyte progenitors, 10.5% (±1.1%) from monocytes, and 12.1% (±0.7%) from lymphocytes. The main solid

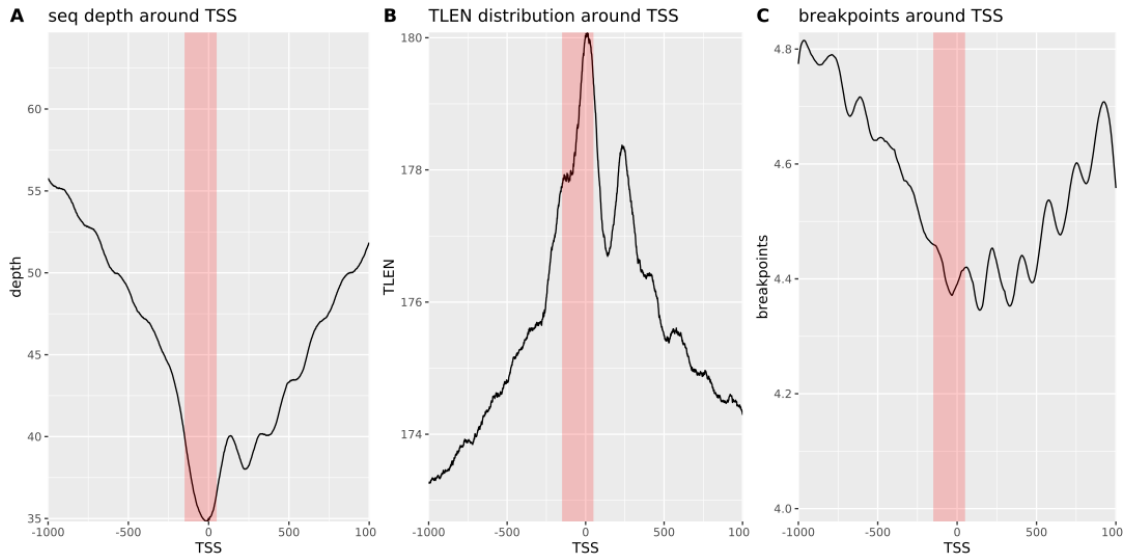


Figure 3.3: Different Features of ±1 kb around First TSS. A) Sequencing depth indicates a lower coverage region around NDR (red rectangle). **B)** TLEN distribution shows that in NDR regions, DNA fragments are longer, with an average length significantly longer than 167 bp, suggesting that only longer fragments can traverse this region. **C)** Breakpoints distribution indicates fewer breakpoints around NDR. Clear periodicity with the size of nucleosome downstream of NDR is observed in all three features.

tissue sources of cfDNA were vascular endothelial cells (8.6±0.9%) and hepatocytes (1.2±0.4%) (Moss et al. 2018). These five blood cells contribute the most frequent contents in cfDNA. We retrieved RNA sequencing data from Blueprint and stratified genes by their mean TPM into quartiles. Each group contained an equal number of genes. Lower expressed genes exhibited higher coverage, indicating well-positioned nucleosomes downstream. Higher expressed genes showed clearer breakpoints amplitude (Fig. 3.4).

In our model, most fragments are the same size as one or two nucleosomes, suggesting that cfDNA circulates wrapped around the nucleosome core particle. Thus, we extracted reads with specific lengths and counted their contributions to the coverage distribution. While reads of single nucleosome size show a distribution similar to all reads, those of double nucleosome size provide a clearer signal at the NDR region. Lower coverage and relatively higher breakpoints counts may be due to the weak linkage between the -1 and +1 nucleosome, where nucleotides are exposed to transcription machinery (Fig. 3.5).



Figure 3.4: Different Features around First TSS. Genes are stratified by their TPM. The top expressed group includes genes with TPM above 11.5 (inclusive); the high expressed group, between 1.98 (inclusive) to 11.5; the median expressed group, between 0.058 (inclusive) to 1.98; and the low expressed group, below 0.058. All data are normalised to the median TPM of each category. **A)** Mean read depth of the four categories of expression levels. Lower expressed genes show higher coverage, indicating a well-protected region and well-positioned nucleosomes downstream. Higher expressed genes exhibit clearer amplitude, suggesting active cleavage at the NDR rather than upstream and downstream. **B)** Mean insert size of the four categories of expression levels. Higher expressed genes have longer read lengths around NDR. Lower expressed genes display a more uniform distribution across ±1kb around TSS. **C)** Mean breakpoints counts of the four categories of expression levels. Lower expressed genes show flatter patterns, while higher expressed genes exhibit clearer amplitude. A clear peak at the edge of the NDR indicates the preferred cleavage site.

3.2.1.5 Reads Spanning Over Single or Double Nucleosomes

Combining gene expression data with read length, we compared the distribution of reads of varying lengths under different gene expression levels. We stratified genes into four categories, as previously mentioned, and mapped the selected reads. Reads with lengths corresponding to single nucleosomes exhibit a pattern similar to that of all reads. Those with lengths corresponding to double nucleosomes display a more intricate pattern, potentially attributable to a small sample size. Nonetheless, it is evident that genes with higher expression levels exhibit reduced coverage in the Nucleosome Depleted Region (NDR) (Fig. 3.6).

We also examined the distribution of breakpoints for reads of different lengths.

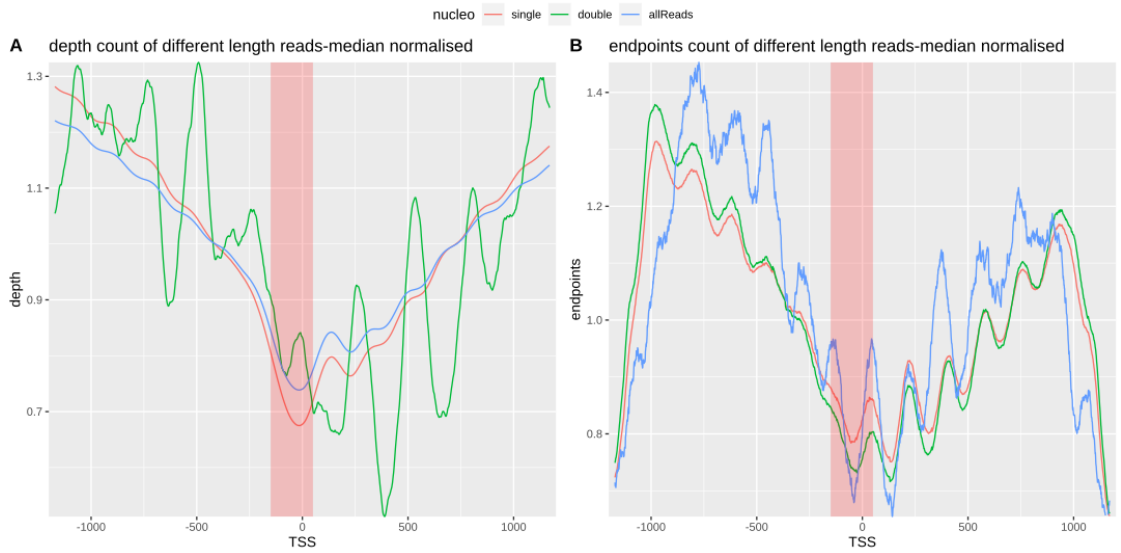


Figure 3.5: The Depth and Breakpoints Distribution of Different Fragment Lengths. Reads of 150-170bp in length are considered the size of a single nucleosome. Reads of 300-400bp are considered the size of dinucleosomes. Both datasets are normalised to the median. **A)** Read coverage around TSS stratified by their length. **B)** Fragments breakpoints distribution around TSS.

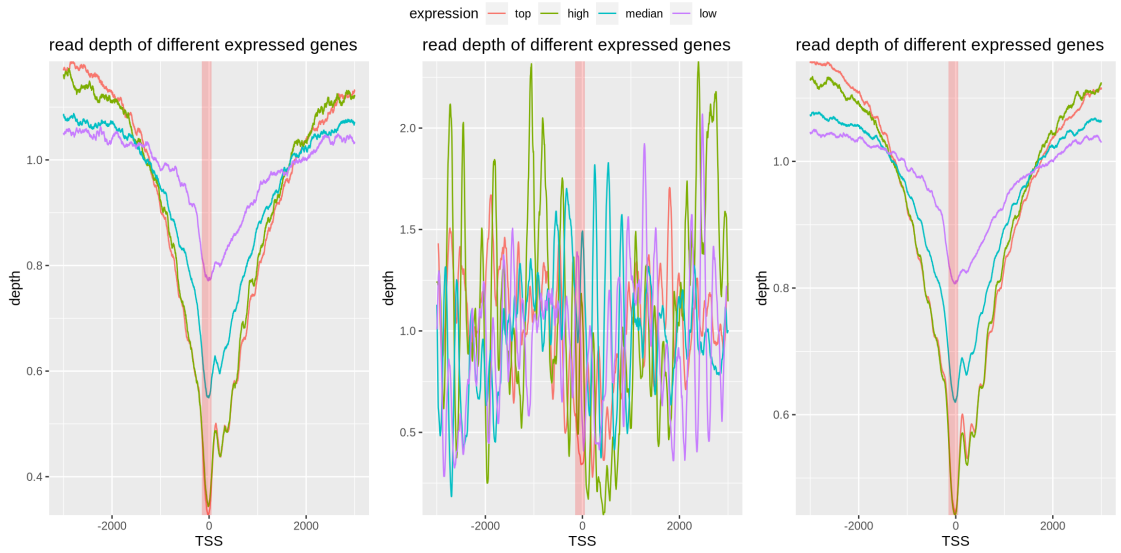


Figure 3.6: Depth Distribution of Different Read Lengths. A) Reads with the length of single nucleosome size, stratified by gene expression levels. **B)** Reads with the length of double nucleosome size, stratified by gene expression levels. **C)** Distribution of all reads. Reads are mapped and normalised to their respective expression category within a ±1.2kb region, with only ±1kb displayed.

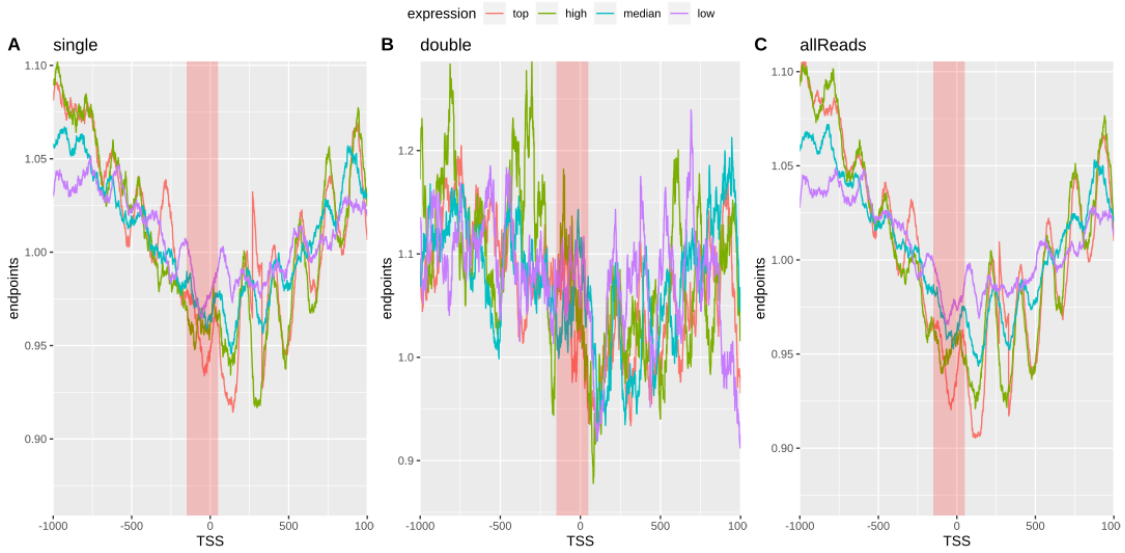


Figure 3.7: Breakpoints Distribution of Different Read Lengths. A) Reads with the length of single nucleosome size, stratified by gene expression levels. **B)** Reads with the length of double nucleosome size, stratified by gene expression levels. **C)** Distribution of all reads. Reads are mapped and normalised to their respective expression category within a ±1.2kb region, with only ±1kb displayed.

Although the sample size for breakpoints is smaller than for depth, and thus provides less extensive information, a distinct decrease at the NDR boundary is observable, indicative of a preferred cleavage site, particularly within the double nucleosome dimer (Fig. 3.7).

3.2.2 TAPS-trial Data

We conducted analogous analyses on the existing TAPS-trial data. This tumour sample was derived from the plasma of a patient with late-stage oesophageal adenocarcinoma (OAC) and was sequenced at a 40-fold coverage. Generally, the distribution is akin to that observed in pooled healthy samples. In the vicinity of the transcription start site (TSS) of all protein-coding genes, an NDR and periodicity downstream of the TSS are noted (Fig. 3.8).

Subsequently, we quantified reads of specific lengths and normalised them to the median. The reads exhibit a distribution analogous to that of healthy controls. As anticipated, reads with lengths corresponding to dinucleosomes are more likely

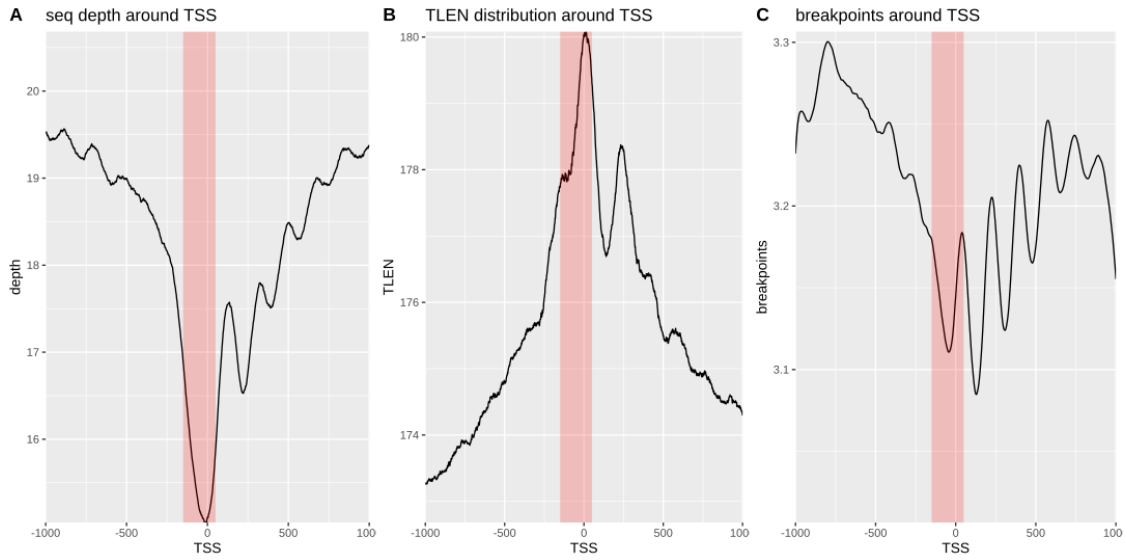


Figure 3.8: Characteristics of ±1 kb around the First TSS in Tumour Patient Samples. A) Sequence depth indicates a lower coverage region around the NDR (red rectangle). **B)** The distribution of fragment lengths (TLEN) shows that within NDR regions, DNA fragments are longer, with an average length significantly exceeding 167 bp, suggesting that only longer fragments can traverse this area. **C)** The distribution of breakpoints indicates a scarcity of breakpoints around the NDR. Pronounced periodicity with the nucleosome size is observed downstream of the NDR in all three characteristics.

to bypass the NDR, with a distinct cleavage pattern evident at the downstream edge of the NDR region (Fig. 3.9).

Consequently, we pooled 22 tumour patient cfDNA samples and 24 non-tumour patient cfDNA samples from the TAPS-trial dataset. We compared the read distribution between tumour and healthy samples. Both tumour and healthy samples in our TAPS-trial exhibit a pattern similar to that of the published deep-sequenced dataset. Among reads with lengths corresponding to double nucleosomes, the differences between lowly expressed and highly expressed genes are even more pronounced (Fig. 3.10).

Tumour and non-tumour samples in our TAPS-trial also display a pattern in fragment endpoints distribution akin to the deep-sequenced dataset. A significant decrease at the downstream of the TSS and the highest amplitude in the dinucleosome group are observable (Fig. 3.11).

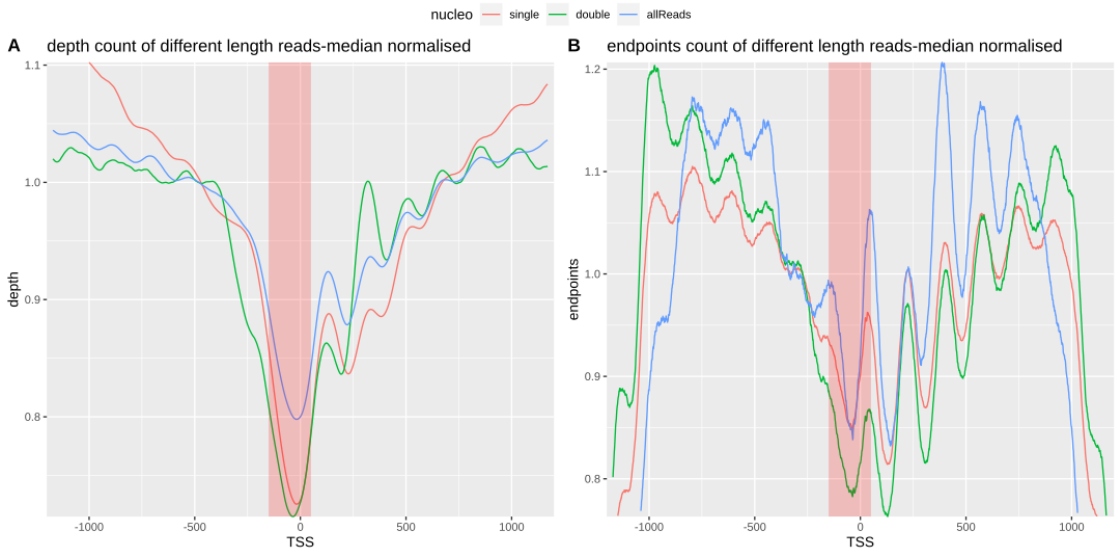


Figure 3.9: Tumour Sample, Characteristics of ±1 kb around the First TSS. **A)** Read coverage around the TSS, stratified by their length. **B)** Distribution of fragment breakpoints around the TSS.

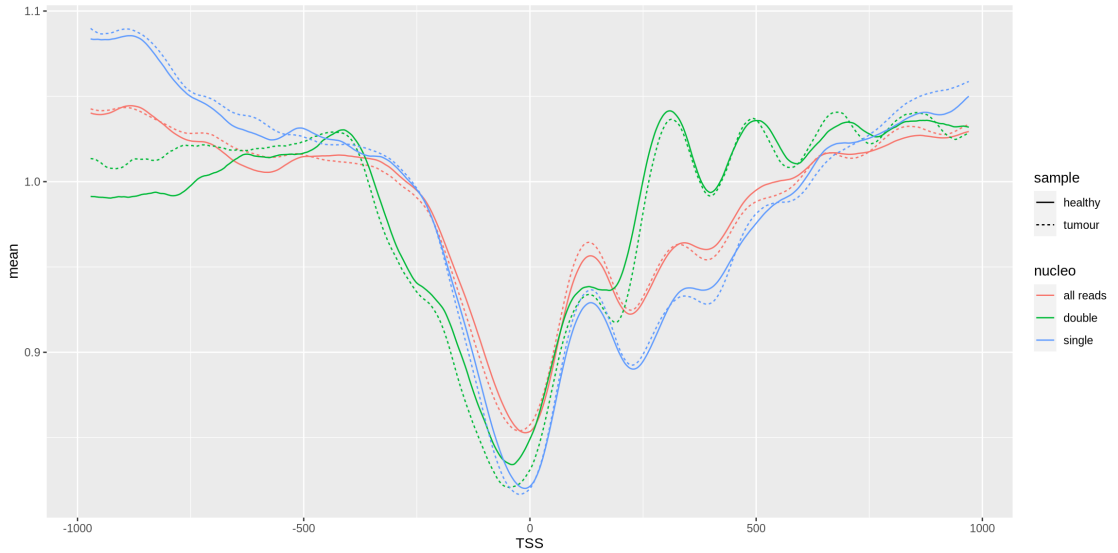


Figure 3.10: Depth Pattern of All Healthy Samples versus Tumour Samples. 22 tumour samples and 24 non-tumour samples are pooled separately and normalised to their respective group median.

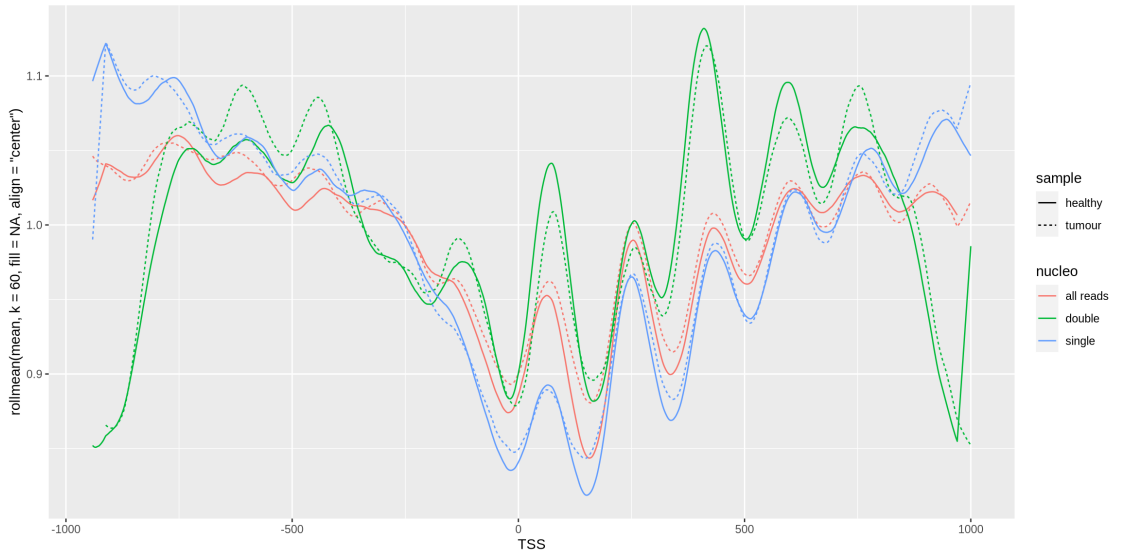


Figure 3.11: Fragmentation Pattern of All Healthy Samples versus Tumour Samples

3.3 Conclusion and Discussion

Our findings from the deep-sequenced cfDNA dataset substantiate the model wherein nucleic acid molecules encircling nucleosomes are safeguarded, and cleavage predominantly occurs between nucleosomes.

Concurrently, there exists DNA helix-related cleavage associated with nucleosome binding.

The distribution of cfDNA fragments across the genome exhibits discernible patterns.

In proximity to the TSS, there is a paucity of reads enriched in the NDR, whereas downstream of the NDR, a favoured cleavage site is frequently observed, highlighting the nucleosome's position.

Dinucleosomes are more inclined to bypass the NDR than single nucleosomes, suggesting that the NDR is a region where nucleosomes are less densely packed.

The coverage and cleavage patterns of cfDNA are correlated with gene expression levels.

This robustly supports the notion that cfDNA can serve as an indicator for tissue-specific gene expression.

In individual sequencing data with lower coverage from various library preparations, we continue to observe the aforementioned phenomenon, thereby confirming its reproducibility.

Additionally, we have discerned that the cfDNA signature of tumour samples aligns with solid tumour gene expression.

4

Tumour Patients' Cell-free DNA Sequencing Reveals Tumour Frequency in Plasma

4.1 Introduction

Circulating tumour DNA (ctDNA), which contains information about tumour-specific expressed genes, is a pivotal biomarker in the clinical context of non-invasive testing. Our study aimed to investigate whether the plasma of tumour patients exhibits abnormally elevated levels of tumour-specific expression genes, potentially allowing for tumour localization through their tissue specificity.

We hypothesized that the cfDNA of tumour patients would differ significantly from that of non-tumour patients in terms of the tumour-specific expressed gene set. To test this, we correlated the cfDNA sequencing data of tumour patients with the RNA sequencing data of solid tumour tissues. Through statistical testing of coverage and range, we observed that the cfDNA of tumour patients indeed showed differences compared to non-tumour patients, although the distinction was not pronounced, possibly due to inter-individual variations obscuring the tumour-specific signature.

This chapter presents a comprehensive exploration of our research approach and the insights we have gleaned, starting with the stratification of genes based on their

tumour-specific expression and advancing to an analysis of sequencing features that are correlated with these genes. We have correlated the cfDNA sequencing data from tumour patients with RNA sequencing data from solid tumour tissues, shedding light on how ctDNA content influences cfDNA sequencing patterns. Through a statistical analysis of gene expression, we have tested the coverage and range of gene expression in cfDNA, highlighting tumour-specific expressed genes that may indicate the presence of tumour-specific DNA.

Our stratification of genes by their tumour-specific expression has enabled us to identify genes that are uniquely expressed in tumour cells. This categorization has been instrumental in our analysis of sequencing features, where we have compared the sequencing depth and fragment length between tumour patients and non-tumour patients, using as a reference genes that are not expressed in either group.

Furthermore, we have delved into the correlation between the coverage range and tissue RNA sequencing data, particularly focusing on the sequencing depth range around the transcription start site (TSS) and its relation to gene expression levels. This analysis has been supplemented by our prediction of ctDNA content in the cfDNA of tumour patient samples through Copy Number Variation (CNV) analysis, allowing us to sort these samples according to ctDNA frequency.

Lastly, we have combined coverage data with breakpoints to calculate the Windowed Protection Score (WPS), a parameter that provides a quantitative description of the positioning of nucleosomes and their correlation with gene expression. This innovative approach has offered us additional insights into the complex interplay between nucleosome positioning and gene expression within the context of tumour-specific DNA.

4.2 Results

4.2.1 Stratifying the Genes by Their Tumour-Specific Expression

The cfDNA sequencing pattern is related to gene expression. To verify whether cfDNA can be used as a tumour indicator, we divided the genes in the human

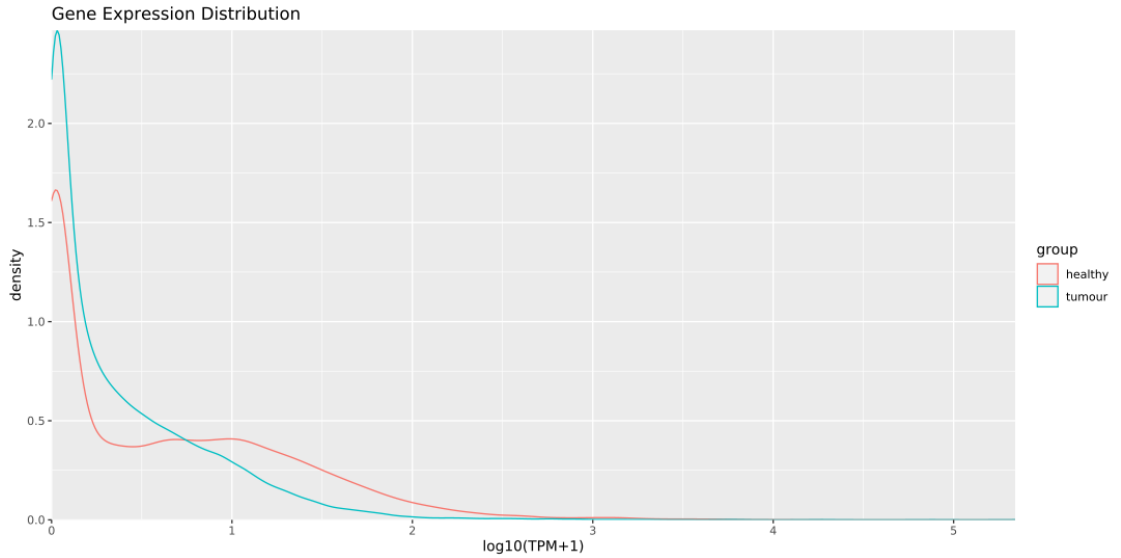


Figure 4.1: The Gene Expression Distribution in Both Tumour and Healthy Blood Cells. The TPM of both groups of genes are dense below 10. Very few genes have high expression levels with TPM over 1000.

genome into four groups according to their expression differences in healthy cells and tumour cells. We used five types of healthy blood cells gene expression data and calculated their gene mean TPM. Concurrently, we used 22 solid tumour tissue sample RNA sequencing data from the TAPS-trial and calculated the mean TPM as tumour TPM. Initially, we calculated the gene expression distribution in both tumour and healthy blood cells. Most genes have TPM lower than 10, while few genes have TPM higher than 1000 (Fig. 4.1).

Some genes exhibit different expression patterns in tumour and healthy cells. To identify genes that are specifically expressed in tumour cells, we compared their expression in tumour samples and healthy blood cells. Thus, we define genes with TPM equal to 0 across all samples as unexpressed in the group, and genes with mean TPM larger than 10 as expressed genes in the group. After removing genes that were abnormally expressed within their group, we classified the remaining genes into three expression patterns: all expressed, all unexpressed, and tumour-specific expressed (Fig. 4.2).

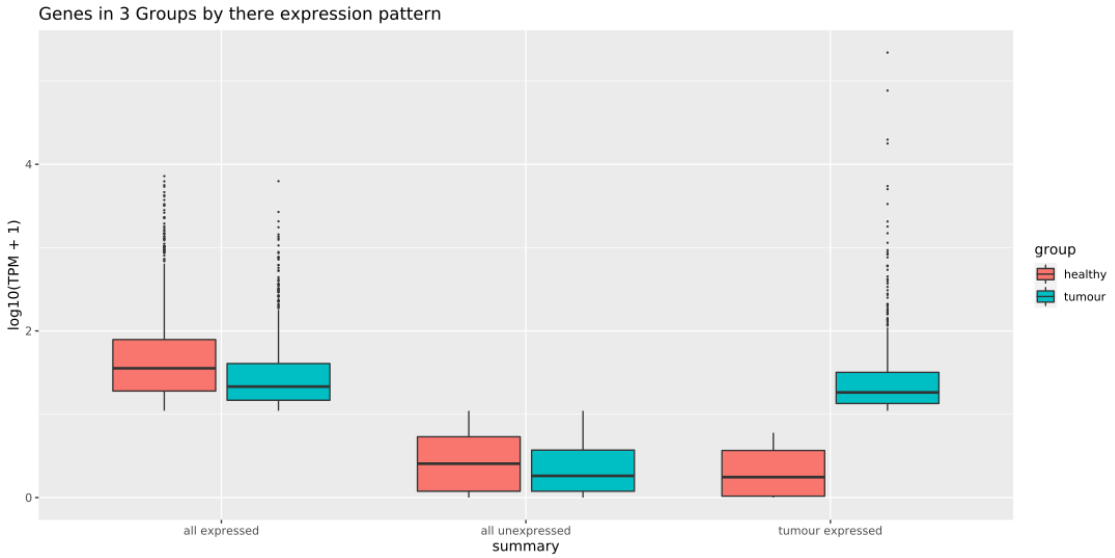


Figure 4.2: Genes into 3 Different Expression Patterns. Only genes expressed stably within the same sample group have been selected. For tumour-specific genes, we define the tumour cell gene TPM larger than 10 and in healthy blood cells TPM lower than 5.

4.2.1.1 Sequencing Features Correlating with Tumour-Specific Expressed Genes

In tumour patients, the expression of tumour-specific genes increases, and nucleosomes at the transcription start site (TSS) become more active. Consequently, cfDNA from tumour patients will contain evidence of increased expression of these genes. We compared the sequencing depth and fragment length between tumour patients and non-tumour patients (Fig. 4.3). We used genes unexpressed in both groups as a reference. In tumour patients, the tumour-specific genes are more likely to show a difference compared to the unexpressed gene group. However, we cannot ignore that the difference in the overall cfDNA sequencing pattern between tumour patients and non-tumour patients is not significant. This may be due to the fact that DNA from tumour cells only constitutes a small fraction of the cfDNA of tumour patients.

4.2.1.2 Coverage Range Correlates with Tissue RNA Sequencing Data

We observed that around the TSS, the range in sequencing depth is related to the expression level of the gene. To find quantitative evidence, we tested the

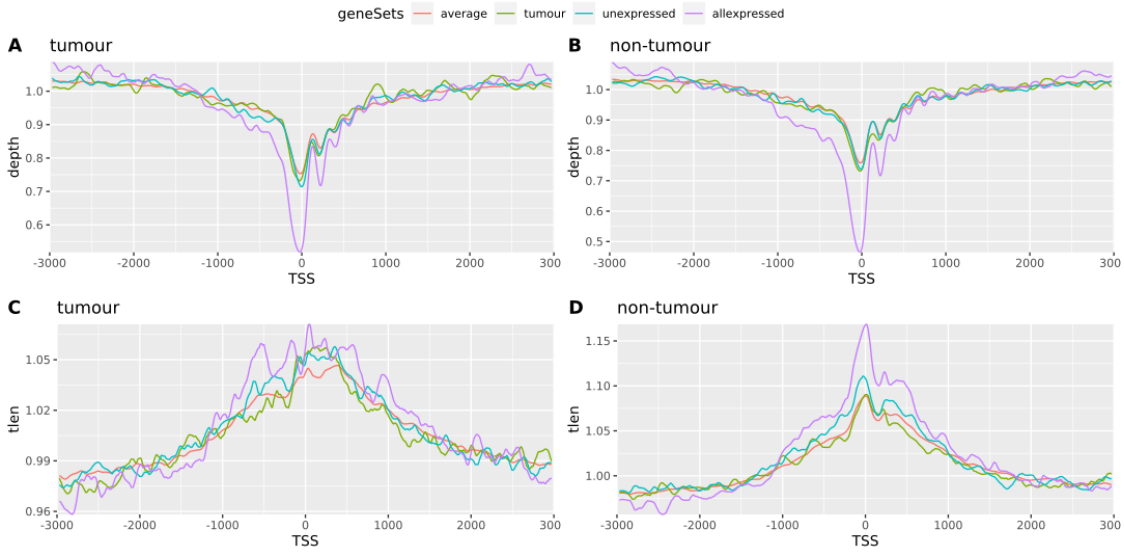


Figure 4.3: Sequencing Features Correlating with Tumour-Specific Expressed Genes. A) Tumour patient cfDNA sequencing depth grouped by gene expression pattern. **B)** Non-tumour patient cfDNA sequencing depth grouped by gene expression pattern. **C)** Tumour patient cfDNA fragments distribution grouped by gene expression pattern. **D)** Non-tumour patient cfDNA fragments distribution grouped by gene expression pattern. We use read insert length (TLEN) as read fragment length. Enzymatic end repair during library preparation removes 5ánd 3óverhangs and may obscure true cleavage sites.

correlation between the range and gene expression. We calculated the range of per-base sequencing depth within ±1kb around the TSS and correlated it to gene expression. In non-tumour samples, there is a significant difference between the all-expressed gene set and the all-unexpressed gene set. Meanwhile, we found no statistical evidence of a difference between globally unexpressed genes and tumour-expressed genes. We also observed a quantitative correlation between the depth range and gene TPM.

Then we compared tumour samples with non-tumour samples. In both tumour samples and non-tumour samples, the depth range in the all-expressed gene set correlates with TPM. However, in the tumour-specific expressed gene set, the correlation between TPM and tumour samples (Fig. 4.4)is more significant than the correlation with non-tumour samples (Fig. 4.5).

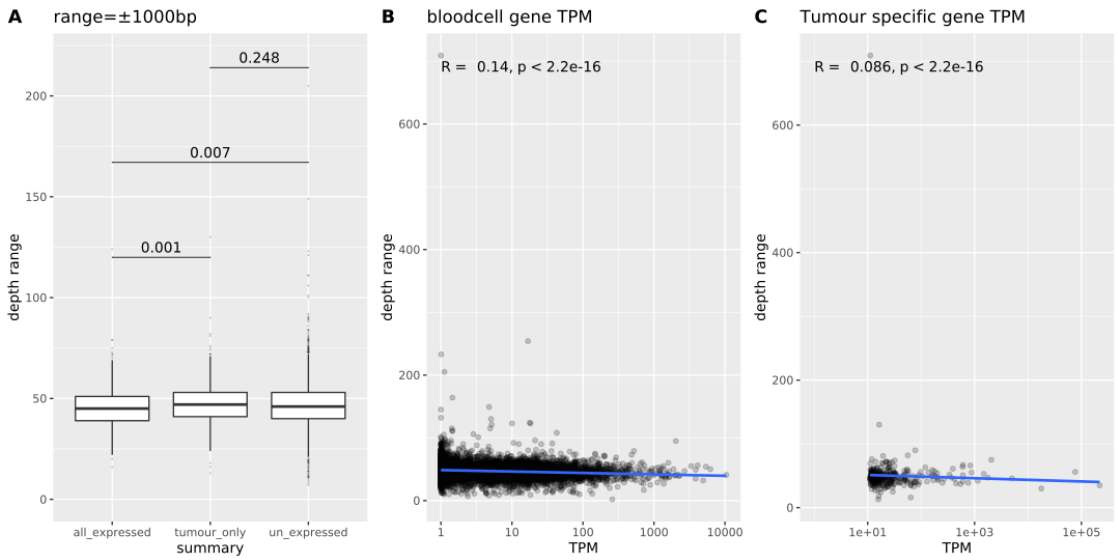


Figure 4.4: Gene Expression Pattern in Healthy Blood Cells Correlates with Depth Range in Tumour Samples. A) Depth range distribution in three gene sets. p-values of Wilcoxon pairwise tests are shown on the horizon line. **B)** Depth range distribution correlates with healthy blood cell TPM. **C)** Depth range distribution correlates with tumour tissue TPM, counting only tumour-specific expressed genes. Spearman's rank correlation coefficient and p-value are shown on the top right.

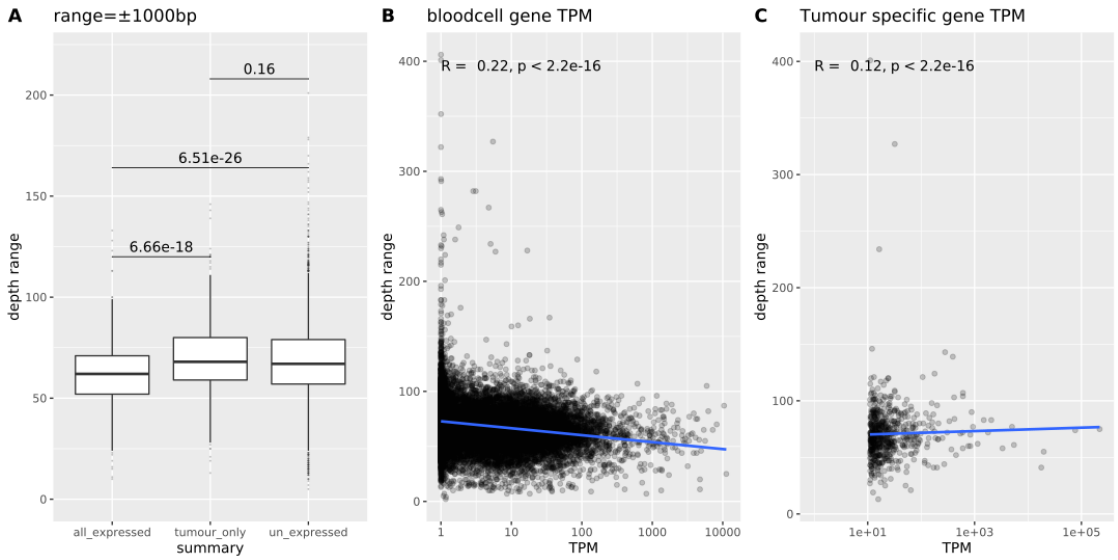


Figure 4.5: Gene Expression Pattern in Healthy Blood Cells Correlates with Depth Range in Non-Tumour Samples. A) Depth range distribution in three gene sets. p-values of Wilcoxon pairwise tests are shown on the horizon line. **B)** Depth range distribution correlates with healthy blood cell TPM. **C)** Depth range distribution correlates with tumour tissue TPM, counting only tumour-specific expressed genes. Spearman's rank correlation coefficient and p-value are shown on the top right.

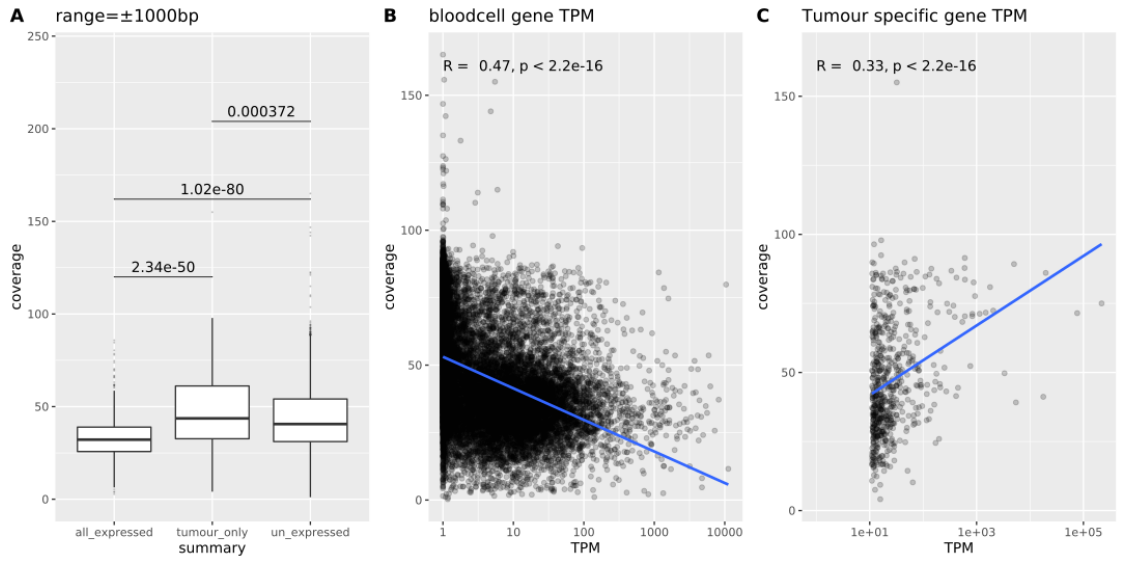


Figure 4.6: Gene Expression Pattern in Tumour Tissues Correlates with Coverage within ± 1kb Around TSS in Non-Tumour Samples. A) Coverage distribution in three gene sets. p-values of Wilcoxon pairwise tests are shown on the horizon line. **B)** Coverage distribution correlates with healthy blood cell TPM. **C)** Depth range distribution correlates with tumour tissue TPM, counting only tumour-specific expressed genes. Spearman's rank correlation coefficient and p-value are shown on the top right.

4.2.1.3 Coverage Around Transcription Start Sites Correlates with Gene Expression Level

We also calculated the correlation between coverage and TPM in specific regions. In non-tumour samples, there was a correlation between coverage in nucleosome-depleted regions (NDRs) and healthy blood cell TPM (Fig. 4.6). This correlation appears to be weakened in the tumour sample (Fig. 4.7).

4.2.1.4 Correlation of the Tumour Frequency and the Sequencing Data

Although there are certain statistical differences in sequencing features between different gene sets, it remains challenging to quantitatively describe the differences in gene expression in different cell samples through cfDNA sequencing data.

The primary reason is that in the serum of tumour patients, ctDNA from tumour cells only constitutes a small part of the total cfDNA. Therefore, in all cfDNA sequencing, cfDNA from healthy blood cells will mask the ctDNA signal.

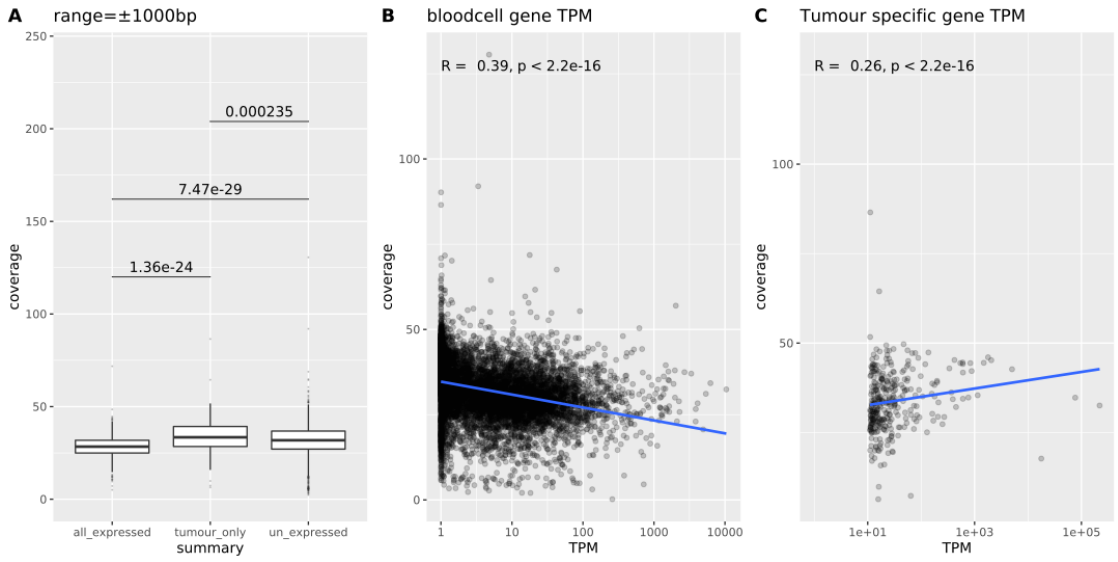


Figure 4.7: Gene Expression Pattern in Tumour Tissues Correlates with Coverage within ± 1kb Around TSS in Tumour Samples. A) Coverage distribution in three gene sets. p-values of Wilcoxon pairwise tests are shown on the horizon line. **B)** Coverage distribution correlates with healthy blood cell TPM. **C)** Depth range distribution correlates with tumour tissue TPM, counting only tumour-specific expressed genes. Spearman's rank correlation coefficient and p-value are shown on the top right.

We predicted the ctDNA content in the cfDNA of our tumour patient samples through Copy Number Variation (CNV) and sorted these samples according to ctDNA frequency. Among them, we can observe certain sample specificity (Fig. 4.8). Lower expressed genes have higher amplitude between 1kb around TSS, which indicates well-positioned genes, and well-positioned nucleosomes tend to give a clear pattern. In general, samples with higher tumour content frequency have a higher amplitude in the NDR region.

We extracted reads of specific lengths to determine the distribution of nucleosomes in this region. As can be seen, single nucleosomes(Fig. 4.9) usually appear downstream of the NDR, while dinucleosomes(Fig. 4.10) usually skip the NDR. In samples with high tumour frequency, this phenomenon is more significant on tumour-specific expressed genes.

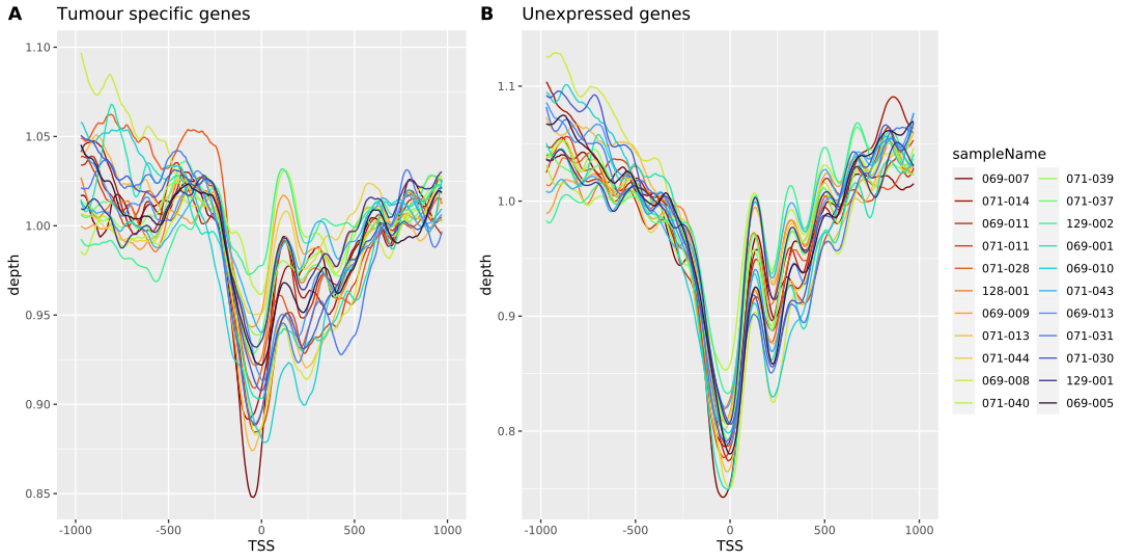


Figure 4.8: All Tumour Samples Read Depth in Different Gene Sets. A) Reads aligned with tumour-specific genes. **B)** Reads aligned with globally unexpressed genes. Tumour content frequency decreases from red to purple. Depth is normalized to the median of their own group.

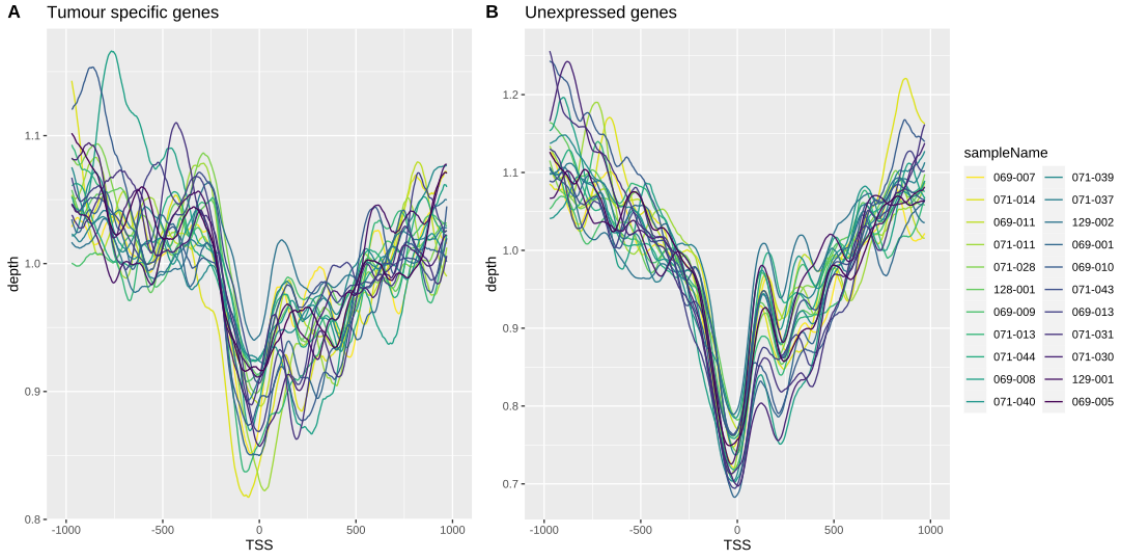


Figure 4.9: Depth of Tumour Samples Reads with Length of Single Nucleosome. A) Reads aligned with tumour-specific genes. **B)** Reads aligned with globally unexpressed genes. Tumour content frequency decreases from yellow to purple. Depth is normalized to the median of their own group.

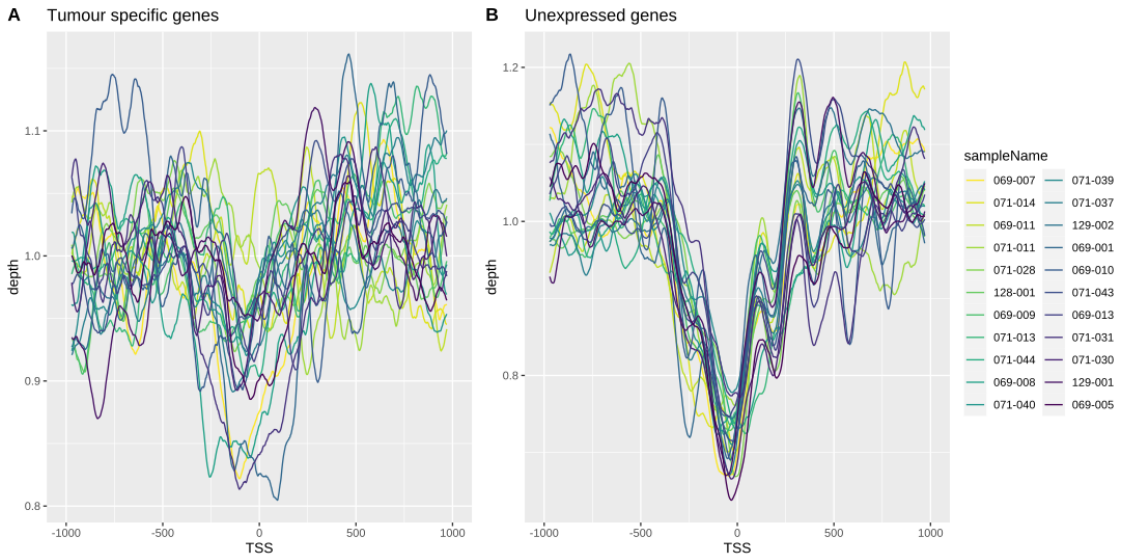


Figure 4.10: Depth of Tumour Samples Reads with Length of Double Nucleosome. A) Reads aligned with tumour-specific genes. **B)** Reads aligned with globally unexpressed genes. Tumour content frequency decreases from yellow to purple. Depth is normalized to the median of their own group.

4.2.1.5 Different Expression Level in EAC for Different Read Length for All Tumour Samples

Lower expressed genes have higher amplitude between 1kb around TSS, which indicates well-positioned genes, while well-positioned nucleosomes tend to give a clear pattern. As tumour frequency increases, the correlation with tumour-specific expressed genes becomes significant (Fig. 4.11).

Similar to the previous conclusion, the depletion of single nucleosomes in the NDR region is more significant in samples with higher tumour frequency. A more obvious difference can be observed when comparing within the tumour-expressed gene set between all reads and single nucleosome reads (Fig. 4.12).

Due to the small sample size, the difference between the double nucleosomes group is not obvious, but we can still observe that double nucleosomes tend to skip NDR and enrich at downstream with periodicity(Fig. 4.13).

4.2.2 Signal Strength Correlates with Tumour Frequency

We related the coverage range in each sample to the tumour frequency of the predicted sample and obtained the correlation between depth range or coverage

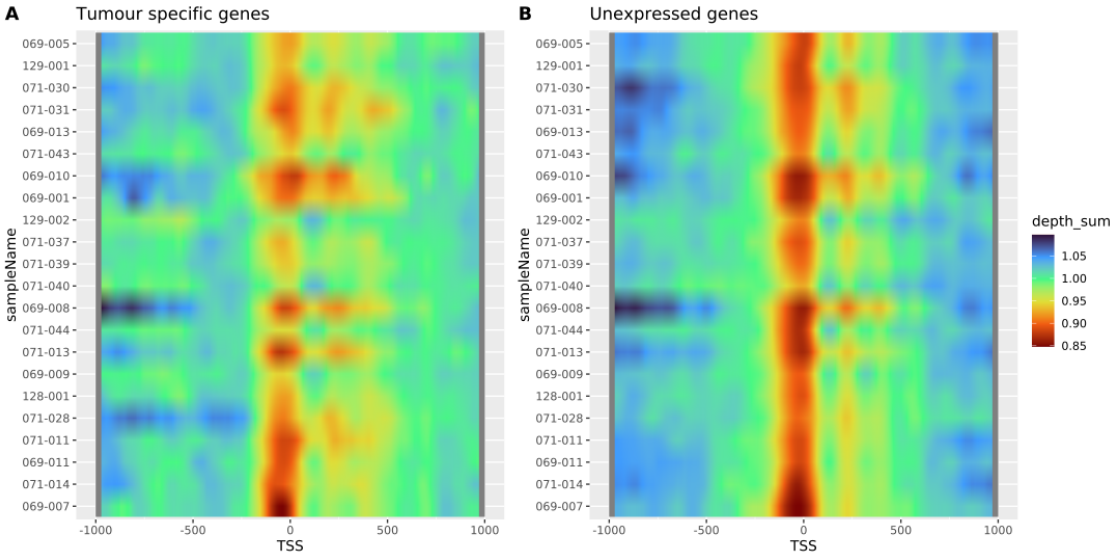


Figure 4.11: Coverage of Tumour Samples Reads. A) Reads aligned with tumour-specific genes. **B)** Reads aligned with globally unexpressed genes. Tumour content frequency increases from top to bottom. Depth is normalized to the median of their own group.

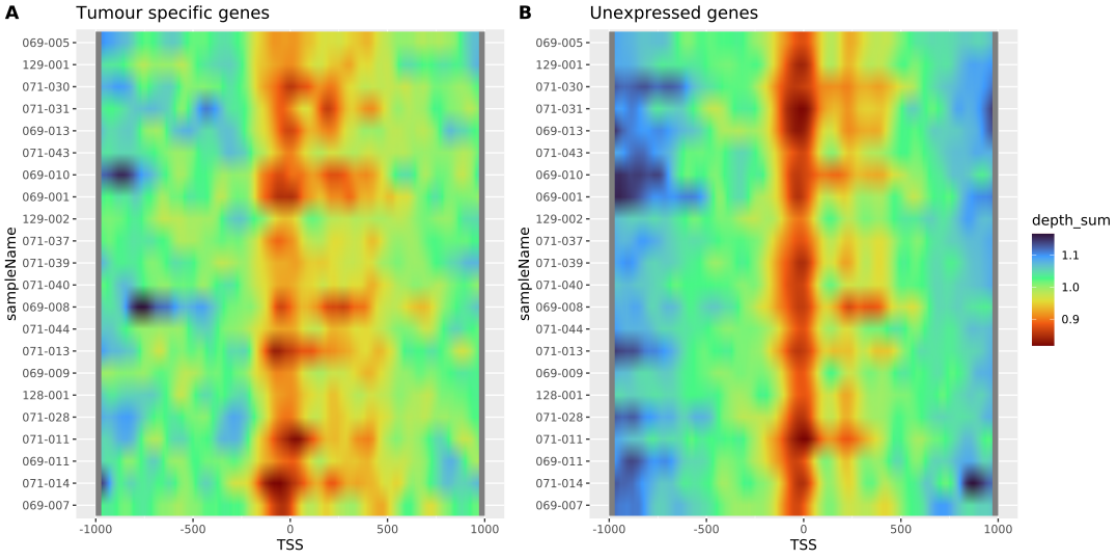


Figure 4.12: Coverage of Tumour Samples Reads with Single Nucleosome Length. A) Reads aligned with tumour-specific genes. **B)** Reads aligned with globally unexpressed genes. Tumour content frequency increases from top to bottom. Depth is normalized to the median of their own group.

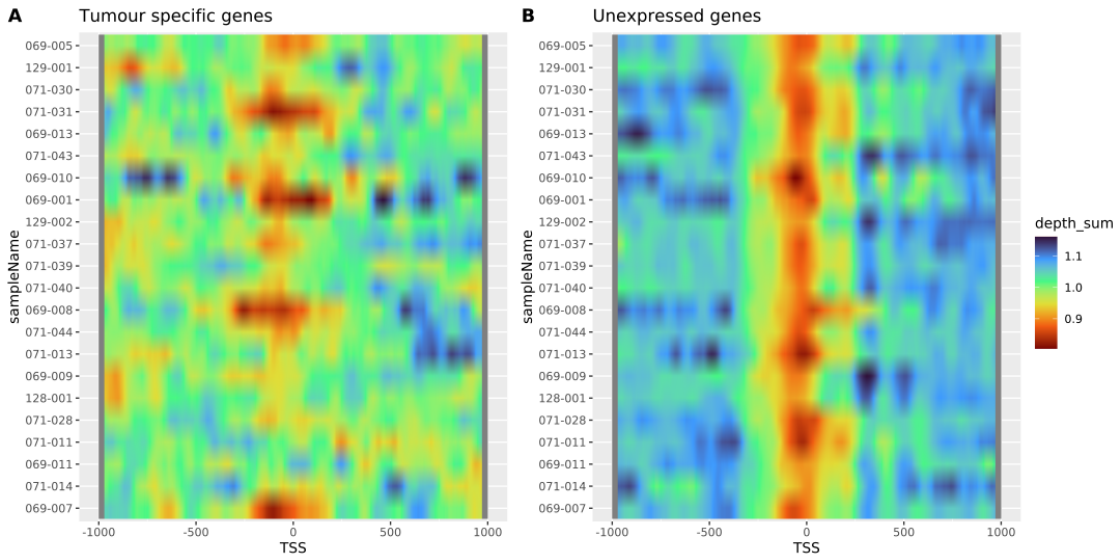


Figure 4.13: Coverage of Tumour Samples Reads with Double Nucleosome Length. A) Reads aligned with tumour-specific genes. **b)** Reads aligned with globally unexpressed genes. Tumour content frequency increases from top to bottom. Depth is normalized to the median of their own group.

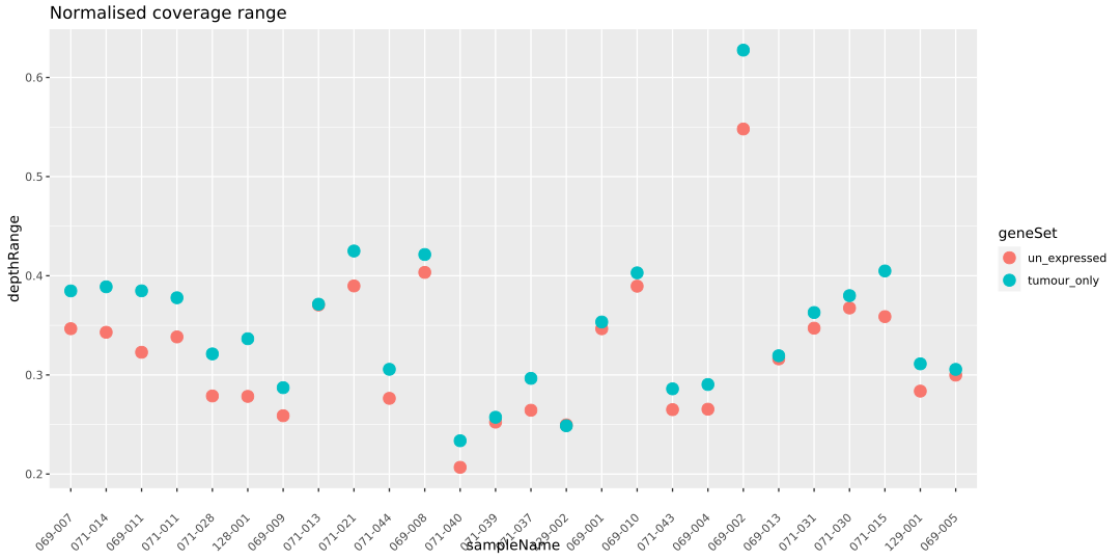


Figure 4.14: Coverage and Range within ±1kb Around TSS. Values are normalized to the median of their own group. Tumour content frequency decreases from left to right.

and tumour frequency. We sorted the tumour samples according to their tumour content frequency and tested the coverage and range within ±1kb around the TSS. Both values were normalized to the median of their own group(Fig. 4.14).

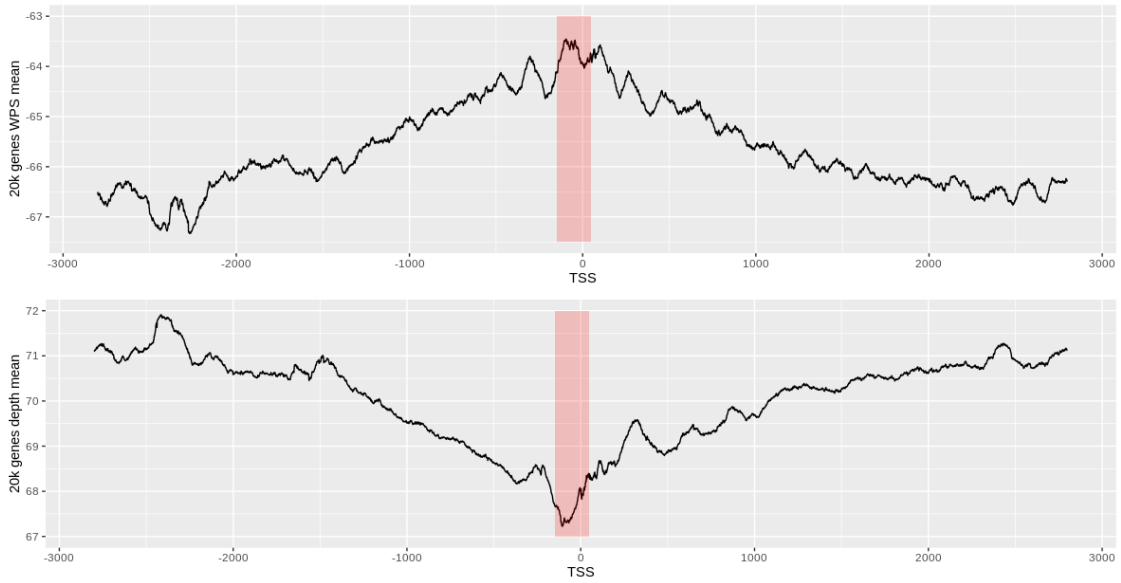


Figure 4.15: Read WPS and Depth. WPS has similar characteristics to depth and can reflect the position of the nucleosome to a certain extent.

4.2.3 Coverage and Breakpoints Combination

To combine coverage and breakpoints, we adopted a parameter quantitatively describing this as the Windowed Protection Score (WPS), which is the number of DNA fragments completely spanning a 120 bp window centered at a given genomic coordinate minus the number of fragments with an endpoint within that same window (Snyder et al. 2016). The WPS exhibits a strong signal around the TSS correlating with the endpoint counts and the depth (Fig. 4.15).

Both features exhibit a significant signal at the NDR as well as periodicity upstream and downstream of the TSS.

We examined highly expressed genes and unexpressed genes and compared their depth and WPS. In contrast, highly expressed genes have lower coverage, which may be because open chromatin generates shorter fragments that are unlikely to be recovered during library preparation. Conversely, unexpressed genes have lower WPS with more pronounced periodicity and larger amplitude, inferring strongly positioned nucleosomes (Fig. 4.16).



Figure 4.16: Read WPS and Depth Stratified by Gene Expression Levels. We sorted genes by their expression TPM order, stratified the genes into four categories by their TPM.

4.3 Conclusion and Discussion

In conclusion, our study underscores the potential of cfDNA sequencing as a non-invasive diagnostic tool for cancer detection. By stratifying genes based on their differential expression in tumour cells versus healthy haematopoietic cells, we identified tumour-specific expression genes through cfDNA sequencing. Although the sequencing patterns of these genes showed differences between tumour and non-tumour patients, the distinction was less pronounced than anticipated, potentially due to inter-individual variations in cfDNA composition.

Our analysis indicated that the sequencing depth and fragment length around the TSS of tumour-specific genes are influenced by gene expression levels. This correlation was more evident in non-tumour samples, suggesting that ctDNA from tumour cells in the plasma of tumour patients may be obscured by cfDNA from healthy cells.

CNV analysis allowed us to predict ctDNA content and sort samples by ctDNA frequency, revealing a correlation between tumour frequency and sequencing patterns. Samples with higher tumour content frequency showed greater amplitude in NDRs,

indicating a link between nucleosome positioning and gene expression levels.

The WPS, which combines coverage data with breakpoints, provides a quantitative measure of nucleosome positioning and its correlation with gene expression. WPS exhibited characteristics akin to sequencing depth and reflected nucleosome positioning to a certain extent.

Overall, our research highlights the complexity of utilising cfDNA sequencing for cancer detection and the need for further investigation with larger cohorts and more advanced analytical techniques. Integrating cfDNA sequencing with other diagnostic methods may enhance sensitivity and specificity in cancer detection, potentially leading to improved patient outcomes.

5 Conclusion

Nucleosomes, as the fundamental units of chromatin, play a pivotal role in the regulation of gene expression through steric hindrance mechanisms. By mapping the dynamic positioning of nucleosomes, we gain insights into the gene expression profiles within cells. Circulating free DNA (cfDNA) in plasma, often encapsulated within nucleosomes, provides a non-invasive window into the genomic locations of nucleosomes at the time of their release from cells, thereby offering clues to the expression levels of specific genes.

In this thesis, we have explored the correlation between cfDNA sequencing patterns and gene expression. We observed that in highly expressed genes, Nucleosome Depleted Regions (NDRs) exhibit reduced coverage and broader ranges. Notably, there are several distinct periodic peaks downstream of the transcription start site (TSS), indicating the presence of multiple, well-aligned nucleosomes. In non-cancer patients' plasma, the sequencing patterns of tumour-specific genes mirror those of non-expressed genes, and they markedly differ from the patterns observed in globally expressed genes. Conversely, in cancer patients' plasma, a discernible gap emerges between the patterns of tumour-specific expressed and non-expressed genes. Employing semi-quantitative analysis, we have been able to differentiate between cancer and non-cancer patients to a certain degree.

The non-invasive nature of cfDNA analysis is particularly crucial for cancer detection, as it offers a less intrusive approach compared to traditional biopsy methods. This is especially important for patients who may be unable to undergo invasive procedures or for whom repeated biopsies are not feasible. The potential clinical applications of our findings are significant, as they could contribute to the development of early cancer detection methods that are more accessible and less risky for patients.

However, the variability in cfDNA among individuals poses challenges in distinguishing those with varying tumour frequencies. This variability also complicates the use of cfDNA coverage data for the prediction of tumour content within samples. One of the limitations of this study is the heterogeneity in the amount of circulating tumour DNA (ctDNA) within cfDNA, which can mask signals and make it difficult to accurately predict tumour presence and progression. Despite these challenges, our findings shed light on the potential of cfDNA sequencing as a tool for gene expression analysis and cancer diagnostics.

Looking forward, the methods developed in this study could be further improved by incorporating advanced bioinformatics tools to better account for individual variability and to enhance the sensitivity and specificity of cancer detection. Additionally, the integration of our findings with other molecular markers and imaging techniques could provide a more comprehensive approach to cancer diagnosis and monitoring.

In conclusion, while our study has its limitations, particularly in the area of inter-individual variability in ctDNA content within cfDNA, the potential for cfDNA sequencing to serve as a non-invasive diagnostic tool for gene expression and cancer detection is promising. Our methods could complement existing ctDNA-based early cancer detection approaches, offering a valuable addition to the arsenal of cancer diagnostic tools. Further research and development are necessary to refine these techniques and to fully realize their clinical potential.

References

[1] Stanislav Volik et al. “Cell-free DNA (cfDNA): clinical significance and utility in cancer shaped by emerging technologies”. In: *Molecular Cancer Research* 14.10 (2016), pp. 898–908.

[2] C Harker Rhodes, Charles Honsinger, and George D Sorenson. “Detection of tumor-derived DNA in cerebrospinal fluid”. In: *Journal of Neuropathology & Experimental Neurology* 53.4 (1994), pp. 364–368.

[3] Krishna B Sriram et al. “Pleural fluid cell-free DNA integrity index to identify cytologically negative malignant pleural effusions including mesotheliomas”. In: *BMC cancer* 12 (2012), pp. 1–12.

[4] Yan-yan Yan et al. “Cell-free DNA: hope and potential application in cancer”. In: *Frontiers in cell and developmental biology* 9 (2021), p. 639233.

[5] Kun Sun et al. “Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments”. In: *Proceedings of the National Academy of Sciences* 112.40 (2015), E5503–E5512.

[6] Edward John Filippone and John L Farber. “The Monitoring of Donor-derived Cell-free DNA in Kidney Transplantation”. In: *Transplantation* 105.3 (2021), pp. 509–516.

[7] Sabine Jahr et al. “DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells”. In: *Cancer research* 61.4 (2001), pp. 1659–1665.

[8] SA Leon et al. “Free DNA in the serum of cancer patients and the effect of therapy”. In: *Cancer research* 37.3 (1977), pp. 646–650.

[9] Heidi Schwarzenbach et al. “Detection and monitoring of cell-free DNA in blood of patients with colorectal cancer”. In: *Annals of the New York Academy of Sciences* 1137.1 (2008), pp. 190–196.

[10] P Mandel. “Les acides nucleiques du plasma sanguin chez 1 homme”. In: *CR Seances Soc Biol Fil* 142 (1948), pp. 241–243.

[11] EM Tan et al. “Deoxybonucleic acid (DNA) and antibodies to DNA in the serum of patients with systemic lupus erythematosus.” In: *The Journal of clinical investigation* 45.11 (1966), pp. 1732–1740.

[12] Florian J Bock and Joel S Riley. “When cell death goes wrong: inflammatory outcomes of failed apoptosis and mitotic cell death”. In: *Cell Death & Differentiation* 30.2 (2023), pp. 293–303.

[13] Peiyong Jiang and YM Dennis Lo. “The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics”. In: *Trends in Genetics* 32.6 (2016), pp. 360–371.

[14] Stephanie CY Yu et al. “High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing”. In: *Clinical chemistry* 59.8 (2013), pp. 1228–1237.

[15] Peter Ulz et al. “Inferring expressed genes by whole-genome sequencing of plasma DNA”. In: *Nature genetics* 48.10 (2016), pp. 1273–1278.

[16] Bhargavi Duvvuri and Christian Lood. “Cell-free DNA as a biomarker in autoimmune rheumatic diseases”. In: *Frontiers in immunology* 10 (2019), p. 502.

[17] Frank Diehl et al. “Circulating mutant DNA to assess tumor dynamics”. In: *Nature medicine* 14.9 (2008), pp. 985–990.

[18] Thierry Lecomte et al. “Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis”. In: *International journal of cancer* 100.5 (2002), pp. 542–548.

[19] Jonathan CM Wan et al. “Liquid biopsies come of age: towards implementation of circulating tumour DNA”. In: *Nature Reviews Cancer* 17.4 (2017), pp. 223–238.

[20] Florent Mouliere et al. “High fragmentation characterizes tumour-derived circulating DNA”. In: *PloS one* 6.9 (2011), e23418.

[21] Catherine Alix-Panabières, Heidi Schwarzenbach, and Klaus Pantel. “Circulating tumor cells and circulating tumor DNA”. In: *Annual review of medicine* 63 (2012), pp. 199–215.

[22] Hakan Akca et al. “Utility of serum DNA and pyrosequencing for the detection of EGFR mutations in non-small cell lung cancer”. In: *Cancer genetics* 206.3 (2013), pp. 73–80.

[23] Caitlin M Stewart and Dana WY Tsui. “Circulating cell-free DNA for non-invasive cancer management”. In: *Cancer Genetics* 228 (2018), pp. 169–179.

[24] Liying Jiang et al. “Current research status of tumor cell biomarker detection”. In: *Microsystems & Nanoengineering* 9.1 (2023), p. 123.

[25] Catherine Alix-Panabieres, Dario Marchetti, and Julie E Lang. “Liquid biopsy: from concept to clinical application”. In: *Scientific Reports* 13.1 (2023), p. 21685.

[26] Patrick R Jordan et al. “The relevance of liquid biopsy in surgical oncology: The application of perioperative circulating nucleic acid dynamics in improving patient outcomes”. In: *The Surgeon* 20.4 (2022), e163–e173.

[27] M. Murtaza et al. “Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA”. In: *Nature* 497.7447 (2013), pp. 108–112.

[28] A. M. Newman et al. “Tumor Mutational Burden and Circulating Tumor DNA as Predictors of Response to Pembrolizumab in Non-Small-Cell Lung Cancer”. In: *Cancer Discovery* 10.7 (2014), pp. 822–829.

[29] D. Zardavas et al. “Plasma ctDNA profiling offers insights into therapy response and resistance in breast cancer”. In: *Clinical Cancer Research* 21.20 (2015), pp. 4618–4627.

[30] K. C. A. Chan et al. “Early Detection of Cancer: Liquid Biopsy in the Clinic”. In: *Journal of the American Medical Association* 318.10 (2017), pp. 956–966.

[31] Chunyan Cai et al. “A dual biomarker detection platform for quantitating circulating tumor DNA (ctDNA)”. In: *Nanotheranostics* 2.1 (2018), p. 12.

[32] J. Lung et al. “Non-invasive early detection and monitoring of cancer using liquid biopsies”. In: *Nature Medicine* 26.4 (2020), pp. 541–553.

[33] Donald E Olins and Ada L Olins. “Nucleosomes: The Structural Quantum in Chromosomes: Virtually all the DNA of eukaryotic cells is organized into a repeating array of nucleohistone particles called nucleosomes. These chromatin subunits are close-packed into higher-order fibers and are modified during chromosome expression”. In: *American Scientist* 66.6 (1978), pp. 704–711.

[34] Roger D Kornberg. “Structure of chromatin”. In: *Annual review of biochemistry* 46.1 (1977), pp. 931–954.

[35] Bruce Alberts et al. *Molecular Biology of the Cell*. 4th. New York: Garland Science, 2002.

[36] MHF Wilkins. “Physical studies of the molecular structure of deoxyribose nucleic acid and nucleoprotein”. In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 21. Cold Spring Harbor Laboratory Press. 1956, pp. 75–90.

[37] Vittorio Luzzati and A Nicolaieff. “The structure of nucleohistones and nucleoprotamines”. In: *Journal of molecular biology* 7.2 (1963), pp. 142–163.

[38] Ada L Olins and Donald E Olins. “Spheroid chromatin units (ν bodies)”. In: *Science* 183.4122 (1974), pp. 330–332.

[39] Roger D Kornberg and Yahli Lorch. “Chromatin structure and transcription”. In: *Annual review of cell biology* 8.1 (1992), pp. 563–587.

[40] Min Han and Michael Grunstein. “Nucleosome loss activates yeast downstream promoters in vivo”. In: *Cell* 55.6 (1988), pp. 1137–1145.

[41] Chris D Clark-Adams et al. “Changes in histone gene dosage alter transcription in yeast.” In: *Genes & development* 2.2 (1988), pp. 150–159.

[42] Vladimir B. Teif and Christopher T. Clarkson. *Nucleosome Positioning*. 2019.

[43] Mariya Shtumpf et al. “NucPosDB: a database of nucleosome positioning in vivo and nucleosomics of cell-free DNA”. In: *Chromosoma* 131.1 (2022), pp. 19–28.

[44] Karolin Luger et al. “Crystal structure of the nucleosome core particle at 2.8 Å resolution”. In: *Nature* 389.6648 (1997), pp. 251–260.

[45] Sari Pennings et al. “Formation, stability and core histone positioning of nucleosomes reassembled on bent and other nucleosome-derived DNA”. In: *Journal of molecular biology* 207.1 (1989), pp. 183–192.

[46] Curt A Davey et al. “Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution”. In: *Journal of molecular biology* 319.5 (2002), pp. 1097–1113.

[47] Juan Ausio, Feng Dong, and KE Van Holde. “Use of selectively trypsinized nucleosome core particles to analyze the role of the histone “tails” in the stabilization of the nucleosome”. In: *Journal of molecular biology* 206.3 (1989), pp. 451–463.

[48] Cizhong Jiang and B Franklin Pugh. “Nucleosome positioning and gene regulation: advances through genomics”. In: *Nature Reviews Genetics* 10.3 (2009), pp. 161–172.

[49] Thomas Schalch et al. “X-ray structure of a tetranucleosome and its implications for the chromatin fibre”. In: *Nature* 436.7047 (2005), pp. 138–141.

[50] Jan Bednar et al. “Structure and dynamics of a 197 bp nucleosome in complex with linker histone H1”. In: *Molecular cell* 66.3 (2017), pp. 384–397.

[51] B. Alberts et al. *Molecular biology of the cell*. Ed. by Nigel Chaffey. New York, 2003.

[52] Kazuhiro Maeshima et al. “Nucleosomal arrays self-assemble into supramolecular globular structures lacking 30-nm fibers”. In: *The EMBO journal* 35.10 (2016), pp. 1115–1132.

[53] Lu Bai and Alexandre V Morozov. “Gene regulation by nucleosome positioning”. In: *Trends in genetics* 26.11 (2010), pp. 476–483.

[54] Ashish Kumar Singh and Felix Mueller-Planitz. “Nucleosome positioning and spacing: from mechanism to function”. In: *Journal of Molecular Biology* 433.6 (2021), p. 166847.

[55] Vladimir B Teif and Karsten Rippe. “Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities”. In: *Nucleic acids research* 37.17 (2009), pp. 5641–5655.

[56] Marta Radman-Livaja and Oliver J Rando. “Nucleosome positioning: how is it established, and why does it matter?” In: *Developmental biology* 339.2 (2010), pp. 258–266.

[57] Peter B Becker and Wolfram Hörz. “ATP-dependent nucleosome remodeling”. In: *Annual review of biochemistry* 71.1 (2002), pp. 247–273.

[58] Peter B Becker and Jerry L Workman. “Nucleosome remodeling and epigenetics”. In: *Cold Spring Harbor perspectives in biology* 5.9 (2013), a017905.

[59] Janet G Yang et al. “The chromatin-remodeling enzyme ACF is an ATP-dependent DNA length sensor that regulates nucleosome spacing”. In: *Nature structural & molecular biology* 13.12 (2006), pp. 1078–1083.

[60] Maud De Dieuleveult et al. “Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells”. In: *Nature* 530.7588 (2016), pp. 113–116.

[61] Yahli Lorch, Janice W LaPointe, and Roger D Kornberg. “Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones”. In: *Cell* 49.2 (1987), pp. 203–210.

[62] Rhishikesh Bargaje et al. “Proximity of H2A. Z containing nucleosome to the transcription start site influences gene expression levels in the mammalian liver and brain”. In: *Nucleic acids research* 40.18 (2012), pp. 8965–8978.

[63] Gang G Wang, C David Allis, and Ping Chi. “Chromatin remodeling and cancer, Part II: ATP-dependent chromatin remodeling”. In: *Trends in molecular medicine* 13.9 (2007), pp. 373–380.

[64] Jakob Skou Pedersen et al. “Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome”. In: *Genome research* 24.3 (2014), pp. 454–466.

[65] Dustin E Schones et al. “Dynamic regulation of nucleosome positioning in the human genome”. In: *Cell* 132.5 (2008), pp. 887–898.

[66] Guo-Cheng Yuan et al. “Genome-scale identification of nucleosome positions in S. cerevisiae”. In: *Science* 309.5734 (2005), pp. 626–630.

[67] Artem Barski et al. “High-resolution profiling of histone methylations in the human genome”. In: *Cell* 129.4 (2007), pp. 823–837.

[68] M. W. Snyder et al. “Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin”. In: *Cell* 164.1-2 (2016), pp. 57–68.

[69] GTEx Consortium. “The Genotype-Tissue Expression (GTEx) project”. In: *Nature Genetics* 47.6 (2015), pp. 567–576.

[70] J. M. Fernandez et al. “The BLUEPRINT Data Analysis Portal”. In: *Cell systems* 3.5 (2016), 491–495.e5.

[71] Viktor A. Adalsteinsson et al. “Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors”. In: *Nature Communications* 8 (2017), p. 965.

[72] Daniel J Gaffney et al. “Controls of nucleosome positioning in the human genome”. In: *PLoS genetics* 8.11 (2012), e1003036.

[73] Joshua Moss et al. “Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease”. In: *Nature communications* 9.1 (2018), pp. 1–12.