Title

Characterizing Plasma Ultrashort Single-Stranded Cell-Free DNA in Non-Small Cell Lung Carcinoma

Author
Cheng, Jordan C

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Characterizing Plasma Ultrashort Single-Stranded Cell-Free DNA
in Non-Small Cell Lung Carcinoma

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Oral Biology

by

Jordan Chi-Ho Cheng

2023

ABSTRACT OF THE DISSERTATION

Characterizing Plasma Ultrashort Single-Stranded Cell-Free DNA in Non-Small Cell Lung Carcinoma

by

Jordan Chi-Ho Cheng

Doctor of Philosophy in Oral Biology

University of California, Los Angeles, 2023

Professor David T. Wong, Chair

Recent advances in liquid biopsy analysis have gravitated towards the utilization of cell-free DNA (cfDNA) in biofluids for cancer screening and treatment guidance. Many non-mutation aspects of cell-free DNA, such as fragment size and methylation, have demonstrated promising clinical utility for cancer detection. However, the presenting populations of cfDNA are influenced by pre-analytical steps such as DNA extraction and library preparation. We hypothesized that conventional workflows excluded a substantial portion of short fragment cfDNA in plasma. In this thesis, we detail the development of a next-generation sequencing pipeline, "Broad Range Cell-free DNA Sequencing" (BRcfDNA-Seq), which combines low-molecular weight nucleic acid extraction with a single-stranded library preparation to circumvent fragment size and strandedness limitations of conventional sequencing. In plasma, using BRcfDNA-Seq revealed the presence of ~50nt ultrashort cell-free DNA (uscfDNA) in addition to 167bp double-stranded mononucleosomal cell-free DNA (mncfDNA). Molecular and genomic analysis showcased that uscfDNA is distinct from mncfDNA in that it is single-stranded, is enriched in regulatory elements such as promoters, exons, and introns, and contains substantial G-quadruplex sequences. We examine if these unique uscfDNA features could be used as potential biomarkers to differentiate between plasma samples from non-cancer and non-small cell lung carcinoma subjects (NSCLC). We observed significant differences in functional element peaks, fragmentomics, end-motif profiles, and G-quadruplex abundance between the uscfDNA of these cohorts. Next, we investigated the methylation characteristics of uscfDNA by introducing a novel method of appending single-stranded premethylated adapters to cfDNA fragments prior to bisulfite conversion, preventing degraded genomic DNA from artificially occluding shorter cfDNA. We observed that the CpG sites of uscfDNA fragments were 15% hypomethylated compared to mncfDNA. Using a deconvolution algorithm, we inferred that uscfDNA derives from eosinophils, neutrophils, and monocytes. Later, we showed that the methylation characteristics of uscfDNA can be used to distinguish NSCLC and non-cancer subjects through differentially methylated region analysis and hypermethylated patterns of promoters, 5' UTR, and exon elements. Collectively, these studies support the uniqueness of the uscfDNA population from mncfDNA. Both genomic and epigenetic characteristics of uscfDNA demonstrate its potential clinical utility as an additional biomarker for liquid biopsy for NSCLC detection.

The dissertation of Jordan Chi-Ho Cheng is approved.

Wayne W. Grody

Udayan Guha

Yong Kim

Igor Spigelman

Benjamin M. Wu

David T. Wong, Committee Chair

University of California, Los Angeles

2023

DEDICATIONS

A kite flyer never actually flies a kite alone. During take-off, as the pilot runs forward with the string, there is a greater chance of success when another colleague runs alongside to launch the kite in the air. Even when the kite finally becomes airborne, wind, humidity, or air density ultimately will affect the flight's turbulence and duration.

In many ways, my experience as a graduate student encompasses that of a kite flyer. My interactions and experiences with my mentors and colleagues have contributed tremendously to my scientific and philosophical trajectory.

I would like to thank my advisor Dr. David Wong for his eternal optimism and enthusiasm toward the scientific pursuit.   Throughout the years, he has always given limitless time and effort toward my training and has constantly shared new ideas and opportunities, which has taken our science forward. I am continually inspired by his passion and vision. It has been this mentor-mentee synergy that has made my graduate experience both exciting and enriching.

Additionally, I would like to thank my committee members for their valuable insights, suggestions, and patience throughout my training. Our collective meetings were always an enjoyable experience. I want to thank Dr. Grody for his great positivity and support of the work, Dr. Guha for his constant willingness to support and expertise, Dr. Kim for his astuteness and impromptu scientific dialogue, Dr. Spigelman for his input, enthusiasm, and curiosity, and Dr. Wu for providing dynamic inquiries and invigorating words.

Dr. Chia has provided countless patience and guidance throughout my time at UCLA. His vast perspective consistently opens my eyes to other aspects of biology that I should consider.

My gratitude goes to Dr. Ravi Shah for offering guidance along each step of the Ph.D. journey. The countless conversations regarding central and peripheral topics have cultivated my curiosity to learn about nature and the humans that interact with it. I now realize the importance of being both the bird and the frog. Thank you for the reassurance to fearlessly walk through the next door that opens.

Over the past six years, I've tremendously enjoyed working with colleagues locally and internationally. There have been many intriguing discussions that stimulate further creativity and vigor. In no particular order, I would like to thank Marco Morselli, Matteo Pellegrini, Allen Huang, Samantha Chiang, Neeti Swarup, Taichiro Nonaka, Susan Kim, Karolina Kaczor-Urbanowicz, Shannon Rao, Sandra Perez, Feng Li, Feng Wei, Michael Tu, and Mohammad Aziz. Thank you for your friendship, insights, support, and encouragement.

To my mom, dad, sister, and Putt, I want to thank you for supporting me through this unknown and transformative journey. Knowing your love and support, and confidence has always kept me inspired and moving forward.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

My graduate studies and part of this work were funded by the UCLA Clinical and Translational Science Institute Core Voucher, Canadian Institute of Health Research Doctoral Foreign Study Award, Tobacco-Related Disease Research Program (TRDRP) Predoctoral Fellowship, Jonsson Comprehensive Cancer Center Predoctoral Fellowship, and the National Cancer Institute F99/K00 Predoctoral to Postdoctoral Fellowship.

Aspects of Chapter 1 are published in the chapter "ctDNA and Lung Cancer." Cheng J, Hu Y, Patel AA, Wong DTW. *Circulating Tumor Cells/ Advances in Liquid Biopsy Technologies*, Second Edition. Springer, Cham, 511-537, 2023. I was involved in the literature review and drafting of the manuscript. Hu Y contributed to the clinical implications. Wong DTW and Patel AA were the corresponding PIs and were involved in all chapter elements.

Chapter 2 is published as "Plasma contains ultrashort single-stranded DNA in addition to nucleosomal cell-free DNA." Cheng J, Morselli M, Huang WL, Heo YJ, Pinheiro-Ferreira T, Li F, Wei F, Chia D, Kim Y, He HJ, Cole KD, Su WC, Pellegrini M, Wong DTW. *iScience*. 2022 Jul 15;25(7):104554. doi: 10.1016/j.isci.2022.104554. Morselli M and I were involved in developing BRcfDNA-Seq through experimental design, data analysis, and interpretation. Heo YJ provided bioinformatic support. Pinheriro-Ferreira T and Huang WL generated supporting data. Li F, Wei F, Chia D, Kim Y, He HJ, Cole KD, and Su WC contributed to the data interpretation and manuscript preparation. Wong DTW and Pellegrini M served as the corresponding PIs and were engaged in the various steps of the project.

Chapter 3 has been accepted as "Distinct Features of Plasma Ultrashort Single-Stranded Cell-Free DNA as Biomarkers for Lung Cancer Detection". Cheng J, Swarup N, Li F, Kordi M, Lin

CC, Yang SC, Huang W, Aziz M, Kim Y, Chia D, Yeh YM, Wei F, Zheng D, Zhang L, Pellegrini M, Su W, Wong DTW. *Clinical Chemistry*. Swarup N and I lead the experimental design and analysis. Li F and Aziz M were involved in the library preparation and sequencing. Kordi M and Zheng D were involved in the initial bioinformatic analysis. Huang W, Lin CC, Yang SC, and Su Wu coordinated the sample procurement. Kim Y, Chia D, Yeh YM, Wei F, and Zhang L contributed to the data interpretation and manuscript preparation. Wong DTW and Su W served as the lead PIs and was involved in all aspects of the study.

Chapter 4 is in preparation for publication. Swarup N, Morselli M, and I were involved in study design, library preparation, and bioinformatic analysis. Wong DTW and Pellegrini M served as the corresponding PIs and were engaged in the various steps of the research.

CURRICULUM VITA

EDUCATION

2012	B.Sc. (Hon.) Physiology, University of British Columbia, Vancouver, Canada
2017	Doctor of Dental Medicine, University of British Columbia, Vancouver, Canada

SELECTED AWARDS, SCHOLARSHIPS, FELLOWSHIPS

2019-2021	Doctoral Foreign Study Award Fellowship, Canadian Institute of Health Research, Canada
2019-2021	Predoctoral Fellowship, Tobacco-Related Disease Research Program, Regents of the University of California, United States
2020-2021	Fellowship Award, Jonsson Comprehensive Cancer Center, University of California, United States
2020-2021	Core Services Voucher, UCLA Clinical and Translational Science Institute (CTSI), University of California, United States
2021-2026	NCI F99/K00 Predoctoral to Postdoctoral Fellow Transition Award, Bethesda, United States

PUBLICATIONS

SELECTED MANUSCRIPTS

1. Wei F, Strom CM, Cheng J*, Lin CC, Hsu CY, Soo Hoo GW, Chia D, Kim Y, Li F, Elashoff D, Grognan T, Tu M, Liao W, Xian R, Grody WW, Su WC, Wong DTW. Electric Field-Induced Release and Measurement Liquid Biopsy for Noninvasive Early Lung Cancer Assessment. J Mol Diagn. 2018 Nov;20(6):738-742. *Joint First Author
2. Borsetto D, Cheng J*, Payne K, Nankivell P, Batis N, Rao K, Bhide S, Li F, Kim Y, Mehanna H, Wong D. Surveillance of HPV-Positive Head and Neck Squamous Cell Carcinoma with Circulating and Salivary DNA Biomarkers. Crit Rev Oncog. 2018;23(3-4):235-245. *Joint First Author
3. Cheng J, Nonaka T, Wong DTW. Salivary Exosomes as Nanocarriers for Cancer Biomarker Delivery. Materials (Basel). 2019 Feb 21;12(4).
4. Kaczor-Urbanowicz KE, Wei F, Rao SL, Kim J, Shin H, Cheng J, Tu M, Wong DTW, Kim Y. Clinical validity of saliva and novel technology for cancer detection. Biochim Biophys Acta Rev Cancer. 2019 Aug;1872(1):49-59.
5. Wang Z, Li F, Rufo J, Chen C, Yang S, Li L, Zhang J, Cheng J, Kim Y, Wu M, Abemayor E, Tu M, Chia D, Spruce R, Batis N, Mehanna H, Wong DTW, Huang TJ. Acoustofluidic Salivary Exosome Isolation: A Liquid Biopsy Compatible Approach for Human Papillomavirus-Associated Oropharyngeal Cancer Detection. J Mol Diagn. 2020 Jan;22(1):50-59

6. Tu M, Cheng J*, Chen YL, Jea WC, Chen WL, Chen CJ, Ho CL, Huang WL, Lin CC, Su WC, Ye Q, Deignan J, Grody W, Li F, Chia D, Wei F, Liao W, Wong DTW, Strom CM. Electric Field-Induced Release and Measurement (EFIRM): Characterization and Technical Validation of a Novel Liquid Biopsy Platform in Plasma and Saliva. J Mol Diagn. 2020 Aug;22(8):1050-1062. *Joint first author
7. Kaczor-Urbanowicz KE, Cheng J, King JC, Sedarat A, Pandol SJ, Farrell JJ, Wong DTW, Kim Y. Reviews on Current Liquid Biopsy for Detection and Management of Pancreatic Cancers. Pancreas. 2020 Oct;49(9):1141-1152.
8. Kim C, Xi L, Cultraro CM, Wei F, Jones G, Cheng J, Shafiei A, Pham TH, Roper N, Akoth E, Ghafoor A, Misra V, Monkash N, Strom C, Tu M, Liao W, Chia D, Morris C, Steinberg SM, Bagheri H, Wong DTW, Raffeld M, Guha U. Longitudinal Circulating Tumor DNA Analysis in Blood and Saliva for Prediction of Response to Osimertinib and Disease Progression in EGFR-Mutant Lung Adenocarcinoma. Cancers (Basel). 2021 Jul 3;13(13).
9. Cheng J, Morselli M, Huang WL, Heo YJ, Pinheiro-Ferreira T, Li F, Wei F, Chia D, Kim Y, He HJ, Cole KD, Su WC, Pellegrini M, Wong DTW. Plasma contains ultrashort single-stranded DNA in addition to nucleosomal cell-free DNA. iScience. 2022 Jul 15;25(7):104554.
10. Cheng J, Swarup N, Li F, Kordi M, Lin CC, Yang SC, Huang W, Aziz M, Kim Y, Chia D, Yeh YM, Wei F, Zheng D, Zhang L, Pellegrini M, Su W, Wong DTW. Distinct Features of Plasma Ultrashort Single-Stranded Cell-Free DNA as Biomarkers for Lung Cancer Detection. (In Press).
11. Swarup N, Cheng J*, Choi I, Heo YJ, Kordi M, Li F, Aziz M, Chia D, Wei F, Elashoff D, Zhang L ,Kim S, Yong Kim, and Wong DTW. Multi-Faceted Attributes of Salivary Cell-free DNA as Liquid Biopsy Biomarkers for Gastric Cancer Detection. (In Review). *Joint first author

BOOK CHAPTERS

1. Cheng J, Nonaka T, Ye L, Wei F, Wong D. Salivaomics, Saliva-Exosomics, and Saliva Liquid Biopsy. In: Granger D., Taylor M. (eds) Salivary Bioscience. Springer, Cham, 157-175, 2020.
2. Cheng J, Hu Y, Patel AA, Wong D. ctDNA and Lung Cancer. Circulating Tumor Cells/ Advances in Liquid Biopsy Technologies, Second Edition. Springer, Cham,, 511-537, 2023.

SELECTED PATENTS AND INTELLECTUAL PROPERTY

US 17/988,051	LIQUID BIOPSY PLATFORM IN PLASMA AND SALIVA
US 63/373,369	NEXT-GENERATION SEQUENCING PIPELINE TO DETECT ULTRASHORT SINGLE-STRANDED CELL-FREE DNA
US 63/503,842	PHYSICAL SEPARATION OF ULTRASHORT SINGLE-STRANDED CELL-FREE DNA IN BIOFLUIDS USING DIFFERENTIAL FILTRATION
UCLA 2023-082-1	NEXT-GENERATION SEQUENCING PIPELINE TO DETECT METHYLATION STATUS OF ULTRASHORT SINGLE-STRANDED CELL-FREE DNA

1

INTRODUCTION

1.1 Introduction

1.1.1 Lung Cancer

Globally, lung cancer is the number one leading cause of cancer-related deaths (Sung et al., 2021). Lung cancer can be categorized into two broad types: small cell lung carcinoma (SCLC), which approximately makes up 15% of the cases, and non-small cell lung carcinoma (NSCLC), which makes up the remaining 85% of the cases and can be further stratified into adenocarcinoma and squamous cell carcinoma (Herbst et al., 2018). In cases where NSCLC is identified at an early stage, treatment with surgical resection of the NSCLC has been associated with favorable outcomes. The 5-year survival rate for small localized lesions is as high as 70-90% (Nesbitt et al., 1995; Shah et al., 1996). Most patients, however, present at stages III or IV, and despite developments in oncological management, survival rates remain guarded (Simmons et al., 2015).

In contrast to NSCLC, SCLC behaves even more aggressively, with a poorer prognosis and an overall 5-year survival of 5%. At the time of discovery, the majority of patients (90%) present with locally advanced or distant metastatic disease (stage III/IV) (S. Wang et al., 2017). Despite having a small window for aggressive treatment, surgery for stage I disease can still demonstrate positive outcomes (Harris et al., 2012). 5-year survival rates of early-stage SCLC have been reported to reach 40% with surgery and 52% with adjuvant chemotherapy/radiotherapy (Yang et al., 2016). It is clear that for both subtypes of lung cancer, early-stage disease detection and screening, are crucial for improving clinical outcomes.

1.1.2 Biology of Early Lung Cancer

One strategy to identify useful biomarkers for early detection is to understand the pathophysiology of the changes in the lung environment during the initial stages of the disease. The lung is constantly exposed to a dynamic external environment, facilitating ventilation of air and waste exchange. The airway branches are lined with pseudostratified epithelium filled with ciliated cells and secretory cells with stem-cell abilities that can regenerate and repair the airway following injuries (Hogan et al., 2014). The alveoli are lined with two cell types – the squamous type I alveolar epithelial cells, which comprise 90% of alveolar coverage and are responsible for capillary interaction. Second, the cuboidal type II alveolar epithelial cells secrete lipids and proteins to reduce surface tension during ventilation. Interestingly, cuboidal type II alveolar epithelial cells can act as repair precursors for type I alveolar epithelial cells (Barkauskas et al., 2013).

Lineage tracing experiments indicate that cells responsible for airway repair may be responsible for initiating cells of tumors in the lung. In particular, neuroendocrine cells, which are rare secretory cell populations of the conducting airways, are potential trigger points for SCLC (Karachaliou et al., 2016; Sutherland et al., 2011). For NSCLC, the majorly described cell of origin for Kirsten rat sarcoma virus (KRAS)-driven adenocarcinomas are alveolar type II epithelial cells (Hanna & Onaitis, 2013). The basal cells of the trachea are hypothesized as the cell of origin for squamous cell carcinoma (Hanna & Onaitis, 2013; Hong et al., 2004) are basal cells since they have been shown to over-express, which could be eventually oncogenic (Giangreco et al., 2012; Lu et al., 2010).

The association between chronic, long-term inflammation and the increased risk of cancer development has also been proposed (Crusz & Balkwill, 2015; Kundu & Surh, 2008). Although chronic inflammation may be the initial trigger for only 20% of cancer, innate immune cells, and their associated mediators are found in almost all human malignancies (Mantovani et al., 2008). Inflammatory pathways interplay between pre-malignant and malignant cells since inflammation can cause cancer states that propagate to trigger further inflammation. The developing tumor microenvironment incorporates signals from inflammatory cells and their cytokine, chemokine, and prostaglandin mediators, which affect the behavior of malignant and non-malignant cells (Mantovani et al., 2008). An inflammatory microenvironment can then promote the activity of tumor infiltration inflammatory cells, tumor-associated fibroblasts, and endothelial progenitor cells (Balkwill & Coussens, 2004; Coussens & Werb, 2002). Reported factors in this environment include tumor necrosis factor-alpha (TNF-a), interleukins 6, 1A, inflammatory chemokine CCL2, and CXCL12-CXCR4 signaling cascades (Ancrile et al., 2007;

D'Alterio et al., 2012; Sanmamed et al., 2014; Singer et al., 2003; J. Zhang et al., 2010; Y. M. Zhu et al., 2004). These factors can produce an inflammatory-associated immune response, recruit inflammatory cells, and promote cell growth, survival, and angiogenesis for emerging cancer cells.

1.1.3 New Paradigm for Air Pollution and Adenocarcinoma Promotion

Recent reports have slightly altered the perceived chain of events of tumor development. Traditionally, it was conceptualized that carcinogens promote tumors by directly inducing DNA damage. Recent studies propose that the majority of carcinogens do not cause detectable DNA damage following exposure (Kucab et al., 2019). In contrast, environment particulate matter measuring <2.5 μm ($PM_{2.5}$) can promote lung cancer by manipulating cells that harbor pre-existing oncogenic mutations. By demonstrating in a mouse model, $PM_{2.5}$ draw an influx of macrophages into lung tissue, releasing interleukin-1B, thereby encouraging lung alveolar type II epithelial cells to transform into a progenitor-like cell state that exacerbates tumorigenesis. They found reported that 295 non-cancer individuals across three clinical cohorts were found to have epidermal growth factor receptor (EGFR) (18%) and KRAS (53%) mutations in their healthy tissue samples (Hill et al., 2023).

The proposed interactions describe how tumors, environments, and immune systems are linked and contribute during the early stages of cancer development. Understanding these mechanisms could aid in yielding potential biomarker targets for lung cancer detection.

1.2 Methods for Lung Cancer Screening

1.2.1 Chest X-Ray for Lung Cancer Screening

Several randomized control studies in the 1970s examined the use of chest X-rays and sputum cytology for early cancer detection. There did not appear to be any difference in mortality between tri-annual screening versus annual screening (Fontana et al., 1975). Two studies compared plain chest X-ray with or without sputum cytology, showing that when both are used, 20% are detectable by cytology alone and were determined to be early-stage squamous cell carcinoma (Frost et al., 1984; Melamed et al., 1984). These patterns were further supported by a prostate, lung, colorectal, and ovarian cancer screening trial, suggesting minor mortality benefits with X-ray methodology (Oken et al., 2011).

1.2.2 Low-dose CT Screening

The emergence of computerized tomography (CT) provided more detailed images of the chest region than the chest X-ray, making it a potentially useful tool for cancer detection. However, the CT was accompanied by a 100-fold greater radiation dose than the chest X-ray, so the potential for early diagnosis had to be balanced by increased radiation exposure. In 1990, CT was validated to be used at a lower radiation dose (22% of the standard amount (Larke et al., 2011)) and titled "low dose CT" (LDCT), rejuvenating interest in using CT as a screening mechanism (Naidich et al., 1990).

Initial studies focused on at-risk populations defined according to age and smoking (20-30 pack years). Two Italian studies (DANTE (Infante et al., 2009) (n=2472) MILD (Pastorino et al., 2012) (n=4099)) and a Danish-lead study (DLSCST (Saghir et al., 2012) n =4104) compared

LDCT with a control arm with yearly medical reviews. Although increased detection of early-stage lung cancer lesions occurred, there was no reduction in mortality. Later, a seminal study with a larger population (n = 53454 participants) at risk of lung cancer (55-74 years, > 30 pack years smoked within 15 years) to either annual LDCT or chest X-ray demonstrated a 20% reduction in cancer mortality and total mortality of 6.7% (National Lung Screening Trial Research Team et al., 2011). Compared to the earlier studies, there was also a decrease in late-stage diagnosis showing the influence of early-stage disease detection. This finding triggered the US Preventative Task Force (USPSTF) to publish recommendations for LDCT screening and raise the upper age of screening to 80 years. Although relatively successful, the NLST study raised several issues. Two hundred thirty-one early stages (stage 1A and 1B) were diagnosed in the LDCT group, where 93% of cases were resected. However, only 79 fewer lung cancer-related deaths were recorded in the LDCT group. This finding may have resulted from a relatively high recurrence rate even after successful surgeries. LDCT is also prone to high positivity rates where any nodules larger than 4mm are referred for further investigation despite not being lung related.

1.2.3 Bronchoscopy Screening

Bronchoscopy is another strategy that has a role in the early detection of lung cancer but only a sensitivity of detection from 35% to 88% depending on the size and position of the tumor (Rivera et al., 2013). When applied in a screening context for patients at risk of lung cancer without suspicious radiological imaging, the sensitivity suffers even more. Bronchoscopy appears to perform better by combining it with RNA expression of the histologically normal bronchial epithelium sampled at the time of bronchoscopy. For example, in one study, an 80-gene expression classifier trained on 77 smokers with or without lung cancer demonstrated a sensitivity of 80% and specificity of 84% in an independent validation cohort (Spira et al., 2007). This supported the concept that combining molecular clues with histological methods would be helpful.

1.3 Liquid Biopsy

1.3.1 Liquid Biopsy Introduction

Liquid biopsy is the sampling of non-solid biological tissue, such as blood, for biomarkers that indicate aspects of cancer status. Potential biomarkers in the blood include circulating nucleic acids (cfDNA), proteins, or circulating tumor cells (Ignatiadis et al., 2021). The advantages of liquid biopsy derive from its noninvasive nature and that it is highly repeatable compared to surgical biopsies. These attributes allow liquid biopsy to be potentially useful for screening, treatment monitoring, and disease tracking (Rolfo et al., 2018). Although other liquid biopsy biomolecular types have been explored extensively in the literature, this thesis will focus primarily on cell-free DNA.

1.3.2 History of Cell-free DNA

Extracellular nucleic acids were first observed in the plasma by the French scientists Mandel and Métais several years before even elucidating the double helix model of DNA (Mandel,P & Metais, P, 1948). However, little attention was paid to this finding until later. Within several decades, the serum of cancer patients was demonstrated to contain higher concentrations of cell-free DNA in comparison to healthy serum (Leon et al., 1977). Further investigation showed that plasma cell-free DNA from cancer patients possessed greater double-strand instability (Stroun et al., 1989). As technology advanced with polymerase chain reaction (PCR) capability, tumor-specific gene aberrations were detected in the cell-free DNA of blood plasma in patients with myelodysplastic syndrome and acute myelogenous leukemia (Vasioukhin et al., 1994). This work served as the initial proof of concept that circulating tumor DNA (ctDNA) – the tumor-derived portion of cell-free DNA (cfDNA) – can potentially be used to non-invasively evaluate a tumor's genetic features. Soon after, it was discovered that fetal (placental) DNA could be identified in the maternal circulation during pregnancy, eventually leading to applications that enabled the detection of genetic anomalies in the fetus without requiring invasive amniocentesis protocols (Lo et al., 1997). The concept of liquid biopsy has since expanded to include other informative analytes (e.g., microRNAs, exosomes, circulating tumor cells, etc.) in various biofluids for cancer detection.

1.3.3 Cell-free DNA Biology

Sources of cfDNA

Understanding the origins and features of cfDNA is critical to the rational development of clinical applications. Although many details of cfDNA biogenesis remain unclear, multiple lines of evidence indicate that cfDNA is shed into the bloodstream from numerous cell types and from several physiological and pathological processes (Figure 1.1). For example, during intense exercise and psychosocial stress, cell-free DNA is observed to increase in the circulation (Hummel et al., 2018). The most often described form of cfDNA is a linear double-stranded DNA of 150–180 base pair in length. This length corresponds to 147bp of DNA that is wrapped around a nucleosome particle plus a short, variable linker segment of DNA that stretches between nucleosomes (Diaz & Bardelli, 2014; Fan et al., 2008). This cfDNA structure is often represented by a 167bp peak in size-fragment diagrams called mononucleosomal cell-free DNA (Sanchez et al., 2021). Adherence to the nucleosome particle protects cfDNA from nuclease-mediated degradation in circulation. In healthy individuals, the majority of cfDNA originates from the physiologic turnover of hematopoietic cells via apoptosis. However, in patients with solid malignancies, typically, a small proportion of cfDNA is contributed by cancer cells undergoing apoptotic or necrotic cell death (Lui et al., 2002; Razavi et al., 2019).

Lipoprotein-associated cfDNA

The electrostatic properties of nucleic acids promote their binding to circulating proteins such as albumin, immunoglobulins, fibronectin, or C1q complements (Chelobanov et al., 2006; Rykova et al., 2012; Rykova EYu et al., 1994). DNA is also found on the cell surface (Bryzgunova et al., 2015). One study showed that cultured cells could have DNA fragments as large as 20kbps on the cell surface which require mild trypsin treatment to completely detach, suggesting an intended anchoring (Morozkin et al., 2004). The concept of "virtosomes," which are non-membranous macromolecular DNA/RNA-lipoprotein complexes, has been coined by Anker, Stroun, and Gahan which may play a physiological role in cellular homeostasis as an intercellular messenger (Gahan & Stroun, 2010). There is some evidence that they are released in an energy-dependent step from livings cells in a controlled manner, suggesting a role in cell-to-cell communication (Adams et al., 1997; Adams & McIntosh, 1985).

Cell-free Mitochondrial DNA

Cell-free mitochondrial DNA is another source of cfDNA found in circulation. In contrast to the nuclear genome, the mitochondrial genome is only 16,000 bp in length, circular, and unprotected by histones. Similar to nuclear DNA, however, cell-free mitochondrial DNA is thought to be released from apoptotic or necrotic cells (Kohler et al., 2009). In circulation, mitochondrial DNA is highly fragmented, appearing as 30-60bp fragments at very high copy numbers due to their short-length genomes (R. Zhang et al., 2016). Mitochondrial DNA exists as naked DNA or can be associated with internal or external membrane fragments(Chiu et al., 2003). Cell-free mitochondrial DNA has been explored as a possible biomarker for several disease conditions, including cancer, stroke, and myocardial infarction (Kohler et al., 2009; Rainer et al., 2003; L. Wang et al., 2015).

Neutrophil and Eosinophil Extracellular Traps

Another pathway in which nucleic acids can enter the circulation is by immune cell release of extracellular nucleic acid traps. Neutrophils release neutrophil extracellular traps (NETs), which can capture and kill bacteria and pathogens in a process called netosis (de Bont et al., 2019). Netosis is a complex process requiring chromatin recondensation and lysis of nuclear and cell membranes to release the extracellular DNA trap. Netosis appears to be essential for innate immunity. It is also associated with autoimmune inflammatory responses, thrombotic disease, sepsis, and cancer (Fuchs et al., 2010; Kaplan & Radic, 2012; Luo et al., 2014). Eosinophils have also been shown to release extracellular DNA traps but uniquely release exclusively mitochondrial DNA, which could be another source of cfDNA in the circulation (Yousefi et al., 2008). In cancer, NETs have been shown to sequester with circulating tumor cells suggesting that an increase in NETs concentration could indirectly signal for the progression of cancer (Cools-Lartigue et al., 2013).

Extrachromosomal Circular DNA

Small circular-form DNA known as extrachromosomal circular DNA (ecDNA) has been observed intracellularly and extracellularly. These extrachromosomal circular DNAs are found in two classes: very small-sized microDNA, which is usually less than 10 kB in length, or larger ecDNA, which can be >1 MB in length (Yan et al., 2020). It has been shown that several hundred ecDNAs can exist in a cell. Examination of the ecDNA in cancer cells has revealed that they often contain homeostatic genes, regulatory regions, or oncogenes sequences. EcDNA has been shown to have high transcriptional accessibility, suggesting that higher expression of genes on ecDNA may provide survival or proliferation advantages to cancer cells. Cell-free ecDNA is highly consistent with the tumor burden in lung cancer patients (Kumar et al., 2017). Although ecDNA concentration in the circulation is much lower than that of linear cfDNA, it nonetheless holds promise as a cancer biomarker (Sin et al., n.d.; J. Zhu et al., 2018)

Extracellular Vesicles

Another source of DNA in the circulation is extracellular vesicles (EVs). EVs can be categorized as exosomes, microvesicles, and apoptotic bodies (Fernando et al., 2017). EVs contain material reflecting the contents of their cell and tissue sources, such as proteins, lipids, mRNA, and microRNA (Tetta et al., 2013). Cancer cells demonstrate a higher level of extracellular vesicle release (Xavier et al., 2020). Double-stranded DNA has been found both on the outside (>2.5kBp) and inside (100-2.5kbp) of exosomes (Thakur et al., 2014). Mitochondrial DNA was also found in a study of glioblastoma and astrocyte exosomes

(Guescini et al., 2010). Mutations in KRAS and p53 have been detected in circulating exosomes of patients with pancreatic cancer (Kahlert et al., 2014). Single-stranded DNA encoding oncogenes can be identified in microvesicles (200-1000 nm in diameter) released in the serum of glioblastoma tumor-bearing mice (Balaj et al., 2011). Larger apoptotic bodies between 1 to 5 μm have been found to contain cytoskeletal elements and degraded chromosomal DNA (Kakarla et al., 2020).

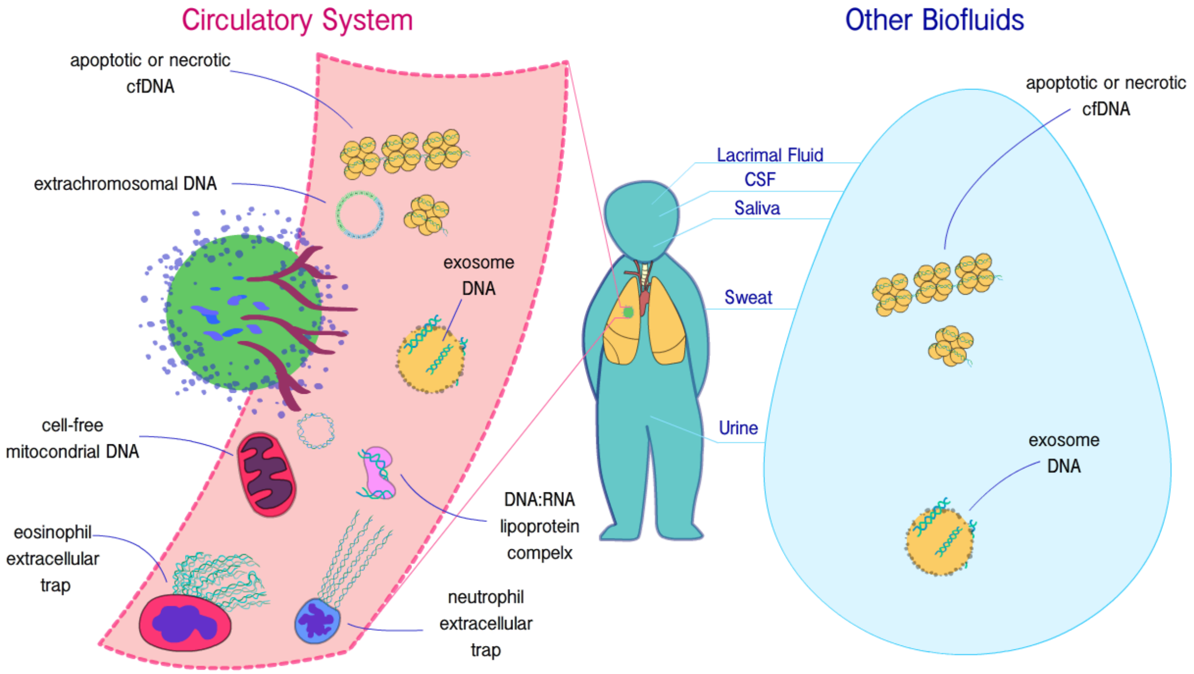
Circulatory System
Other Biofluids
apoptotic or necrotic cfDNA
extrachromosomal DNA
exosome DNA
cell-free mitochondrial DNA
eosinophil extracellular trap
DNA:RNA lipoprotein compelx
neutrophil extracellular trap
Lacrimal Fluid
CSF
Saliva
Sweat
Urine
apoptotic or necrotic cfDNA
exosome DNA

Figure 1.1. Sources of cell-free DNA in blood and other biofluids. Various cellular sources contribute to the pool of observed cell-free DNA in circulatory blood and other biofluids. Cell-free DNA can originate from hemopoietic cells, healthy solid tissue cells (including the placenta), or tumor cells.

1.4 Circulating Tumor DNA in Lung Cancer

The dysregulated proliferation of cells in non-small cell lung cancer is often driven by genetic aberrations, which if identified, can be targeted with specific medications such as tyrosine kinase inhibitors (TKIs). For example, approximately 20% of patients with NSCLC adenocarcinoma possess somatic mutations in the ATP-binding region of the epidermal growth factor receptor gene (EGFR) (Kris et al., 2014) leading to Ras/Raf/MAPK and PI3K pathway activation and driving carcinogenesis. TKIs such as gefitinib, erlotinib, and osimertinib target these dysregulated signaling pathways and have demonstrated excellent clinical efficacy (Rosell et al., 2012). However, in virtually all patients treated with such TKIs, resistance eventually emerges, often driven by new mutations (e.g., T790M from gefitinib/erlotinib treatment and C797S from osimertinib treatment).

Since the somatic mutation status of the tumor is correlated with treatment efficacy, clinical guidelines recommend treatment selection based on mutation profiling from a tissue biopsy (Ettinger et al., 2017). Obtaining a tissue biopsy, however, can sometimes lead to clinical complications (such as pneumothorax) and will sometimes yield insufficient tissue for molecular testing (Boskovic et al., 2014). Biofluid-based testing, more commonly known as liquid biopsy, has enabled the noninvasive determination of tumor-derived somatic mutations in EGFR and other NSCLC driver mutations from fluids such as blood and saliva. In addition to cell-free DNA, there is a myriad of cancer-associated biomarkers in biofluids that can be harvested as informative liquid biopsy signals. Peripheral blood proteins, circulating tumor cells, exosomes, platelet RNA, and microRNAs potentially harbor information that can aid the diagnosis and management of NSCLC (Rolfo et al., 2018). However, the tumor-derived component of cell-free DNA (cfDNA), known as circulating tumor DNA (ctDNA), has been heavily investigated and has matured into a clinically valuable biomarker for NSCLC detection. As sequencing information can be derived from these fragments, it is a promising companion diagnostic tool alongside tissue biopsy.

1.5 Potential Non-somatic Mutation-based Biomarker Features of cfDNA

1.5.1 Introduction

Despite the many virtues of ctDNA, its detection is as difficult as finding a needle in a haystack. ctDNA is present at extremely low concentrations compared to cfDNA of non-tumor origin. Ratios range from >5-10% in late stages to <0.01 to 0.1% in early stages (or after surgical intervention)(Bettegowda et al., 2014). As an alternative, non-ctDNA-based liquid biopsy markers have emerged as potential promising contributors. These features of cfDNA: fragmentomics, END-motifs, and topological characteristics, can potentially be done in conjunction with ctDNA analysis to bolster information for cancer detection.

1.5.2 Fragment Size and Fragmentomics

Many researchers have shown observed that size information could help achieve improved performance for cancer detection. For example, the positive predictive value of detected tumor-derived mutations in hepatocellular carcinoma patients could be improved by 85% if the shorter size of tumor-derived DNA was considered (Jiang et al., 2015). This strategy could also distinguish clonal hematopoiesis from tumor-derived mutations in the plasma (Marass et al., 2020). Another group showed that physically selecting for short DNA molecules will enrich the ctDNA ratio compared to cfDNA (Mouliere et al., 2018). Typically, ctDNA has been continuously reported present with a fragment size in the range of ~150-180bp, corresponding to the length of DNA wrapped around the mononucleosomal complex with or without the linker DNA (Jahr et al., 2001; Sanchez et al., 2021; Snyder et al., 2016). Necrosis-derived cell-free DNA is believed to be released from cells in a longer form (>1kb), but recent evidence suggests that such DNA becomes fragmented by nucleases in the circulation to a size distribution resembling apoptotic derived cfDNA (Rostami et al., 2020). Tumor-derived DNA fragments in the circulation have been found to have a size distribution that is shifted smaller in some (but not all) tumors. Indeed, a recent study found that tumor-derived DNA sequences were enriched within cell-free DNA fragments in the size range of 90-150 bp (Mouliere et al., 2018). By borrowing single-strand library preparation technology from paleogenomics which incorporates damaged or nicked DNA (Gansauge & Meyer, 2013), Shendure and colleagues showed that a high proportion of cfDNA fragments are shorter than 160bp length (Snyder et al., 2016). Additional studies using similar ssDNA library preparation methods confirmed this observation and revealed the presence of both shorter nuclear cfDNA and circulating mitochondria DNA <100bp (Burnham et al., 2016). In one study, a genome-wide assessment of cfDNA fragment size revealed substantial variability in size across different genomic regions but remarkable consistency within a given genomic region across different healthy individuals (Cristiano et al., 2019). Deviations from such genomic region-specific fragmentation size patterns are observed in patients with cancer, and this signal is being probed to enable early cancer diagnosis in an approach called DELFI (DNA Evaluation of fragments for early interception) (Cristiano et al., 2019). The concept of fragmentomics has also demonstrated proof-of-concept in liver cancer (Foda et al., 2022), lung cancer (S. Wang et al., 2023), and osteosarcoma (Udomruk et al., 2023).

1.5.3 Single-stranded Vs. Double-stranded cfDNA Ratios

Investigators have also examined how DNA strandedness could be a valuable feature of cfDNA for cancer differentiation. A colorectal (Song et al., 2021) and gastric cancer (Huang et al., 2020) study revealed that increased ssDNA concentration was associated with cancer recurrence. Whereas in the other study, gastric cancer samples had a lower single-stranded to double-stranded ratio than healthy individuals. While most studies in the field have focused on the analysis of double-stranded cell-free DNA fragments, there is growing interest in probing for cancer-specific signals in single-stranded DNA fragments.

1.5.4 Preferred and Fragment Ends

Since there appears to be a relationship between cell-free DNA from different tissue sources and their size, it has been hypothesized that these molecules might have different ends than the DNA from blood cells usually present in the blood. Deep sequencing of plasma DNA ends shows that DNA fragmentation is a nonrandom process in which certain genomic regions are more prone to be cleaved and found at the end of plasma DNA fragments called "preferred end sites" (Chan et al., 2016). Using liver cancer and liver transplant as clinical proof-of-concept models, tumor-associated preferred ends are more pervasive than mutations alone and thus could be an additional cfDNA signature that is useful in cancer detection.

Additionally, double-stranded cell-free DNA fragments have been shown to commonly have single-stranded ends (overhangs) in both plasma (Jiang, Xie, et al., 2020) and urine (Zhou et al., 2021). Such DNA ends have been referred to as "jagged ends," and it has been shown that tumor-derived cfDNA fragments appear to have increased "jaggedness" compared to non-tumor cfDNA. Investigators have also studied the nucleotides at the proximal 5′ end of the cfDNA molecule, which is called k-mer end motifs (k being any length of bp). These studies have shown that the 5′end motifs preferentially start with the C nucleotides (Jiang, Sun, et al., 2020; Serpas et al., 2019). Mechanistic studies involving nuclease knock-out mouse models revealed that plasma size and end motif were related to DNA nuclease activity (Serpas et al., 2019). The top six 4-mer end motifs all started with "CC" in wild type mice, which would decline in Dnase1L3 mice. Alterations in Dnase1l3 "CC" signatures were present in human subjects with DNase1l3 deficiency. This pattern was also found in patients with familiar systematic lupus erythematosus (SLE) and many cancers are also associated with downregulated DNase 1L3 expression. Some cancers described were hepatocellular carcinoma, colorectal cancer, lung cancer, nasopharyngeal cancer, and head and neck squamous cell carcinoma(Jiang, Sun, et al., 2020). Using all 256 signatures could achieve an AUC of 0.86 with and without cancers. Other groups have also started using these with some success (Bao et al., 2022; Guo et al., 2022; Zhitnyuk et al., 2022). These nucleases-associated cutting patterns provide another diagnostic tool for monitoring disease based on DNase aberrations.

1.5.5 Cell-free DNA Methylation

Methylation of DNA is an epigenetic modification that is found most at the fifth carbon of cytosine nucleobases (producing 5-methylcytosine). Such methylation typically occurs at cytosine residues in the sequence context of 5′ -C-phosphate-G-3′ (CpG) and is mediated by a family of enzymes known as DNA methyltransferases (Jones, 2012). In most regions of the genome of human somatic cells, CpG sites are highly methylated (~75%) (Smith & Meissner, 2013). In contrast, regions of high CpG density, known as CpG islands (CGIs), are typically hypomethylated (Suzuki & Bird, 2008). Patterns of methylation across the genome are highly consistent in the cell-free DNA of healthy individuals, and aberrant methylation patterns become apparent in cfDNA derived from cancer cells (Chan et al., 2013). Cancer-specific alterations in methylation include both hypermethylation of CpG islands (often as a means of silencing gene expression) and hypomethylation, which occurs more broadly across the genome (contributing to genomic instability) (Esteller, 2008). Examination of cancer-specific methylation patterns in cell-free DNA has emerged as a promising approach for cancer detection, determining tissue of origin, and identifying minimal residual disease following therapy (M. C. Liu et al., 2020; Parikh et al., 2021; Shen et al., 2018).

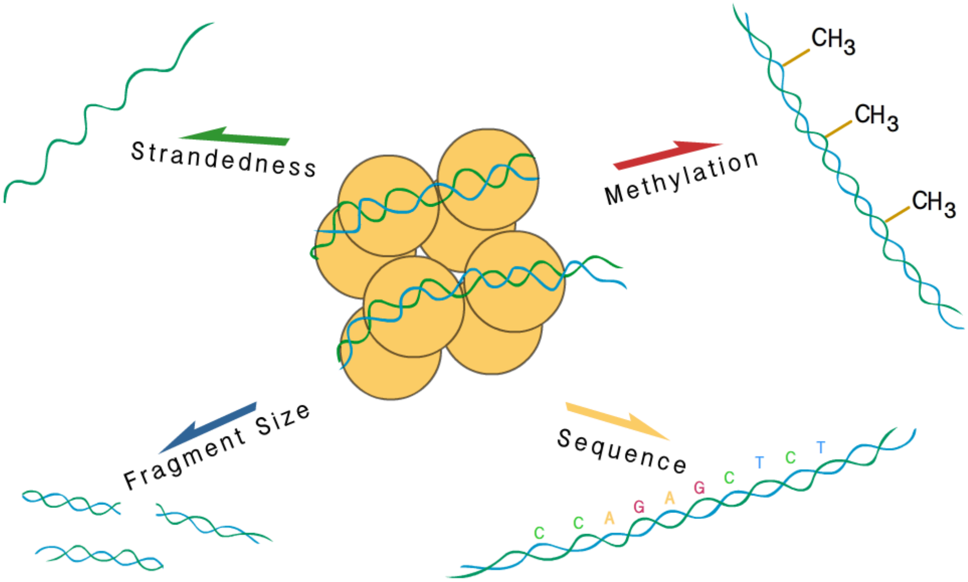

Figure 1.2 Characteristics of cell-free DNA that can be utilized in liquid biopsy detection of cancer. Cancer-specific signatures of cell-free DNA can be found in features that include sequences (somatic mutations, end-motif, G-Quad), methylation patterns, fragment size differences, and strandedness.

1.6 Influence of the Preprocessing on Plasma cfDNA Traits

For next-generation sequencing description of cfDNA, the apparent characteristics of the reported cfDNA depend on the preprocessing steps. Historically, cell-free DNA is assumed to be present as large at 40kb molecules but is now associated with the 167bp (or multiples of) length of DNA packed around nucleosomal structures (Holdenrieder et al., 2005). At the time, before next-generation sequencing methods were widely used, one study used quantitative PCR (Q-PCR) to examine fragment size demonstrating that not only is cfDNA highly fragmented (Diehl et al., 2005; Mouliere et al., 2011), but cell-free DNA fragment could potentially be as small as 45bp (Mouliere et al., 2013, 2014). In contrast to these Q-PCR-based reports, next-generation sequencing suggested consistently obtaining a fragment size profile of 166-167bp (Jiang et al., 2015; Sanchez et al., 2018; Snyder et al., 2016). There is no optimal single method for cfDNA fragment size analysis, and the accepted cfDNA size profile is constantly in flux. Mainly blood collection protocols (Wong et al., 2016), blood centrifugation (Rikkert et al., 2018), DNA extraction methodology (Jorgez et al., 2006; Lampignano et al., 2020; Markus et al., 2018; Pérez-Barrios et al., 2016; Sorber et al., 2017), and library preparation (Burnham et al., 2016; Sanchez et al., 2021; Snyder et al., 2016; van der Pol et al., 2022) all affect the cfDNA in the final presentation.

Traditionally, cfDNA analysis has been focused on double-stranded DNA (Jiang et al., 2015; Mouliere et al., 2018; Newman et al., 2016). The emergence of alternative protocols, in particular single-stranded library protocols from ancient DNA analysis for the recovery of single-stranded DNA within samples, can portray the encompassing DNA differently (Gansauge et al., 2017; Gansauge & Meyer, 2013; Meyer et al., 2012). In cell-free DNA, the use of single-stranded libraries has demonstrated an elevation in cfDNA molecules shorter than 100bp (Burnham et al., 2016; Sanchez et al., 2021; Snyder et al., 2016; Vong et al., 2017). In regard to improving ctDNA analysis, single-strand libraries have also presented mixed results on whether they increase the yield of ctDNA. Several studies report an increase (X. Liu et al., 2019; J. Zhu et al., 2020), while others reported no differences (Moser et al., 2017). For biological understanding, cfDNA fragments <100bp in length have been demonstrated to represent regulatory mechanisms (Esfahani et al., 2022; Snyder et al., 2016; Ulz et al., 2016, 2019). Exploring additional structural features of cfDNA through more inclusive technology will improve our understanding of cfDNA biology. This will then provide additional opportunities for diagnostic strategies and concepts.

1.7 Thesis Objective and Chapters Overview

The overarching goal of this thesis is to explore the hypothesis that there is ultrashort single-stranded cell-free DNA in plasma and document characteristics that may make it a viable new biomarker for NSCLC detection. As prior work by other investigators has hinted that single-stranded cell-free DNA libraries reveal a population of cell-free DNA below 70bp, there may be even shorter cell-free DNA that has previously been ignored (Burnham et al., 2016; Snyder et al., 2016).

To address this hypothesis, in Chapter 2, we designed a method titled "Broad Range Cell-free DNA Sequencing" (BRcfDNA-Seq), which combines low-molecular weight nucleic acid extraction and single-stranded library preparation. By processing non-cancer plasma through BRcfDNA-Seq, we report the presence of single-stranded cell-free DNA in plasma.

Chapter 3's intent is to bioinformatically analyze the sequenced data generated from BRcfDNA-Seq to compare the genomic characteristics between uscfDNA and mncfDNA. Here we observe that uscfDNA possesses the unique tendency to form functional element peaks and has enriched in potential G-Quad sequences compared to mncfDNA. As a proof of concept, we compare the uscfDNA and mncfDNA of plasma samples from non-cancer and NSCLC subjects to detect differences in functional element peaks, fragmentomics, end-motif profiles, and G-Quad abundance. We show that these features can be potential biomarkers in uscfDNA.

In Chapter 4, we investigate the methylation characteristics of uscfDNA compared to mncfDNA. Here we introduce the "5-mC Adapter Bisulfite Sequencing" (5mCAdpBS-SEq) method that circumvents genomic DNA degradation during bisulfite conversion from occluding signal from the uscfDNA. Using this method, we characterize the methylation pattern of uscfDNA compared to mncfDNA. Later, we explore if there is biomarker potential by looking at the differences in genomic patterns between NSCLC and non-cancer.

Finally, in Chapter 5, we describe the potential future directions that can be pursued regarding uscfDNA. Here, we dissect the technological, biological, and clinical aspects of uscfDNA, which can emerge with the advancement of further research.

1.8 References

Adams, D. H., Diaz, N., & Gahan, P. B. (1997). In vitro stimulation by tumour cell media of [3H]-thymidine incorporation by mouse spleen lymphocytes. Cell Biochemistry and Function, 15(2), 119–126. https://doi.org/10.1002/(SICI)1099-0844(19970601)15:2<119::AID-CBF731>3.0.CO;2-C

Adams, D. H., & McIntosh, A. A. (1985). Studies on the cytosolic DNA of chick embryo fibroblasts and its uptake by recipient cultured cells. The International Journal of Biochemistry, 17(10), 1041–1051. https://doi.org/10.1016/0020-711x(85)90035-7

Ancrile, B., Lim, K.-H., & Counter, C. M. (2007). Oncogenic Ras-induced secretion of IL6 is required for tumorigenesis. Genes & Development, 21(14), 1714–1719. https://doi.org/10.1101/gad.1549407

Balaj, L., Lessard, R., Dai, L., Cho, Y.-J., Pomeroy, S. L., Breakefield, X. O., & Skog, J. (2011). Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. Nature Communications, 2, 180. https://doi.org/10.1038/ncomms1180

Balkwill, F., & Coussens, L. M. (2004). Cancer: An inflammatory link. Nature, 431(7007), 405–406. https://doi.org/10.1038/431405a

Bao, H., Wang, Z., Ma, X., Guo, W., Zhang, X., Tang, W., Chen, X., Wang, X., Chen, Y., Mo, S., Liang, N., Ma, Q., Wu, S., Xu, X., Chang, S., Wei, Y., Zhang, X., Bao, H., Liu, R., … Shao, Y. (2022). Letter to the Editor: An ultra-sensitive assay using cell-free DNA fragmentomics for multi-cancer early detection. Molecular Cancer, 21(1), 129. https://doi.org/10.1186/s12943-022-01594-w

Barkauskas, C. E., Cronce, M. J., Rackley, C. R., Bowie, E. J., Keene, D. R., Stripp, B. R., Randell, S. H., Noble, P. W., & Hogan, B. L. M. (2013). Type 2 alveolar cells are stem cells in adult lung. The Journal of Clinical Investigation, 123(7), 3025–3036. https://doi.org/10.1172/JCI68782

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., Antonarakis, E. S., Azad, N. S., Bardelli, A., Brem, H., Cameron, J. L., Lee, C. C., Fecher, L. A., Gallia, G. L., Gibbs, P., … Diaz, L. A. (2014). Detection of circulating tumor DNA in early- and late-stage human malignancies. Science Translational Medicine, 6(224), 224ra24. https://doi.org/10.1126/scitranslmed.3007094

Boskovic, T., Stanic, J., Pena-Karan, S., Zarogoulidis, P., Drevelegas, K., Katsikogiannis, N., Machairiotis, N., Mpakas, A., Tsakiridis, K., Kesisis, G., Tsiouda, T., Kougioumtzi, I., Arikas, S., & Zarogoulidis, K. (2014). Pneumothorax after transthoracic needle biopsy of lung lesions under CT guidance. Journal of Thoracic Disease, 6(Suppl 1), S99–S107. https://doi.org/10.3978/j.issn.2072-1439.2013.12.08

Bryzgunova, O. E., Tamkovich, S. N., Cherepanova, A. V., Yarmoshchuk, S. V., Permyakova, V. I., Anykeeva, O. Y., & Laktionov, P. P. (2015). Redistribution of Free- and Cell-Surface-Bound DNA in Blood of Benign and Malignant Prostate Tumor Patients. Acta Naturae, 7(2), 115–118.

Burnham, P., Kim, M. S., Agbor-Enoh, S., Luikart, H., Valantine, H. A., Khush, K. K., & De Vlaminck, I. (2016). Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Scientific Reports, 6. https://doi.org/10.1038/srep27859

Chan, K. C. A., Jiang, P., Chan, C. W. M., Sun, K., Wong, J., Hui, E. P., Chan, S. L., Chan, W. C., Hui, D. S. C., Ng, S. S. M., Chan, H. L. Y., Wong, C. S. C., Ma, B. B. Y., Chan, A. T. C., Lai, P. B. S., Sun, H., Chiu, R. W. K., & Lo, Y. M. D. (2013). Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proceedings of the National Academy of Sciences of the United States of America, 110(47), 18761–18768. https://doi.org/10.1073/pnas.1313995110

Chan, K. C. A., Jiang, P., Sun, K., Cheng, Y. K. Y., Tong, Y. K., Cheng, S. H., Wong, A. I. C., Hudecova, I., Leung, T. Y., Chiu, R. W. K., & Lo, Y. M. D. (2016). Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proceedings of the National Academy of Sciences of the United States of America, 113(50), E8159–E8168. https://doi.org/10.1073/pnas.1615800113

Chelobanov, B. P., Laktionov, P. P., & Vlasov, V. V. (2006). Proteins involved in binding and cellular uptake of nucleic acids. Biochemistry. Biokhimiia, 71(6), 583–596. https://doi.org/10.1134/s0006297906060010

Cheng, J., Hu, Y., Wong, D. T. W., & Patel, A. A. (2023). CtDNA and Lung Cancer. In R. J. Cote & E. Lianidou (Eds.), Circulating Tumor Cells: Advances in Liquid Biopsy Technologies (pp. 511–537). Springer International Publishing. https://doi.org/10.1007/978-3-031-22903-9_20

Cheng, J., Morselli, M., Huang, W.-L., Heo, Y. J., Pinheiro-Ferreira, T., Li, F., Wei, F., Chia, D., Kim, Y., He, H.-J., Cole, K. D., Su, W.-C., Pellegrini, M., & Wong, D. T. W. (2022). Plasma contains ultrashort single-stranded DNA in addition to nucleosomal cell-free DNA. IScience, 25(7), 104554. https://doi.org/10.1016/j.isci.2022.104554

Chiu, R. W. K., Chan, L. Y. S., Lam, N. Y. L., Tsui, N. B. Y., Ng, E. K. O., Rainer, T. H., & Lo, Y. M. D. (2003). Quantitative analysis of circulating mitochondrial DNA in plasma. Clinical Chemistry, 49(5), 719–726. https://doi.org/10.1373/49.5.719

Cools-Lartigue, J., Spicer, J., McDonald, B., Gowing, S., Chow, S., Giannias, B., Bourdeau, F., Kubes, P., & Ferri, L. (2013). Neutrophil extracellular traps sequester circulating tumor cells and promote metastasis. The Journal of Clinical Investigation. https://doi.org/10.1172/JCI67484

Coussens, L. M., & Werb, Z. (2002). Inflammation and cancer. Nature, 420(6917), 860–867. https://doi.org/10.1038/nature01322

Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., Jensen, S. Ø., Medina, J. E., Hruban, C., White, J. R., Palsgrove, D. N., Niknafs, N., Anagnostou, V., Forde, P., Naidoo, J., Marrone, K., Brahmer, J., Woodward, B. D., Husain, H., … Velculescu, V. E. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature, 1. https://doi.org/10.1038/s41586-019-1272-6

Crusz, S. M., & Balkwill, F. R. (2015). Inflammation and cancer: Advances and new agents. Nature Reviews. Clinical Oncology, 12(10), 584–596. https://doi.org/10.1038/nrclinonc.2015.105

D'Alterio, C., Barbieri, A., Portella, L., Palma, G., Polimeno, M., Riccio, A., Ieranò, C., Franco, R., Scognamiglio, G., Bryce, J., Luciano, A., Rea, D., Arra, C., & Scala, S. (2012). Inhibition of stromal CXCR4 impairs development of lung metastases. Cancer Immunology, Immunotherapy, 61(10), 1713–1720. https://doi.org/10.1007/s00262-012-1223-7

de Bont, C. M., Boelens, W. C., & Pruijn, G. J. M. (2019). NETosis, complement, and coagulation: A triangular relationship. Cellular & Molecular Immunology, 16(1), Article 1. https://doi.org/10.1038/s41423-018-0024-0

Diaz, L. A., & Bardelli, A. (2014). Liquid Biopsies: Genotyping Circulating Tumor DNA. Journal of Clinical Oncology, 32(6), 579–586. https://doi.org/10.1200/JCO.2012.45.2011

Diehl, F., Li, M., Dressman, D., He, Y., Shen, D., Szabo, S., Diaz, L. A., Goodman, S. N., David, K. A., Juhl, H., Kinzler, K. W., & Vogelstein, B. (2005). Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proceedings of the National Academy of Sciences, 102(45), 16368–16373. https://doi.org/10.1073/pnas.0507904102

Esfahani, M. S., Hamilton, E. G., Mehrmohamadi, M., Nabet, B. Y., Alig, S. K., King, D. A., Steen, C. B., Macaulay, C. W., Schultz, A., Nesselbush, M. C., Soo, J., Schroers-Martin, J. G., Chen, B., Binkley, M. S., Stehr, H., Chabon, J. J., Sworder, B. J., Hui, A. B.-Y., Frank, M. J., … Alizadeh, A. A. (2022). Inferring gene expression from cell-free DNA fragmentation profiles. Nature Biotechnology, 40(4), 585–597. https://doi.org/10.1038/s41587-022-01222-4

Esteller, M. (2008). Epigenetics in cancer. The New England Journal of Medicine, 358(11), 1148–1159. https://doi.org/10.1056/NEJMra072067

Ettinger, D. S., Wood, D. E., Aisner, D. L., Akerley, W., Bauman, J., Chirieac, L. R., D'Amico, T. A., DeCamp, M. M., Dilling, T. J., Dobelbower, M., Doebele, R. C., Govindan, R., Gubens, M. A., Hennon, M., Horn, L., Komaki, R., Lackner, R. P., Lanuti, M., Leal, T. A., … Hughes, M. (2017). Non–Small Cell Lung Cancer, Version 5.2017, NCCN Clinical Practice Guidelines in Oncology. Journal of the National Comprehensive Cancer Network, 15(4), 504–535. https://doi.org/10.6004/jnccn.2017.0050

Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L., & Quake, S. R. (2008). Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proceedings of the National Academy of Sciences, 105(42), 16266–16271. https://doi.org/10.1073/pnas.0808319105

Fernando, M. R., Jiang, C., Krzyzanowski, G. D., & Ryan, W. L. (2017). New evidence that a large proportion of human blood plasma cell-free DNA is localized in exosomes. PLOS ONE, 12(8), e0183915. https://doi.org/10.1371/journal.pone.0183915

Foda, Z. H., Annapragada, A. V., Boyapati, K., Bruhm, D. C., Vulpescu, N. A., Medina, J. E., Mathios, D., Cristiano, S., Niknafs, N., Luu, H. T., Goggins, M. G., Anders, R. A., Sun, J., Mehta, S. H., Thomas, D. L., Kirk, G. D., Adleff, V., Phallen, J., Scharpf, R. B., … Velculescu, V. E. (2022). Detecting liver cancer using cell-free DNA fragmentomes. Cancer Discovery, CD-22-0659. https://doi.org/10.1158/2159-8290.CD-22-0659

Fontana, R. S., Sanderson, D. R., Woolner, L. B., Miller, W. E., Bernatz, P. E., Payne, W. S., & Taylor, W. F. (1975). The Mayo Lung Project for early detection and localization of bronchogenic carcinoma: A status report. Chest, 67(5), 511–522. https://doi.org/10.1378/chest.67.5.511

Frost, J. K., Ball, W. C., Levin, M. L., Tockman, M. S., Baker, R. R., Carter, D., Eggleston, J. C., Erozan, Y. S., Gupta, P. K., & Khouri, N. F. (1984). Early lung cancer detection: Results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study. The American Review of Respiratory Disease, 130(4), 549–554. https://doi.org/10.1164/arrd.1984.130.4.549

Fuchs, T. A., Brill, A., Duerschmied, D., Schatzberg, D., Monestier, M., Myers, D. D., Wrobleski, S. K., Wakefield, T. W., Hartwig, J. H., & Wagner, D. D. (2010). Extracellular DNA traps promote thrombosis. Proceedings of the National Academy of Sciences of the United States of America, 107(36), 15880–15885. https://doi.org/10.1073/pnas.1005743107

Gahan, P. B., & Stroun, M. (2010). The virtosome-a novel cytosolic informative entity and intercellular messenger. Cell Biochemistry and Function, 28(7), 529–538. https://doi.org/10.1002/cbf.1690

Gansauge, M.-T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., Riehl, L. M., Schmidt, A., & Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. Nucleic Acids Research, 45(10), e79. https://doi.org/10.1093/nar/gkx033

Gansauge, M.-T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nature Protocols, 8(4), 737–748. https://doi.org/10.1038/nprot.2013.038

Giangreco, A., Lu, L., Vickers, C., Teixeira, V. H., Groot, K. R., Butler, C. R., Ilieva, E. V., George, P. J., Nicholson, A. G., Sage, E. K., Watt, F. M., & Janes, S. M. (2012). β-Catenin determines upper airway progenitor cell fate and preinvasive squamous lung cancer progression by modulating epithelial-mesenchymal transition. The Journal of Pathology, 226(4), 575–587. https://doi.org/10.1002/path.3962

Guescini, M., Genedani, S., Stocchi, V., & Agnati, L. F. (2010). Astrocytes and Glioblastoma cells release exosomes carrying mtDNA. Journal of Neural Transmission (Vienna, Austria: 1996), 117(1), 1–4. https://doi.org/10.1007/s00702-009-0288-8

Guo, W., Chen, X., Liu, R., Liang, N., Ma, Q., Bao, H., Xu, X., Wu, X., Yang, S., Shao, Y., Tan, F., Xue, Q., Gao, S., & He, J. (2022). Sensitive detection of stage I lung adenocarcinoma using plasma cell-free DNA breakpoint motif profiling. EBioMedicine, 81. https://doi.org/10.1016/j.ebiom.2022.104131

Hanna, J. M., & Onaitis, M. W. (2013). Cell of origin of lung cancer. Journal of Carcinogenesis, 12, 6. https://doi.org/10.4103/1477-3163.109033

Harris, K., Khachaturova, I., Azab, B., Maniatis, T., Murukutla, S., Chalhoub, M., Hatoum, H., Kilkenny, T., Elsayegh, D., Maroun, R., & Alkaied, H. (2012). Small Cell Lung Cancer Doubling Time and its Effect on Clinical Presentation: A Concise Review. Clinical Medicine Insights. Oncology, 6, 199–203. https://doi.org/10.4137/CMO.S9633

Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature, 553(7689), 446–454. https://doi.org/10.1038/nature25183

Hill, W., Lim, E. L., Weeden, C. E., Lee, C., Augustine, M., Chen, K., Kuan, F.-C., Marongiu, F., Evans, E. J., Moore, D. A., Rodrigues, F. S., Pich, O., Bakker, B., Cha, H., Myers, R., van Maldegem, F., Boumelha, J., Veeriah, S., Rowan, A., … Swanton, C. (2023). Lung adenocarcinoma promotion by air pollutants. Nature, 616(7955), 159–167. https://doi.org/10.1038/s41586-023-05874-3

Hogan, B. L. M., Barkauskas, C. E., Chapman, H. A., Epstein, J. A., Jain, R., Hsia, C. C. W., Niklason, L., Calle, E., Le, A., Randell, S. H., Rock, J., Snitow, M., Krummel, M., Stripp, B. R., Vu, T., White, E. S., Whitsett, J. A., & Morrisey, E. E. (2014). Repair and regeneration of the respiratory system: Complexity, plasticity, and mechanisms of lung stem cell function. Cell Stem Cell, 15(2), 123–138. https://doi.org/10.1016/j.stem.2014.07.012

Holdenrieder, S., Mueller, S., & Stieber, P. (2005). Stability of nucleosomal DNA fragments in serum. Clinical Chemistry, 51(6), 1026–1029. https://doi.org/10.1373/clinchem.2005.048454

Hong, K. U., Reynolds, S. D., Watkins, S., Fuchs, E., & Stripp, B. R. (2004). Basal Cells Are a Multipotent Progenitor Capable of Renewing the Bronchial Epithelium. The American Journal of Pathology, 164(2), 577–588.

Huang, X., Zhao, Q., An, X., Pan, J., Zhao, L., Shen, L., Xu, Y., & Yuan, D. (2020). The Ratio of ssDNA to dsDNA in Circulating Cell-Free DNA Extract is a Stable Indicator for Diagnosis of Gastric Cancer. Pathology & Oncology Research, 26(4), 2621–2632. https://doi.org/10.1007/s12253-020-00869-1

Hummel, E. M., Hessas, E., Müller, S., Beiter, T., Fisch, M., Eibl, A., Wolf, O. T., Giebel, B., Platen, P., Kumsta, R., & Moser, D. A. (2018). Cell-free DNA release under psychosocial and physical stress conditions. Translational Psychiatry, 8(1), Article 1. https://doi.org/10.1038/s41398-018-0264-x

Ignatiadis, M., Sledge, G. W., & Jeffrey, S. S. (2021). Liquid biopsy enters the clinic—Implementation issues and future challenges. Nature Reviews Clinical Oncology, 18(5), Article 5. https://doi.org/10.1038/s41571-020-00457-x

Infante, M., Cavuto, S., Lutman, F. R., Brambilla, G., Chiesa, G., Ceresoli, G., Passera, E., Angeli, E., Chiarenza, M., Aranzulla, G., Cariboni, U., Errico, V., Inzirillo, F., Bottoni, E., Voulaz, E., Alloisio, M., Destro, A., Roncalli, M., Santoro, A., … DANTE Study Group. (2009). A randomized study of lung cancer screening with spiral computed tomography: Three-year results from the DANTE trial. American Journal of Respiratory and Critical Care Medicine, 180(5), 445–453. https://doi.org/10.1164/rccm.200901-0076OC

Jahr, S., Hentze, H., Englisch, S., Hardt, D., Fackelmayer, F. O., Hesch, R. D., & Knippers, R. (2001). DNA fragments in the blood plasma of cancer patients: Quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Research, 61(4), 1659–1665.

Jiang, P., Chan, C. W. M., Chan, K. C. A., Cheng, S. H., Wong, J., Wong, V. W.-S., Wong, G. L. H., Chan, S. L., Mok, T. S. K., Chan, H. L. Y., Lai, P. B. S., Chiu, R. W. K., & Lo, Y. M. D. (2015). Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proceedings of the National Academy of Sciences, 112(11), E1317–E1325. https://doi.org/10.1073/pnas.1500076112

Jiang, P., Sun, K., Peng, W., Cheng, S. H., Ni, M., Yeung, P. C., Heung, M. M. S., Xie, T., Shang, H., Zhou, Z., Chan, R. W. Y., Wong, J., Wong, V. W. S., Poon, L. C., Leung, T. Y., Lam, W. K. J., Chan, J. Y. K., Chan, H. L. Y., Chan, K. C. A., … Lo, Y. M. D. (2020). Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discovery, 10(5), 664–673. https://doi.org/10.1158/2159-8290.CD-19-0622

Jiang, P., Xie, T., Ding, S. C., Zhou, Z., Cheng, S. H., Chan, R. W. Y., Lee, W.-S., Peng, W., Wong, J., Wong, V. W. S., Chan, H. L. Y., Chan, S. L., Poon, L. C. Y., Leung, T. Y., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2020). Detection and characterization of jagged ends of double-stranded DNA in plasma. Genome Research, 30(8), 1144–1153. https://doi.org/10.1101/gr.261396.120

Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. Nature Reviews. Genetics, 13(7), 484–492. https://doi.org/10.1038/nrg3230

Jorgez, C. J., Dang, D. D., Simpson, J. L., Lewis, D. E., & Bischoff, F. Z. (2006). Quantity versus quality: Optimal methods for cell-free DNA isolation from plasma of pregnant women. Genetics in Medicine, 8(10), Article 10. https://doi.org/10.1097/01.gim.0000241904.32039.6f

Kahlert, C., Melo, S. A., Protopopov, A., Tang, J., Seth, S., Koch, M., Zhang, J., Weitz, J., Chin, L., Futreal, A., & Kalluri, R. (2014). Identification of double-stranded genomic DNA spanning all chromosomes with mutated KRAS and p53 DNA in the serum exosomes of patients with pancreatic cancer. The Journal of Biological Chemistry, 289(7), 3869–3875. https://doi.org/10.1074/jbc.C113.532267

Kakarla, R., Hur, J., Kim, Y. J., Kim, J., & Chwae, Y.-J. (2020). Apoptotic cell-derived exosomes: Messages from dying cells. Experimental & Molecular Medicine, 52(1), Article 1. https://doi.org/10.1038/s12276-019-0362-8

Kaplan, M. J., & Radic, M. (2012). Neutrophil extracellular traps: Double-edged swords of innate immunity. Journal of Immunology (Baltimore, Md.: 1950), 189(6), 2689–2695. https://doi.org/10.4049/jimmunol.1201719

Karachaliou, N., Pilotto, S., Lazzari, C., Bria, E., de Marinis, F., & Rosell, R. (2016). Cellular and molecular biology of small cell lung cancer: An overview. Translational Lung Cancer Research, 5(1), 2–15. https://doi.org/10.3978/j.issn.2218-6751.2016.01.02

Kohler, C., Radpour, R., Barekati, Z., Asadollahi, R., Bitzer, J., Wight, E., Bürki, N., Diesch, C., Holzgreve, W., & Zhong, X. Y. (2009). Levels of plasma circulating cell free nuclear and mitochondrial DNA as potential biomarkers for breast tumors. Molecular Cancer, 8, 105. https://doi.org/10.1186/1476-4598-8-105

Kris, M. G., Johnson, B. E., Berry, L. D., Kwiatkowski, D. J., Iafrate, A. J., Wistuba, I. I., Varella-Garcia, M., Franklin, W. A., Aronson, S. L., Su, P.-F., Shyr, Y., Camidge, D. R., Sequist, L. V., Glisson, B. S., Khuri, F. R., Garon, E. B., Pao, W., Rudin, C., Schiller, J., … Bunn, P. A. (2014). Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA, 311(19), 1998–2006. https://doi.org/10.1001/jama.2014.3741

Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S. P., Arlt, V. M., Phillips, D. H., & Nik-Zainal, S. (2019). A Compendium of Mutational Signatures of Environmental Agents. Cell, 177(4), 821-836.e16. https://doi.org/10.1016/j.cell.2019.03.001

Kumar, P., Dillon, L. W., Shibata, Y., Jazaeri, A. A., Jones, D. R., & Dutta, A. (2017). Normal and Cancerous Tissues Release Extrachromosomal Circular DNA (eccDNA) into the Circulation. Molecular Cancer Research: MCR, 15(9), 1197–1205. https://doi.org/10.1158/1541-7786.MCR-17-0095

Kundu, J. K., & Surh, Y.-J. (2008). Inflammation: Gearing the journey to cancer. Mutation Research/Reviews in Mutation Research, 659(1), 15–30. https://doi.org/10.1016/j.mrrev.2008.03.002

Lampignano, R., Neumann, M. H. D., Weber, S., Kloten, V., Herdean, A., Voss, T., Groelz, D., Babayan, A., Tibbesma, M., Schlumpberger, M., Chemi, F., Rothwell, D. G., Wikman, H., Galizzi, J.-P., Riise Bergheim, I., Russnes, H., Mussolin, B., Bonin, S., Voigt, C., … Heitzer, E. (2020). Multicenter Evaluation of Circulating Cell-Free DNA Extraction and Downstream Analyses for the Development of Standardized (Pre)analytical Work Flows. Clinical Chemistry, 66(1), 149–160. https://doi.org/10.1373/clinchem.2019.306837

Larke, F. J., Kruger, R. L., Cagnon, C. H., Flynn, M. J., McNitt-Gray, M. M., Wu, X., Judy, P. F., & Cody, D. D. (2011). Estimated radiation dose associated with low-dose chest CT of average-size participants in the National Lung Screening Trial. AJR. American Journal of Roentgenology, 197(5), 1165–1169. https://doi.org/10.2214/AJR.11.6533

Leon, S. A., Shapiro, B., Sklaroff, D. M., & Yaros, M. J. (1977). Free DNA in the serum of cancer patients and the effect of therapy. Cancer Research, 37(3), 646–650.

Liu, M. C., Oxnard, G. R., Klein, E. A., Swanton, C., Seiden, M. V., Liu, M. C., Oxnard, G. R., Klein, E. A., Smith, D., Richards, D., Yeatman, T. J., Cohn, A. L., Lapham, R., Clement, J., Parker, A. S., Tummala, M. K., McIntyre, K., Sekeres, M. A., Bryce, A. H., … Berry, D. A. (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Annals of Oncology, 31(6), 745–759. https://doi.org/10.1016/j.annonc.2020.02.011

Liu, X., Liu, L., Ji, Y., Li, C., Wei, T., Yang, X., Zhang, Y., Cai, X., Gao, Y., Xu, W., Rao, S., Jin, D., Lou, W., Qiu, Z., & Wang, X. (2019). Enrichment of short mutant cell-free DNA fragments enhanced detection of pancreatic cancer. EBioMedicine, 41, 345–356. https://doi.org/10.1016/j.ebiom.2019.02.010

Lo, Y. M., Corbetta, N., Chamberlain, P. F., Rai, V., Sargent, I. L., Redman, C. W., & Wainscoat, J. S. (1997). Presence of fetal DNA in maternal plasma and serum. Lancet (London, England), 350(9076), 485–487. https://doi.org/10.1016/S0140-6736(97)02174-0

Lu, Y., Futtner, C., Rock, J. R., Xu, X., Whitworth, W., Hogan, B. L. M., & Onaitis, M. W. (2010). Evidence That SOX2 Overexpression Is Oncogenic in the Lung. PLOS ONE, 5(6), e11022. https://doi.org/10.1371/journal.pone.0011022

Lui, Y. Y. N., Chik, K.-W., Chiu, R. W. K., Ho, C.-Y., Lam, C. W. K., & Lo, Y. M. D. (2002). Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. Clinical Chemistry, 48(3), 421–427.

Luo, L., Zhang, S., Wang, Y., Rahman, M., Syk, I., Zhang, E., & Thorlacius, H. (2014). Proinflammatory role of neutrophil extracellular traps in abdominal sepsis. American Journal of Physiology. Lung Cellular and Molecular Physiology, 307(7), L586-596. https://doi.org/10.1152/ajplung.00365.2013

Mandel,P, & Metais, P. (1948). Les acides nucléiques du plasma sanguin chez l'Homme. Comptes Rendus Des Seances de La Societe de Biologie et de Ses Filiales, 142, 241–243.

Mantovani, A., Allavena, P., Sica, A., & Balkwill, F. (2008). Cancer-related inflammation. Nature, 454(7203), 436–444. https://doi.org/10.1038/nature07205

Marass, F., Stephens, D., Ptashkin, R., Zehir, A., Berger, M. F., Solit, D. B., Diaz, L. A., & Tsui, D. W. Y. (2020). Fragment Size Analysis May Distinguish Clonal Hematopoiesis from Tumor-Derived Mutations in Cell-Free DNA. Clinical Chemistry, 66(4), 616–618. https://doi.org/10.1093/clinchem/hvaa026

Markus, H., Contente-Cuomo, T., Farooq, M., Liang, W. S., Borad, M. J., Sivakumar, S., Gollins, S., Tran, N. L., Dhruv, H. D., Berens, M. E., Bryce, A., Sekulic, A., Ribas, A., Trent, J. M., LoRusso, P. M., & Murtaza, M. (2018). Evaluation of pre-analytical factors affecting plasma DNA analysis. Scientific Reports, 8(1), 7375. https://doi.org/10.1038/s41598-018-25810-0

Melamed, M. R., Flehinger, B. J., Zaman, M. B., Heelan, R. T., Perchick, W. A., & Martini, N. (1984). Screening for early lung cancer. Results of the Memorial Sloan-Kettering study in New York. Chest, 86(1), 44–53. https://doi.org/10.1378/chest.86.1.44

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science (New York, N.Y.), 338(6104), 222–226. https://doi.org/10.1126/science.1224344

Morozkin, E. S., Laktionov, P. P., Rykova, E. Y., & Vlassov, V. V. (2004). Extracellular nucleic acids in cultures of long-term cultivated eukaryotic cells. Annals of the New York Academy of Sciences, 1022, 244–249. https://doi.org/10.1196/annals.1318.038

Moser, T., Ulz, P., Zhou, Q., Perakis, S., Geigl, J. B., Speicher, M. R., & Heitzer, E. (2017). Single-Stranded DNA Library Preparation Does Not Preferentially Enrich Circulating Tumor DNA. Clinical Chemistry, 63(10), 1656–1659. https://doi.org/10.1373/clinchem.2017.277988

Mouliere, F., Chandrananda, D., Piskorz, A. M., Moore, E. K., Morris, J., Ahlborn, L. B., Mair, R., Goranova, T., Marass, F., Heider, K., Wan, J. C. M., Supernat, A., Hudecova, I., Gounaris, I., Ros, S., Jimenez-Linan, M., Garcia-Corbacho, J., Patel, K., Østrup, O., … Rosenfeld, N. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. Science Translational Medicine, 10(466). https://doi.org/10.1126/scitranslmed.aat4921

Mouliere, F., El Messaoudi, S., Gongora, C., Guedj, A.-S., Robert, B., Del Rio, M., Molina, F., Lamy, P.-J., Lopez-Crapez, E., Mathonnet, M., Ychou, M., Pezet, D., & Thierry, A. R. (2013). Circulating Cell-Free DNA from Colorectal Cancer Patients May Reveal High KRAS or BRAF Mutation Load. Translational Oncology, 6(3), 319–328.

Mouliere, F., El Messaoudi, S., Pang, D., Dritschilo, A., & Thierry, A. R. (2014). Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. Molecular Oncology, 8(5), 927–941. https://doi.org/10.1016/j.molonc.2014.02.005

Mouliere, F., Robert, B., Arnau Peyrotte, E., Del Rio, M., Ychou, M., Molina, F., Gongora, C., & Thierry, A. R. (2011). High fragmentation characterizes tumour-derived circulating DNA. PloS One, 6(9), e23418. https://doi.org/10.1371/journal.pone.0023418

Naidich, D. P., Marshall, C. H., Gribbin, C., Arams, R. S., & McCauley, D. I. (1990). Low-dose CT of the lungs: Preliminary observations. Radiology, 175(3), 729–731. https://doi.org/10.1148/radiology.175.3.2343122

National Lung Screening Trial Research Team, Aberle, D. R., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., Fagerstrom, R. M., Gareen, I. F., Gatsonis, C., Marcus, P. M., & Sicks, J. D. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England Journal of Medicine, 365(5), 395–409. https://doi.org/10.1056/NEJMoa1102873

Nesbitt, J. C., Putnam, J. B., Walsh, G. L., Roth, J. A., & Mountain, C. F. (1995). Survival in early-stage non-small cell lung cancer. The Annals of Thoracic Surgery, 60(2), 466–472. https://doi.org/10.1016/0003-4975(95)00169-l

Newman, A. M., Lovejoy, A. F., Klass, D. M., Kurtz, D. M., Chabon, J. J., Scherer, F., Stehr, H., Liu, C. L., Bratman, S. V., Say, C., Zhou, L., Carter, J. N., West, R. B., Sledge, G. W., Shrager, J.

B., Loo, B. W., Neal, J. W., Wakelee, H. A., Diehn, M., & Alizadeh, A. A. (2016). Integrated digital error suppression for improved detection of circulating tumor DNA. Nature Biotechnology, 34(5), 547–555. https://doi.org/10.1038/nbt.3520

Oken, M. M., Hocking, W. G., Kvale, P. A., Andriole, G. L., Buys, S. S., Church, T. R., Crawford, E. D., Fouad, M. N., Isaacs, C., Reding, D. J., Weissfeld, J. L., Yokochi, L. A., O'Brien, B., Ragard, L. R., Rathmell, J. M., Riley, T. L., Wright, P., Caparaso, N., Hu, P., … PLCO Project Team. (2011). Screening by chest radiograph and lung cancer mortality: The Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. JAMA, 306(17), 1865–1873. https://doi.org/10.1001/jama.2011.1591

Parikh, A. R., Van Seventer, E. E., Siravegna, G., Hartwig, A. V., Jaimovich, A., He, Y., Kanter, K., Fish, M. G., Fosbenner, K. D., Miao, B., Phillips, S., Carmichael, J. H., Sharma, N., Jarnagin, J., Baiev, I., Shah, Y. S., Fetter, I. J., Shahzade, H. A., Allen, J. N., … Corcoran, R. B. (2021). Minimal Residual Disease Detection using a Plasma-only Circulating Tumor DNA Assay in Patients with Colorectal Cancer. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 27(20), 5586–5594. https://doi.org/10.1158/1078-0432.CCR-21-0410

Pastorino, U., Rossi, M., Rosato, V., Marchianò, A., Sverzellati, N., Morosi, C., Fabbri, A., Galeone, C., Negri, E., Sozzi, G., Pelosi, G., & La Vecchia, C. (2012). Annual or biennial CT screening versus observation in heavy smokers: 5-year results of the MILD trial. European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation (ECP), 21(3), 308–315. https://doi.org/10.1097/CEJ.0b013e328351e1b6

Pérez-Barrios, C., Nieto-Alcolado, I., Torrente, M., Jiménez-Sánchez, C., Calvo, V., Gutierrez-Sanz, L., Palka, M., Donoso-Navarro, E., Provencio, M., & Romero, A. (2016). Comparison of methods for circulating cell-free DNA isolation using blood from cancer patients: Impact on biomarker testing. Translational Lung Cancer Research, 5(6), 665–672. https://doi.org/10.21037/tlcr.2016.12.03

Rainer, T. H., Wong, L. K. S., Lam, W., Yuen, E., Lam, N. Y. L., Metreweli, C., & Lo, Y. M. D. (2003). Prognostic use of circulating plasma nucleic acid concentrations in patients with acute stroke. Clinical Chemistry, 49(4), 562–569. https://doi.org/10.1373/49.4.562

Razavi, P., Li, B. T., Brown, D. N., Jung, B., Hubbell, E., Shen, R., Abida, W., Juluru, K., De Bruijn, I., Hou, C., Venn, O., Lim, R., Anand, A., Maddala, T., Gnerre, S., Vijaya Satya, R., Liu, Q., Shen, L., Eattock, N., … Reis-Filho, J. S. (2019). High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. Nature Medicine, 25(12), Article 12. https://doi.org/10.1038/s41591-019-0652-7

Rikkert, L. G., van der Pol, E., van Leeuwen, T. G., Nieuwland, R., & Coumans, F. A. W. (2018). Centrifugation affects the purity of liquid biopsy-based tumor biomarkers. Cytometry.

Part A: The Journal of the International Society for Analytical Cytology, 93(12), 1207–1212. https://doi.org/10.1002/cyto.a.23641

Rivera, M. P., Mehta, A. C., & Wahidi, M. M. (2013). Establishing the diagnosis of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest, 143(5 Suppl), e142S-e165S. https://doi.org/10.1378/chest.12-2353

Rolfo, C., Mack, P. C., Scagliotti, G. V., Baas, P., Barlesi, F., Bivona, T. G., Herbst, R. S., Mok, T. S., Peled, N., Pirker, R., Raez, L. E., Reck, M., Riess, J. W., Sequist, L. V., Shepherd, F. A., Sholl, L. M., Tan, D. S. W., Wakelee, H. A., Wistuba, I. I., … Gandara, D. R. (2018). Liquid Biopsy for Advanced Non-Small Cell Lung Cancer (NSCLC): A Statement Paper from the IASLC. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer, 13(9), 1248–1268. https://doi.org/10.1016/j.jtho.2018.05.030

Rosell, R., Carcereny, E., Gervais, R., Vergnenegre, A., Massuti, B., Felip, E., Palmero, R., Garcia-Gomez, R., Pallares, C., Sanchez, J. M., Porta, R., Cobo, M., Garrido, P., Longo, F., Moran, T., Insa, A., De Marinis, F., Corre, R., Bover, I., … Spanish Lung Cancer Group in collaboration with Groupe Français de Pneumo-Cancérologie and Associazione Italiana Oncologia Toracica. (2012). Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): A multicentre, open-label, randomised phase 3 trial. The Lancet. Oncology, 13(3), 239–246. https://doi.org/10.1016/S1470-2045(11)70393-X

Rostami, A., Lambie, M., Yu, C. W., Stambolic, V., Waldron, J. N., & Bratman, S. V. (2020). Senescence, Necrosis, and Apoptosis Govern Circulating Cell-free DNA Release Kinetics. Cell Reports, 31(13), 107830. https://doi.org/10.1016/j.celrep.2020.107830

Rykova, E. Y., Morozkin, E. S., Ponomaryova, A. A., Loseva, E. M., Zaporozhchenko, I. A., Cherdyntseva, N. V., Vlassov, V. V., & Laktionov, P. P. (2012). Cell-free and cell-bound circulating nucleic acid complexes: Mechanisms of generation, concentration and content. Expert Opinion on Biological Therapy, 12 Suppl 1, S141-153. https://doi.org/10.1517/14712598.2012.673577

Rykova EYu, null, Pautova, L. V., Yakubov, L. A., Karamyshev, V. N., & Vlassov, V. V. (1994). Serum immunoglobulins interact with oligonucleotides. FEBS Letters, 344(1), 96–98. https://doi.org/10.1016/0014-5793(94)00360-2

Saghir, Z., Dirksen, A., Ashraf, H., Bach, K. S., Brodersen, J., Clementsen, P. F., Døssing, M., Hansen, H., Kofoed, K. F., Larsen, K. R., Mortensen, J., Rasmussen, J. F., Seersholm, N., Skov, B. G., Thorsen, H., Tønnesen, P., & Pedersen, J. H. (2012). CT screening for lung cancer brings forward early disease. The randomised Danish Lung Cancer Screening Trial: Status after five annual screening rounds with low-dose CT. Thorax, 67(4), 296–301. https://doi.org/10.1136/thoraxjnl-2011-200736

Sanchez, C., Roch, B., Mazard, T., Blache, P., Dache, Z. A. A., Pastor, B., Pisareva, E., Tanos, R., & Thierry, A. R. (2021). Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics. JCI Insight, 6(7), 144561. https://doi.org/10.1172/jci.insight.144561

Sanchez, C., Snyder, M. W., Tanos, R., Shendure, J., & Thierry, A. R. (2018). New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. NPJ Genomic Medicine, 3, 31. https://doi.org/10.1038/s41525-018-0069-0

Sanmamed, M. F., Carranza-Rua, O., Alfaro, C., Oñate, C., Martín-Algarra, S., Perez, G., Landazuri, S. F., Gonzalez, A., Gross, S., Rodriguez, I., Muñoz-Calleja, C., Rodríguez-Ruiz, M., Sangro, B., López-Picazo, J. M., Rizzo, M., Mazzolini, G., Pascual, J. I., Andueza, M. P., Perez-Gracia, J. L., & Melero, I. (2014). Serum interleukin-8 reflects tumor burden and treatment response across malignancies of multiple tissue origins. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 20(22), 5697–5707. https://doi.org/10.1158/1078-0432.CCR-13-3203

Serpas, L., Chan, R. W. Y., Jiang, P., Ni, M., Sun, K., Rashidfarrokhi, A., Soni, C., Sisirak, V., Lee, W.-S., Cheng, S. H., Peng, W., Chan, K. C. A., Chiu, R. W. K., Reizis, B., & Lo, Y. M. D. (2019). Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. Proceedings of the National Academy of Sciences, 116(2), 641–649. https://doi.org/10.1073/pnas.1815031116

Shah, R., Sabanathan, S., Richardson, J., Mearns, A. J., & Goulden, C. (1996). Results of surgical treatment of stage I and II lung cancer. The Journal of Cardiovascular Surgery, 37(2), 169–172.

Shen, S. Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M. H. A., Chadwick, D., Zuzarte, P. C., Borgida, A., Wang, T. T., Li, T., Kis, O., Zhao, Z., Spreafico, A., Medina, T. da S., Wang, Y., Roulois, D., Ettayebi, I., Chen, Z., Chow, S., … De Carvalho, D. D. (2018). Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature, 563(7732), 579–583. https://doi.org/10.1038/s41586-018-0703-0

Simmons, C. P., Koinis, F., Fallon, M. T., Fearon, K. C., Bowden, J., Solheim, T. S., Gronberg, B. H., McMillan, D. C., Gioulbasanis, I., & Laird, B. J. (2015). Prognosis in advanced lung cancer—A prospective study examining key clinicopathological factors. Lung Cancer (Amsterdam, Netherlands), 88(3), 304–309. https://doi.org/10.1016/j.lungcan.2015.03.020

Sin, S. T. K., Deng, J., Ji, L., Yukawa, M., Chan, R. W. Y., Volpi, S., Vaglio, A., Fenaroli, P., Bocca, P., Cheng, S. H., Wong, D. K. L., Lui, K. O., Jiang, P., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (n.d.). Effects of nucleases on cell-free extrachromosomal circular DNA. JCI Insight, 7(8), e156070. https://doi.org/10.1172/jci.insight.156070

Singer, C. F., Kronsteiner, N., Hudelist, G., Marton, E., Walter, I., Kubista, M., Czerwenka, K., Schreiber, M., Seifert, M., & Kubista, E. (2003). Interleukin 1 system and sex steroid receptor expression in human breast cancer: Interleukin 1alpha protein secretion is correlated with malignant phenotype. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 9(13), 4877–4883.

Smith, Z. D., & Meissner, A. (2013). DNA methylation: Roles in mammalian development. Nature Reviews. Genetics, 14(3), 204–220. https://doi.org/10.1038/nrg3354

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. (2016). Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell, 164(0), 57–68. https://doi.org/10.1016/j.cell.2015.11.050

Song, H. S., Kang, D. H., Kim, H., Ahn, T. S., Kim, T. W., & Baek, M.-J. (2021). Clinical relevance and prognostic role of preoperative cell-free single-stranded DNA concentrations in colorectal cancer patients. Korean Journal of Clinical Oncology, 17(2), 59–67. https://doi.org/10.14216/kjco.21010

Sorber, L., Zwaenepoel, K., Deschoolmeester, V., Roeyen, G., Lardon, F., Rolfo, C., & Pauwels, P. (2017). A Comparison of Cell-Free DNA Isolation Kits: Isolation and Quantification of Cell-Free DNA in Plasma. The Journal of Molecular Diagnostics: JMD, 19(1), 162–168. https://doi.org/10.1016/j.jmoldx.2016.09.009

Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y.-M., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. E., & Brody, J. S. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nature Medicine, 13(3), 361–366. https://doi.org/10.1038/nm1556

Stroun, M., Anker, P., Maurice, P., Lyautey, J., Lederrey, C., & Beljanski, M. (1989). Neoplastic characteristics of the DNA found in the plasma of cancer patients. Oncology, 46(5), 318–322. https://doi.org/10.1159/000226740

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249. https://doi.org/10.3322/caac.21660

Sutherland, K. D., Proost, N., Brouns, I., Adriaensen, D., Song, J.-Y., & Berns, A. (2011). Cell of origin of small cell lung cancer: Inactivation of Trp53 and Rb1 in distinct cell types of adult mouse lung. Cancer Cell, 19(6), 754–764. https://doi.org/10.1016/j.ccr.2011.04.019

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: Provocative insights from epigenomics. Nature Reviews. Genetics, 9(6), 465–476. https://doi.org/10.1038/nrg2341

Tetta, C., Ghigo, E., Silengo, L., Deregibus, M. C., & Camussi, G. (2013). Extracellular vesicles as an emerging mechanism of cell-to-cell communication. Endocrine, 44(1), 11–19. https://doi.org/10.1007/s12020-012-9839-0

Thakur, B. K., Zhang, H., Becker, A., Matei, I., Huang, Y., Costa-Silva, B., Zheng, Y., Hoshino, A., Brazier, H., Xiang, J., Williams, C., Rodriguez-Barrueco, R., Silva, J. M., Zhang, W., Hearn, S., Elemento, O., Paknejad, N., Manova-Todorova, K., Welte, K., … Lyden, D. (2014). Double-stranded DNA in exosomes: A novel biomarker in cancer detection. Cell Research, 24(6), Article 6. https://doi.org/10.1038/cr.2014.44

Udomruk, S., Phanphaisarn, A., Kanthawang, T., Sangphukieo, A., Sutthitthasakul, S., Tongjai, S., Teeyakasem, P., Thongkumkoon, P., Orrapin, S., Moonmuang, S., Klangjorhor, J., Pasena, A., Suksakit, P., Dissook, S., Puranachot, P., Settakorn, J., Pusadee, T., Pruksakorn, D., & Chaiyawat, P. (2023). Characterization of Cell-Free DNA Size Distribution in Osteosarcoma Patients. Clinical Cancer Research, 29(11), 2085–2094. https://doi.org/10.1158/1078-0432.CCR-22-2912

Ulz, P., Perakis, S., Zhou, Q., Moser, T., Belic, J., Lazzeri, I., Wölfler, A., Zebisch, A., Gerger, A., Pristauz, G., Petru, E., White, B., Roberts, C. E. S., John, J. St., Schimek, M. G., Geigl, J. B., Bauernhofer, T., Sill, H., Bock, C., … Speicher, M. R. (2019). Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications, 10, 4666. https://doi.org/10.1038/s41467-019-12714-4

Ulz, P., Thallinger, G. G., Auer, M., Graf, R., Kashofer, K., Jahn, S. W., Abete, L., Pristauz, G., Petru, E., Geigl, J. B., Heitzer, E., & Speicher, M. R. (2016). Inferring expressed genes by whole-genome sequencing of plasma DNA. Nature Genetics, 48(10), 1273–1278. https://doi.org/10.1038/ng.3648

van der Pol, Y., Moldovan, N., Verkuijlen, S., Ramaker, J., Boers, D., Onstenk, W., de Rooij, J., Bahce, I., Pegtel, D. M., & Mouliere, F. (2022). The Effect of Preanalytical and Physiological Variables on Cell-Free DNA Fragmentation. Clinical Chemistry, 68(6), 803–813. https://doi.org/10.1093/clinchem/hvac029

Vasioukhin, V., Anker, P., Maurice, P., Lyautey, J., Lederrey, C., & Stroun, M. (1994). Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. British Journal of Haematology, 86(4), 774–779. https://doi.org/10.1111/j.1365-2141.1994.tb04828.x

Vong, J. S. L., Tsang, J. C. H., Jiang, P., Lee, W.-S., Leung, T. Y., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2017). Single-Stranded DNA Library Preparation Preferentially Enriches Short Maternal DNA in Maternal Plasma. Clinical Chemistry, 63(5), 1031–1037. https://doi.org/10.1373/clinchem.2016.268656

Wang, L., Xie, L., Zhang, Q., Cai, X., Tang, Y., Wang, L., Hang, T., Liu, J., & Gong, J. (2015). Plasma nuclear and mitochondrial DNA levels in acute myocardial infarction patients. Coronary Artery Disease, 26(4), 296–300. https://doi.org/10.1097/MCA.0000000000000231

Wang, S., Meng, F., Li, M., Bao, H., Chen, X., Zhu, M., Liu, R., Xu, X., Yang, S., Wu, X., Shao, Y., Xu, L., & Yin, R. (2023). Multidimensional Cell-Free DNA Fragmentomic Assay for Detection of Early-Stage Lung Cancer. American Journal of Respiratory and Critical Care Medicine, 207(9), 1203–1213. https://doi.org/10.1164/rccm.202109-2019OC

Wang, S., Tang, J., Sun, T., Zheng, X., Li, J., Sun, H., Zhou, X., Zhou, C., Zhang, H., Cheng, Z., Ma, H., & Sun, H. (2017). Survival changes in patients with small cell lung cancer and disparities between different sexes, socioeconomic statuses and ages. Scientific Reports, 7(1), 1339. https://doi.org/10.1038/s41598-017-01571-0

Wong, F. C. K., Sun, K., Jiang, P., Cheng, Y. K. Y., Chan, K. C. A., Leung, T. Y., Chiu, R. W. K., & Lo, Y. M. D. (2016). Cell-free DNA in maternal plasma and serum: A comparison of quantity, quality and tissue origin using genomic and epigenomic approaches. Clinical Biochemistry, 49(18), 1379–1386. https://doi.org/10.1016/j.clinbiochem.2016.09.009

Xavier, C. P. R., Caires, H. R., Barbosa, M. A. G., Bergantim, R., Guimarães, J. E., & Vasconcelos, M. H. (2020). The Role of Extracellular Vesicles in the Hallmarks of Cancer and Drug Resistance. Cells, 9(5). https://doi.org/10.3390/cells9051141

Yan, Y., Guo, G., Huang, J., Gao, M., Zhu, Q., Zeng, S., Gong, Z., & Xu, Z. (2020). Current understanding of extrachromosomal circular DNA in cancer pathogenesis and therapeutic resistance. Journal of Hematology & Oncology, 13(1), 124. https://doi.org/10.1186/s13045-020-00960-9

Yang, C.-F. J., Chan, D. Y., Speicher, P. J., Gulack, B. C., Wang, X., Hartwig, M. G., Onaitis, M. W., Tong, B. C., D'Amico, T. A., Berry, M. F., & Harpole, D. H. (2016). Role of Adjuvant Therapy in a Population-Based Cohort of Patients With Early-Stage Small-Cell Lung Cancer. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 34(10), 1057–1064. https://doi.org/10.1200/JCO.2015.63.8171

Yousefi, S., Gold, J. A., Andina, N., Lee, J. J., Kelly, A. M., Kozlowski, E., Schmid, I., Straumann, A., Reichenbach, J., Gleich, G. J., & Simon, H.-U. (2008). Catapult-like release of mitochondrial DNA by eosinophils contributes to antibacterial defense. Nature Medicine, 14(9), 949–953. https://doi.org/10.1038/nm.1855

Zhang, J., Patel, L., & Pienta, K. J. (2010). Targeting chemokine (C-C motif) ligand 2 (CCL2) as an example of translation of cancer molecular biology to the clinic. Progress in Molecular Biology and Translational Science, 95, 31–53. https://doi.org/10.1016/B978-0-12-385071-3.00003-4

Zhang, R., Nakahira, K., Guo, X., Choi, A. M. K., & Gu, Z. (2016). Very Short Mitochondrial DNA Fragments and Heteroplasmy in Human Plasma. Scientific Reports, 6(1), Article 1. https://doi.org/10.1038/srep36097

Zhitnyuk, Y. V., Koval, A. P., Alferov, A. A., Shtykova, Y. A., Mamedov, I. Z., Kushlinskii, N. E., Chudakov, D. M., & Shcherbo, D. S. (2022). Deep cfDNA fragment end profiling enables cancer detection. Molecular Cancer, 21(1), 26. https://doi.org/10.1186/s12943-021-01491-8

Zhou, Z., Cheng, S. H., Ding, S. C., Heung, M. M. S., Xie, T., Cheng, T. H. T., Lam, W. K. J., Peng, W., Teoh, J. Y. C., Chiu, P. K. F., Ng, C.-F., Jiang, P., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2021). Jagged Ends of Urinary Cell-Free DNA: Characterization and Feasibility Assessment in Bladder Cancer Detection. Clinical Chemistry, 67(4), 621–630. https://doi.org/10.1093/clinchem/hvaa325

Zhu, J., Chen, S., Zhang, F., & Wang, L. (2018). Cell-free eccDNAs: A new type of nucleic acid component for liquid biopsy? Molecular Diagnosis & Therapy, 22(5), 515–522. https://doi.org/10.1007/s40291-018-0348-6

Zhu, J., Huang, J., Zhang, P., Li, Q., Kohli, M., Huang, C.-C., & Wang, L. (2020). Advantages of Single-Stranded DNA Over Double-Stranded DNA Library Preparation for Capturing Cell-Free Tumor DNA in Plasma. Molecular Diagnosis & Therapy, 24(1), 95–101. https://doi.org/10.1007/s40291-019-00429-7

Zhu, Y. M., Webster, S. J., Flower, D., & Woll, P. J. (2004). Interleukin-8/CXCL8 is a growth factor for human lung cancer cells. British Journal of Cancer, 91(11), 1970–1976. https://doi.org/10.1038/sj.bjc.6602227

2

DEVELOPMENT OF BROAD-RANGE CELL-FREE DNA SEQUENCING (BRCFDNA-SEQ)

2.1 Abstract

Next-generation sequencing (NGS) workflows and downstream observations are interconnected. Hence, understanding how nucleic extraction and library preparations directly influence the portrayal of purified cell-free DNA from plasma is crucial. In this chapter, we document the design of an optimal pipeline for recovering low-molecular weight uscfDNA from plasma. Ultimately, we show that coupling the SPRI magnetic bead extraction method or a modified silica column-based Qiagen nucleic acid extraction method with a single-stranded DNA library preparation can reveal the presence of low-molecular cell-free DNA in healthy plasma. Further, treatment with single-stranded and double-stranded specific nucleases suggests that ultrashort cell-free DNA is primarily single-stranded. The resulting development of this NGS workflow is titled "Broad-Range Cell-Free DNA Sequencing" (BRcfDNA-Seq) and is

recommended for future studies for those interested in investigating ultrashort and mononucleosomal cell-free DNA from plasma samples.

2.2 Introduction

2.2.1 cfDNA Processing Workflow

The depicted size of cell-free DNA is directly influenced by the preanalytical workflow, including blood collection and processing. These processes include DNA purification (Lampignano et al., 2020), the type of library preparation (Barlow et al., 2016; Sanchez et al., 2021), and the bioinformatic pipeline used (Chen et al., 2022). To test the hypothesis that ultrashort cell-free DNA in plasma exists, which could also be single-stranded, a workflow must be assembled which is optimized for low molecular nucleic acid weight DNA of all strandedness conformations.

2.2.2 DNA Purification

For proper assessment of cell-free DNA downstream methods, ideally, the cfDNA in biofluids must be purified (Sidransky, 1997). In 1869, the first DNA extraction attempt was performed by Friedrich Miescher, who accidentally isolated DNA when attempting to study the protein of leukocytes derived from the pus collected from surgical bandages. During these observations, he noticed that the consistency of the substance differed from the familiar protein appearance (Dahm, 2005). In fact, the precipitation appeared during the addition of acid but would dissolve in an alkaline solution leading to his coining of precipitation as a “nucleic acid.”

Since that discovery, an assortment of DNA extractions have been developed. Some examples include size-exclusive chromatography (Ellegren & Låås, 1989), ion-exchange chromatography (Budelier & Schorr, 2001), alkaline extraction (Birnboim & Doly, 1979), salting-out methodology (Miller et al., 1988), cetyltrimethylammonium bromite extraction (CTAB)(Aboul-Maaty & Oraby, 2019), magnetic beads(Hawkins et al., 1994), and filter-paper based DNA extraction (Shi et al., 2018).

One method ideal for low-molecular-weight nucleic acid purification is the phenol-chloroform-based DNA extraction method (Kirby, 1956; Sambrook & Russell, 2006). Phenol-chloroform extraction is a liquid-liquid extraction that separates molecules based on their solubility properties (Tshepelevitsh et al., 2017). Chloroform mixed with phenol efficiently denatures proteins, and when mixed with an aqueous solution of a biofluid, it will generate a lower organic phase and an upper aqueous phase (Sambrook & Russell, 2006). In a basic pH phenol-chloroform, nucleic acids (DNA and RNA) reside in the aqueous phase due to their charged phosphate backbone (Avison, 2006). Proteins have various charged and uncharged domains, and combined with phenol, will precipitate in the interface between the two layers. Simultaneously, lipids and other non-polar debris dissolve in the lower organic phase.

Another method is the solid-phase nucleic acid extraction principle used in commercial kits and appreciated for its quickness and efficiency. The Qiagen Circulating Nucleic Acid Kit is one example of a commercial solid-phase extraction kit and is often benchmarked as the gold standard for cell-free DNA extraction (Diefenbach et al., 2018). Under chaotropic conditions, the solid phase sorbent surface (usually in the form of a membrane in a column) will absorb nucleic acids depending on the environmental pH and salt content while repelling other molecules, which are later washed away (McCormick, 1989). Some examples of types of solid phase surfaces are glass particles, diatomaceous earth, or magnetic beads(Ali et al., 2017).

Chaotropic salts disrupt the hydrogen bonding in proteins releasing DNA or RNA-binding proteins but do not disrupt nucleic acids themselves (Farrah et al., 1981). The released DNA is inclined to adsorb to the surface of the silica. Contaminants are removed with a wash buffer, and nucleic acids can be eluted late with an aqueous buffer.

Solid phase reversible immobilization (SPRI) beads are normally used for cleanup of DNA (DeAngelis et al., 1995). SPRI beads have paramagnetic properties (magnetic under a magnetic field). Each bead is made of polystyrene, surrounded by a magnetite layer, and coated with carboxyl molecules. When DNA is in a binding buffer of polyethylene glycol (PEG) and salt, they bind to the carboxyl groups on the bead surface. The binding capacity of SPRI beads is immense, demonstrating the ability to bind 3µg of DNA with only 1uL of beads. Size selection can be performed by SPRI beads by modifying the PEG volume, which affects the DNA fragments binding to the beads. As the ratio of SPRI beads used increases, the total proportion of the amount of PEG and salt increases, promoting the lower molecular weight of DNA to bind. Thus, it can also be used in this manner to purify DNA from biofluids.

Isopropanol is an important component that will affect the size of DNA captured alongside the bead, and greater volumes are used when attempting to capture RNA molecules that are shorter than DNA. Isopropanol lowers the dielectric constant shield around the molecules and allows salts to bind to negative phosphates of the nucleic acids to neutralize them. Resultingly, they become less soluble and prone to precipitate out of the solution (Green & Sambrook, 2017).

2.2.3 DNA Quantification Considerations

Compared to regular genomic DNA residing in the nucleus of cells, the characteristics of cell-free DNA introduce challenges for its extraction. Firstly, the cell-free DNA concentration in plasma appears considerably low at the ~30ng/ml range (1.8-44ng/ml). However, this value is influenced by the quantification method (Fleischhacker & Schmidt, 2007; Perkins et al., 2012). Florescence, spectrometry, and amplicon-based PCR methods will differ in their reported quantification of cfDNA due to their diverse mechanisms of action (Bronkhorst et al., 2019). The inability of quantification methods to differentiate between single-stranded and double-stranded DNA also adds to the ambiguity of cell-free DNA populations (Nakayama et al., 2016).

2.2.4 Library Preparation Considerations

Since the discovery of DNA as a double-stranded nucleic acid molecule, sequencing protocols have evolved rapidly over the last few decades (Lander et al., 2001; Sanger et al., 1977; Watson & Crick, 1953). Next-generation sequencing (NGS) differs from Sanger sequencing by being able to sequence multiple DNA fragments simultaneously in parallel and providing identification of multiple genetic regions per sequencing run (Reis-Filho, 2009). NGS requires common workflow protocols, including sample processing, library preparation, sequencing, and bioinformatics analysis of the data. All sequencing technologies require a specific library preparation for the DNA fragments for loading on the instrument for the sequencing (Hess et al., 2020).

Prototypical library preparation for genomic DNA has the following steps: mechanical fragmentation, enzymatic reactions for adapter ligation, size selection and cleanup, index library amplification, and quantification. For genomic sequencing, fragmentation can be achieved mechanically by utilizing ultrasonic shearing to generate cavitation bubbles such as using a bioruptor, ultrasonication, or with digestion enzymes (Hess et al., 2020). Library preparation for cell-free DNA has a similar workflow but does not require mechanical fragmentation since cell-free DNA already presents in a fragmented form (van der Pol et al., 2022).

After fragmentation, adapters are ligated to the end of the DNA molecules. Traditionally or double-stranded DNA, the ends may contain overhangs that need to be repaired or blunted before phosphorylation of the 5′ prime end and A-tailing of the 3′ ends to allow for ligation of the sequencing adapters (Head et al., 2014).   T4 polynucleotide kinase, T4 DNA polymerase, and Klenow Large fragments are commonly used for this purpose (Head et al., 2014). Finally, the library is amplified using a polymerase chain reaction to increase the DNA bulk for downstream sequencing.

During adapter ligation, a possible byproduct is adapter dimers. Adapter ligation is normally performed at a 10:1 (adapters to native DNA fragment) ratio. Skewing this ratio may promote greater adapter-dimers formation, which will dominate in the PCR amplification steps and lose valuable information (Head et al., 2014). Adapter dimers can be removed with bead-based size-optimized cleanup steps to preserve larger molecular weight DNA while discarding smaller molecules (Head et al., 2014). This bead cleanup step for removing remnant enzymes, buffers (Kircher et al., 2012), and adapter is one aspect that needs to be considered when optimizing for shorter nucleic acids.

2.2.5 Double Versus Single-strand Library Considerations

The predominant type of cfDNA is normally thought to be the 167bp mononucleosomal cell-free DNA (mncfDNA) (Snyder et al., 2016). Hence, double-stranded library construction was typically used in NGS cfDNA studies. As stated earlier, during adapter dimer ligation, for double-stranded library preparation, the overhangs of each dsDNA molecule must be polished, which causes the dsDNA molecule to lose a portion of its original sequence (Avgeris et al., 2021). Additionally, since the adapters are double-stranded, they will ignore single-stranded DNA molecules and exclude them from the final library preparation (Mouliere et al., 2014).

Interestingly, this bias is a common scenario in studying ancient DNA samples where the DNA is usually highly fragmented and at low concentrations (Allentoft et al., 2012). By utilizing a single-stranded library preparation, they were able to sequence the genome of a fossilized extracted DNA, which through time, can become fragmented and single-stranded (Gansauge & Meyer, 2013; Meyer et al., 2012). Single-stranded DNA library preparations heat denature duplex template DNA, separating them into two single-stranded templates prior to adapter ligation. This denaturation allows for the incorporation of both blunt end and nicked dsDNA and ssDNA molecules. Svante Pääbo, who led these studies, eventually received the Noble Prize for Physiology or Medicine in 2022.

Several studies have attempted to evaluate the whole spectrum of cfDNA by comparing a single-stranded library approach was applied in comparison to the conventional library approach on the same DNA extracts (Burnham et al., 2016; Sanchez et al., 2021; Snyder et al., 2016). They demonstrated that the ssDNA library preparation is more sensitive to cfDNA of a broad range of types and lengths (Figure 2.1). The finding in these reports suggested that a considerable fraction of genomic cfDNA is non-nucleosomal and subject to nuclease degradation (Sanchez et al., 2021).

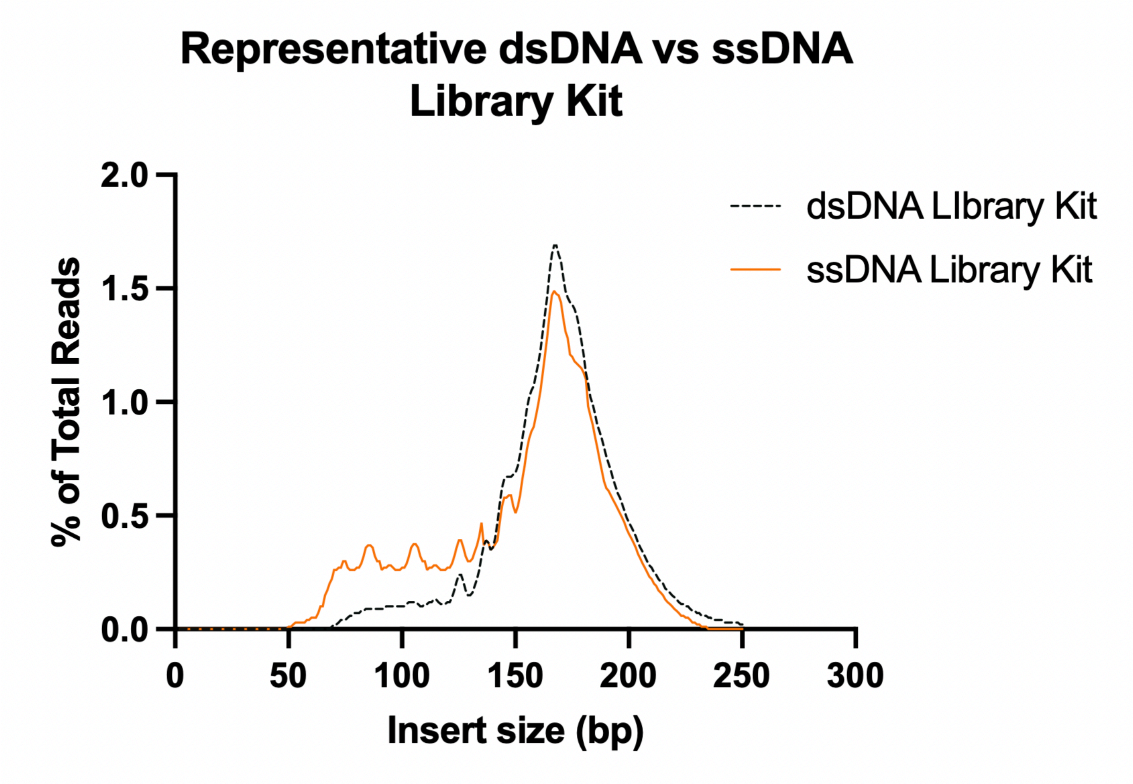

Figure 2.1 Representative fragment profile generated by double-stranded and single-stranded library preparation. Single-stranded library preparations are more sensitive for representing shorter cfDNA fragments below 80bp, which is not readily detectable by double-stranded library preparation (based on (Burnham et al., 2016; Sanchez et al., 2021; Snyder et al., 2016).

2.2.6 Chapter Goals

With these considerations of cell-free DNA extraction and library preparation, in this chapter, we document the development of an ultrashort single-stranded cfDNA-optimized sequencing pipeline (BRcfDNA-Seq) (Figure 2.2A and B). This pipeline incorporates an ultrashort single-stranded cfDNA (uscfDNA) extraction method and single-stranded library preparation. The extraction method utilizes both Solid Phase Reversible Immobilization magnetic beads (SPRI) and phenol:chloroform:isoamyl alcohol to retain low molecular weight fragments in plasma. It leverages a high ratio of isopropanol to create a DNA-phobic environment that precipitates out nucleic acids and proteins before isolating the aqueous nucleic acid-containing portion with phenol:chloroform isoamyl alcohol. Subsequent magnetic bead washes help retain the uscfDNA and reduce unwanted contaminants that may affect downstream library preparation enzymes (Figure 2A).

The SPRI extraction method was compared to two other methods, the standard protocol of the commercial silica column-based extraction kit protocol (QIAGEN QIAamp Circulating Nucleic Acid Kit, referred to as QiaC) and the miRNA protocol of the QIAamp Circulating Nucleic Acid Kit, referred to as QiaM. The QiaM protocol uses an increased volume of isopropanol, lysis, and binding buffers designed for shorter nucleic acid retention (miRNA).

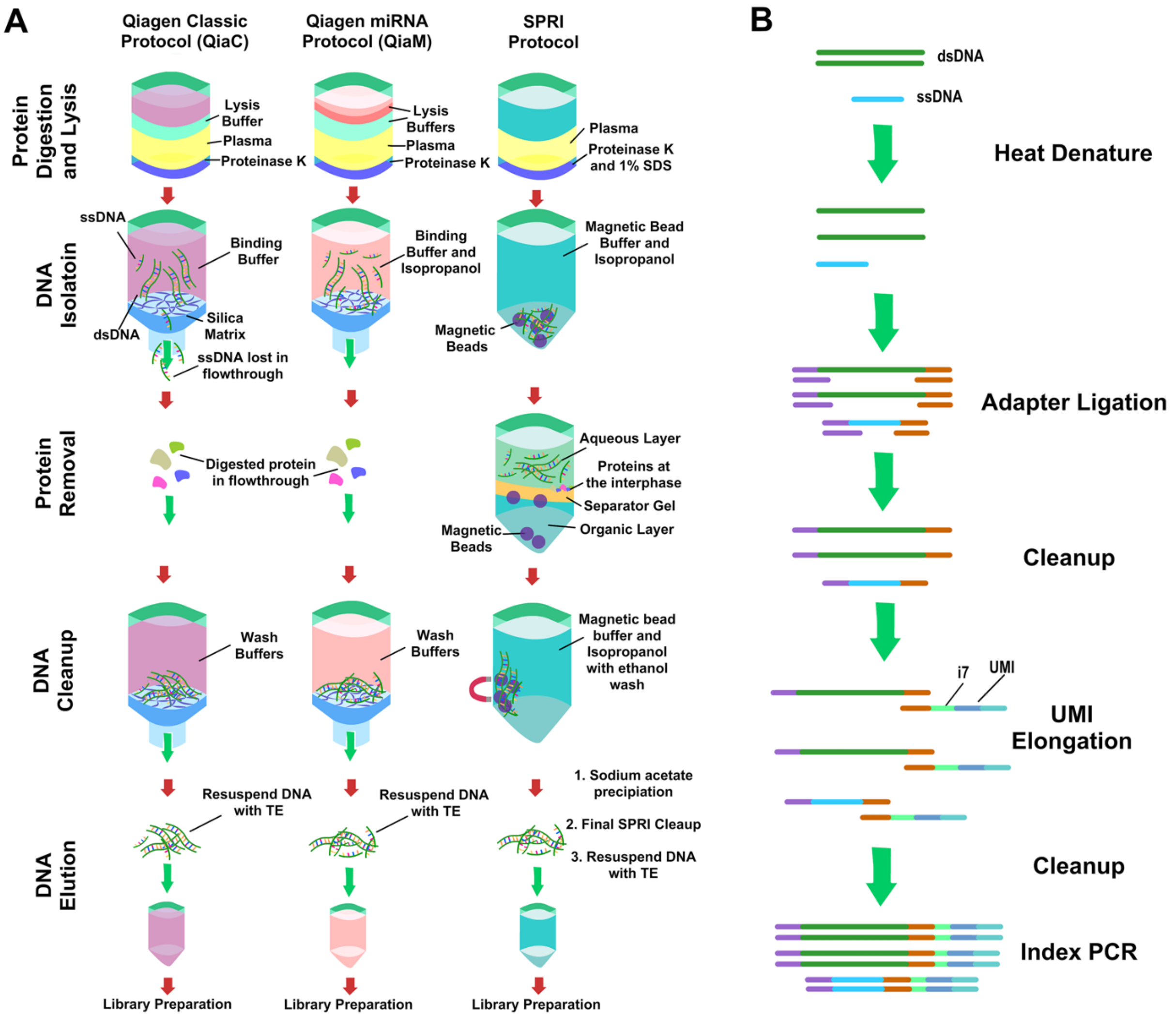

Figure 2.2 BRcfDNA-seq Schematic Protocol. A) Schematic of the three extraction protocols compared in this chapter. QiaC refers to the QIAGEN QIAamp Circulating Nucleic Acid Kit regular protocol. QiaM refers to the miRNA protocol of the QIAamp Circulating Nucleic Acid Kit. SPRI refers to the (Solid Phase Reversible Immobilization) magnetic beads and phenol:chloroform:isoamyl alcohol protocol. Compared to QiaC, QiaM, and SPRI protocols utilize an increased ratio of isopropanol to retain the low-molecular nucleic acids for downstream analysis. B) Single-stranded library preparation can incorporate dsDNA, ssDNA, and nicked DNA into the library. Unique molecular identifiers (UMI) are incorporated during the library preparation to remove PCR duplicates.

2.3 Results

2.3.1 BRcfDNA-Seq Can Purify and Visualize Ultrashort cfDNA in Plasma

Single-stranded libraries were made from cell-free DNA extracted by QiaM and SPRI methods which revealed a distinct cfDNA band at 200bp in the electropherogram corresponding to about 50bp of insert size (the library preparation adds about 150 bp-worth of adapters) compared to QiaC (Figure 2.3A). The mncfDNA peak (300bp before adapter removal) is present in all three extraction methods. This was a reproducible phenomenon with similar observations in multiple donors (Figure 2.3B and C). Upon sequencing and alignment to the human genome (Figure 2.3D-F), we observed size distribution of cell-free DNA with a uscfDNA peak (40-70) for QiaM and SPRI (Figure 2.3E and F) but not QiaC (Figure 2.3D). All three demonstrated the mncfDNA population with a peak at 167bp.

Similarly, we observed that using the QiaM, which incorporates higher isopropanol volume, will enhance the capture of low-molecular nucleic acids (Figure 2.4A and B). Interestingly, the miRNA purification protocol is associated with slower flow through the silica column. Scanning Electron Microscope images of the silica column indicate a reduction in pore size accompanied by sheet-like deposits possibly derived from increased isopropanol precipitation of organic matter in the plasma (Figure 2.4B). As part of BRcfDNA-Seq, these two extraction methods optimized for short DNA are partnered with a single-stranded library construction to fully visualize and examine the cfDNA population smaller than 100bp.

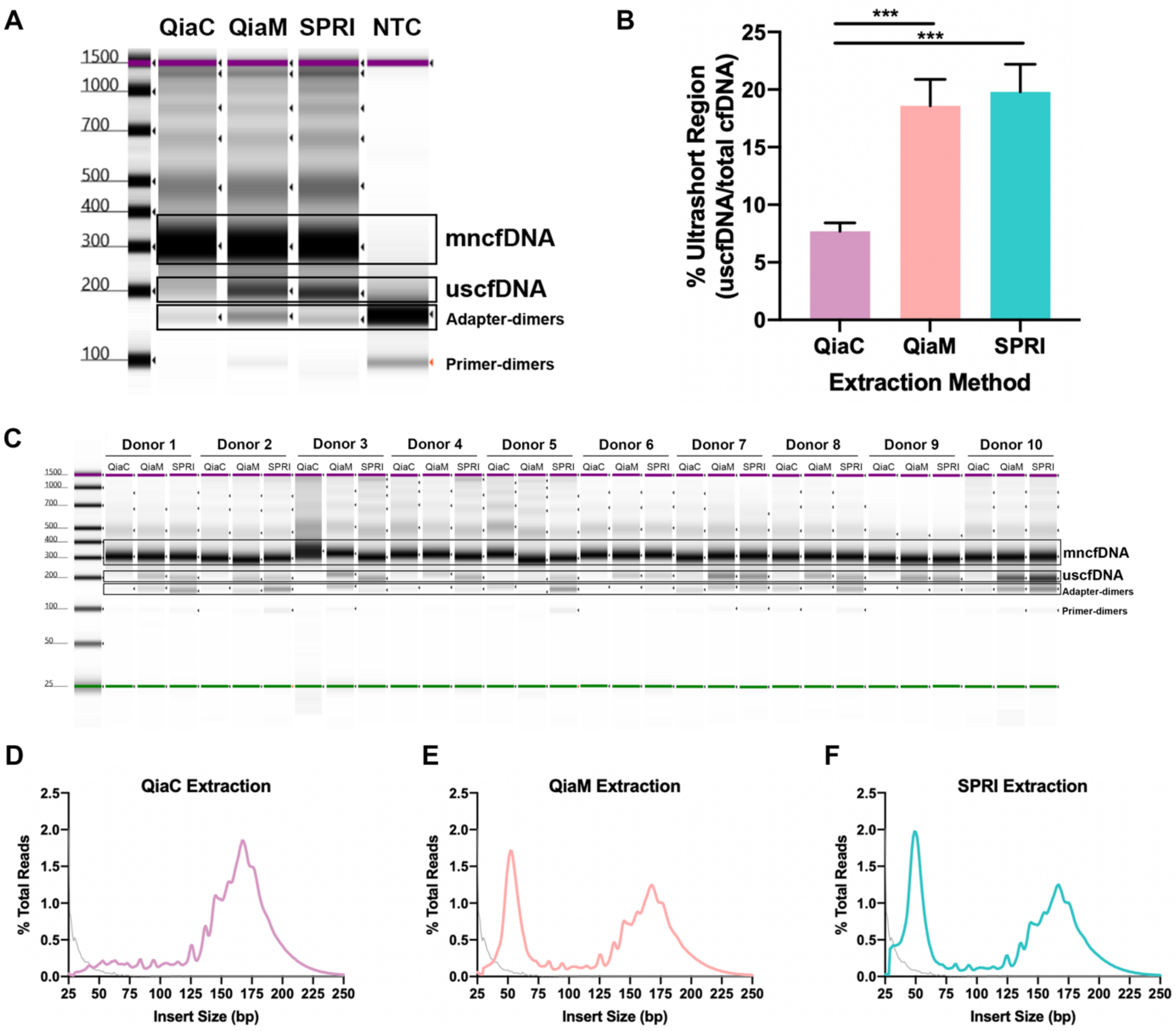

Figure 2.3 BRcfDNA-Seq reveals a population of ultrashort cfDNA fragments at 50nt in the plasma of healthy donors. A) BRcfDNA-seq using QiaM or SPRI reveals a distinct final NGS library uscfDNA band at 200bp (~50bp after adapter dimer subtraction) compared to QiaC. The electropherogram image was cropped for representative sizes. B) QiaM and SPRI extraction method can reproducibly isolate the 200bp fragment (180-250bp region in electropherogram) in ten human donors based on quantification of electrophoresis output (200bp band divided by (200bp + 300bp (250-350bp region)). Note: Bands are elongated with ~150bp of adapters on both sides). *** p < 0.001. The paired two-tailed student-test test was performed after ANOVA analysis. The bar graphs represent the standard error of Mean (SEM). C) Electropherogram images of ten healthy donors extracted with QiaC, QiaM, and SPRI show the presence of uscfDNA. (C) Alignment of total mapped reads from QiaM and SPRI extracted samples exclusively shows the native uscfDNA at 50bp in addition to the mncfDNA peak at ~167bp when adapters are trimmed. Extraction methods: QiaC (D - fuchsia), QiaM (E - pink), and SPRI (F -teal). The grey line represents sequencing of no template control.

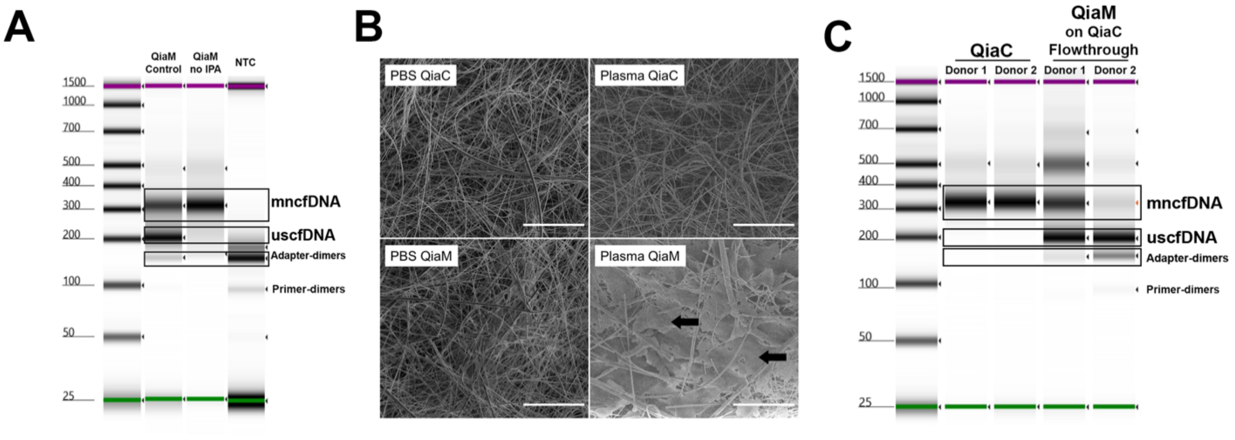

Figure 2.4 Inherent characteristics of the QiaM extraction protocol. A) The increased isopropanol (1.8ml to 2.3ml) is integral to retaining the uscfDNA from plasma. B) Scanning Electron Microscope images of the Qiagen silica filter show global sheet-like deposits (black arrows) only in QiaM extraction of plasma. Scale bars (white line) represent 50uM. C) Using the Qiagen Kit with a centrifuge (as opposed to vacuum-based), the flow through from a QiaC plasma extraction was subsequently extracted with QiaM to reveal the rescue of the uscfDNA band.

In an additional experiment, we used the QiaC protocol with a centrifuge (as opposed to a vacuum) to collect the flow-through of the binding step of the standard QIaC protocol for the presence of low-molecular-weight DNA. The QiaC flow through was subsequently extracted with QiaM (with increased isopropanol and lysis and binding buffers) to reveal that the uscfDNA could be rescued (Figure 2.4C). This also indicates that the QiaC protocol will lose low-molecular DNA.

2.3.2 uscfDNA is Predominantly Single-stranded

To examine the properties of strandedness, the extracted cfDNA supplemented with two control oligos (250 nt single-stranded and 350 bp double-stranded) was subject to strand-specific enzymes. When the DNA extracts were subject to dsDNA-specific DNase (dsDNase) digestion, the mncfDNA (300 bp) and the control dsDNA bands (500+ bp) showed an apparent reduction in intensity, as evidenced by the electrophoresis of the corresponding final libraries (Figure 2.5A and Figure 2.6A). In contrast, digestion by single-strand specific nucleases (S1, Exo 1, and P1) showed a significant reduction in the uscfDNA band and the control ssDNA band (400+bp) while preserving the mncfDNA band and the control dsDNA band (500+bp) in plasma extracted by both the QiaM and SPRI protocols.   Sequencing and alignment of these libraries confirmed the results from the electropherograms (Figure 2.5A, bottom panels, and Figure 2.7). These results strongly indicate the single-stranded nature of the uscfDNA.

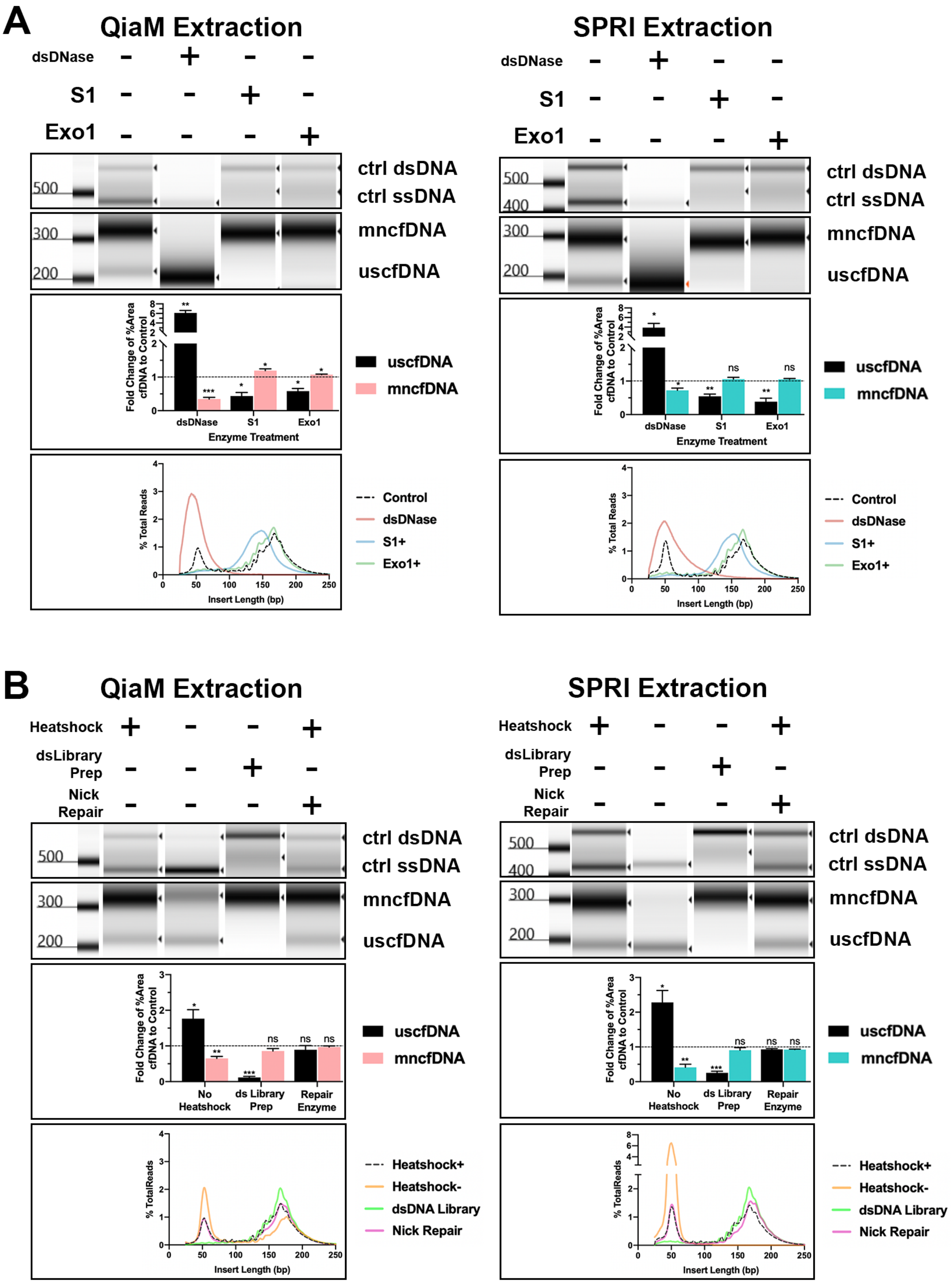
A
QiaM Extraction
SPRI Extraction
dsDNase
S1
Exo1
500
400
300
200
ctrl dsDNA
ctrl ssDNA
mncfDNA
uscfDNA
Fold Change of %Area cfDNA to Control
Enzyme Treatment
dsDNase
S1
Exo1
uscfDNA
mncfDNA
ns
% Total Reads
Insert Length (bp)
Control
dsDNase
S1+
Exo1+
B
Heatshock
dsLibrary Prep
Nick Repair
No Heatshock
ds Library Prep
Repair Enzyme
% TotalReads
Heatshock+
Heatshock-
dsDNA Library
Nick Repair

Figure 2.5 uscfDNA population is predominantly single-stranded. A) Size distribution of final library digestion with cfDNA supplemented with control oligos. B) Size distribution of library preparation variations with cfDNA supplemented with control oligos. For A and B, Top panels: electrophoretic visualization. Middle panels: quantification of the mapped reads belonging to the short (uscfDNA) or long population (mncfDNA). Bottom panels: mapped read size distribution. Reads with insert sizes under 25bp and above 250 were excluded from the plots. Bar graphs composed of plasma from three different human donors. The paired two-tailed student-test test was performed after ANOVA analysis. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$. Sequences from the lambda genome of 460bp dsDNA and 356nt ssDNA were used as positive controls. Adapter dimers have been cropped from the presented electropherograms. The bar graphs represent the standard error of Mean (SEM). Electropherogram images were cropped for representative sizes.

To corroborate the single-stranded nature of this DNA, we leveraged the differences in the adapter ligation chemistry between ssDNA and dsDNA library kits (Figure 2.5B). The uscfDNA peak was absent in the dsDNA library preparation (which only processes intact double-stranded substrates), suggesting that the ultrashort population (uscfDNA) is endogenously single-stranded. By contrast, the ssDNA library kits require initial heat denaturation (98^{o}C for 3 minutes) to efficiently incorporate dsDNA molecules into the library. By skipping this step, the presence of the 200bp population remained, suggesting that the uscfDNA population is mostly single-stranded (Figure 2.5B). Finally, to determine if the source of the uscfDNA derived from nicked dsDNA, we pretreated the extracted nucleic acids with a nick repair enzyme but did not observe a reduction of ultrashort fragments in the final library. This suggests that the vast majority of uscfDNA are not derived from nicked mncfDNA. These observations were consistent among three replicates (Figure 2.6A and B).

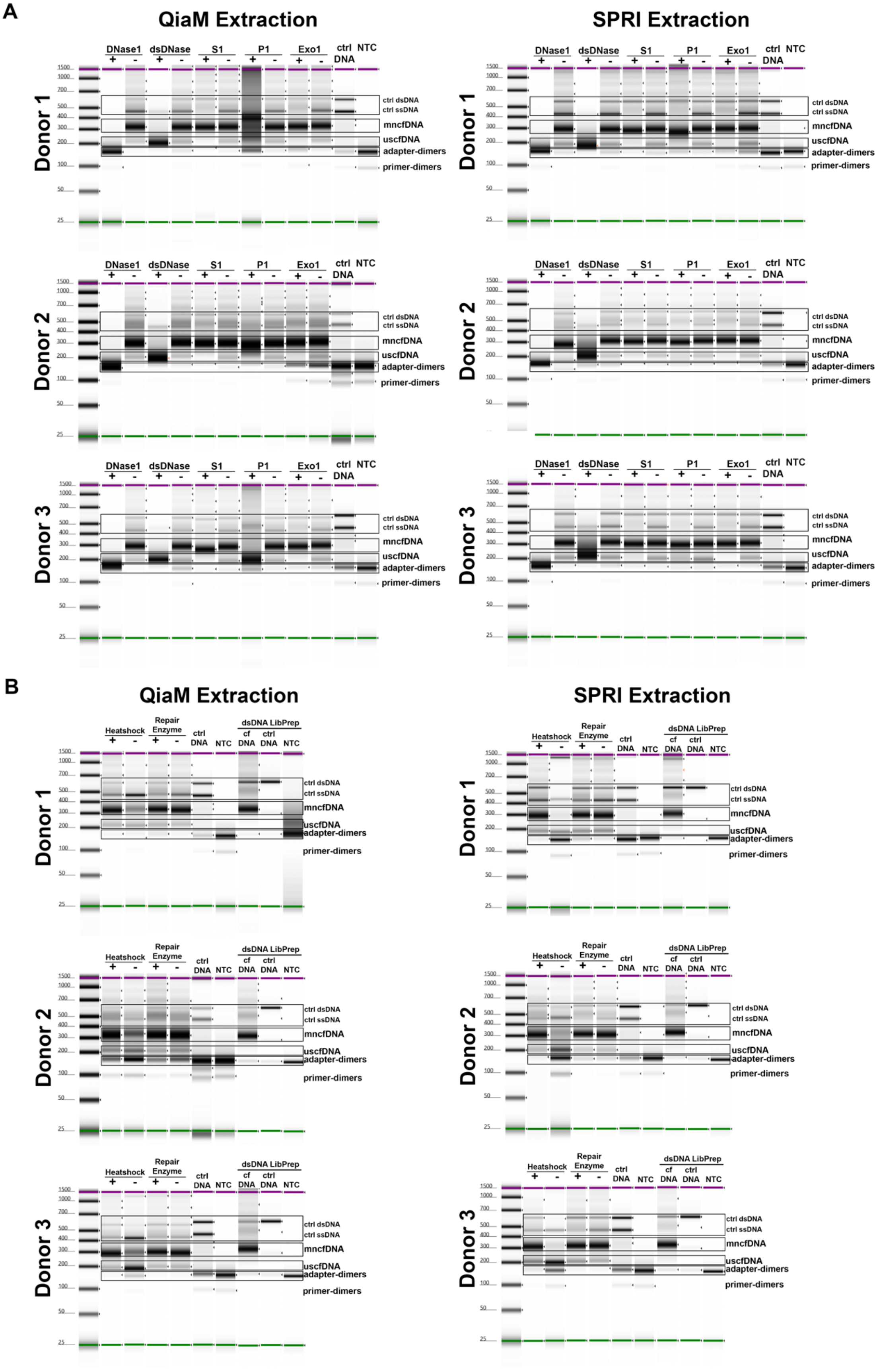
A
QiaM Extraction
SPRI Extraction
DNase1
dsDNase
S1
P1
Exo1
+
-
ctrl DNA
NTC
Donor 1
Donor 2
Donor 3
1500
1000
700
500
400
300
200
100
50
25
ctrl dsDNA
ctrl ssDNA
mncfDNA
uscfDNA
adapter-dimers
primer-dimers
B
QiaM Extraction
SPRI Extraction
Heatshock
Repair Enzyme
ctrl DNA
NTC
dsDNA LibPrep
cf DNA
ctrl DNA
NTC

Figure 2.6 Electropherograms of final libraries prepared from different treatments. A) Electropherograms of final libraries constructed from extracted cfDNA after nuclease digestion. B) Electropherograms of final libraries constructed from extracted cfDNA after undergoing ssDNA, dsDNase library preparation, and nick-repair enzyme treatment. Replicate experiments using plasma from three healthy donors extracted by QiaM and SPRI.

The alignment of sequenced digestion libraries recapitulated the findings previously mentioned with some interesting observations (Figure 2.5A, B, and 2.7A and B). Firstly, the S1-treated samples showed a 10bp downshift in the modality of the mncfDNA peak (from 160 to 150bp). Secondly, the S1 and nick-repair enzyme treatment flattened the periodicity on the left side of the mncfDNA peak. These observations suggest that the 10bp periodicity may result from nicked mncfDNA at certain fragment lengths. The S1 enzyme may also be digesting jagged edges, flanking the mncfDNA.

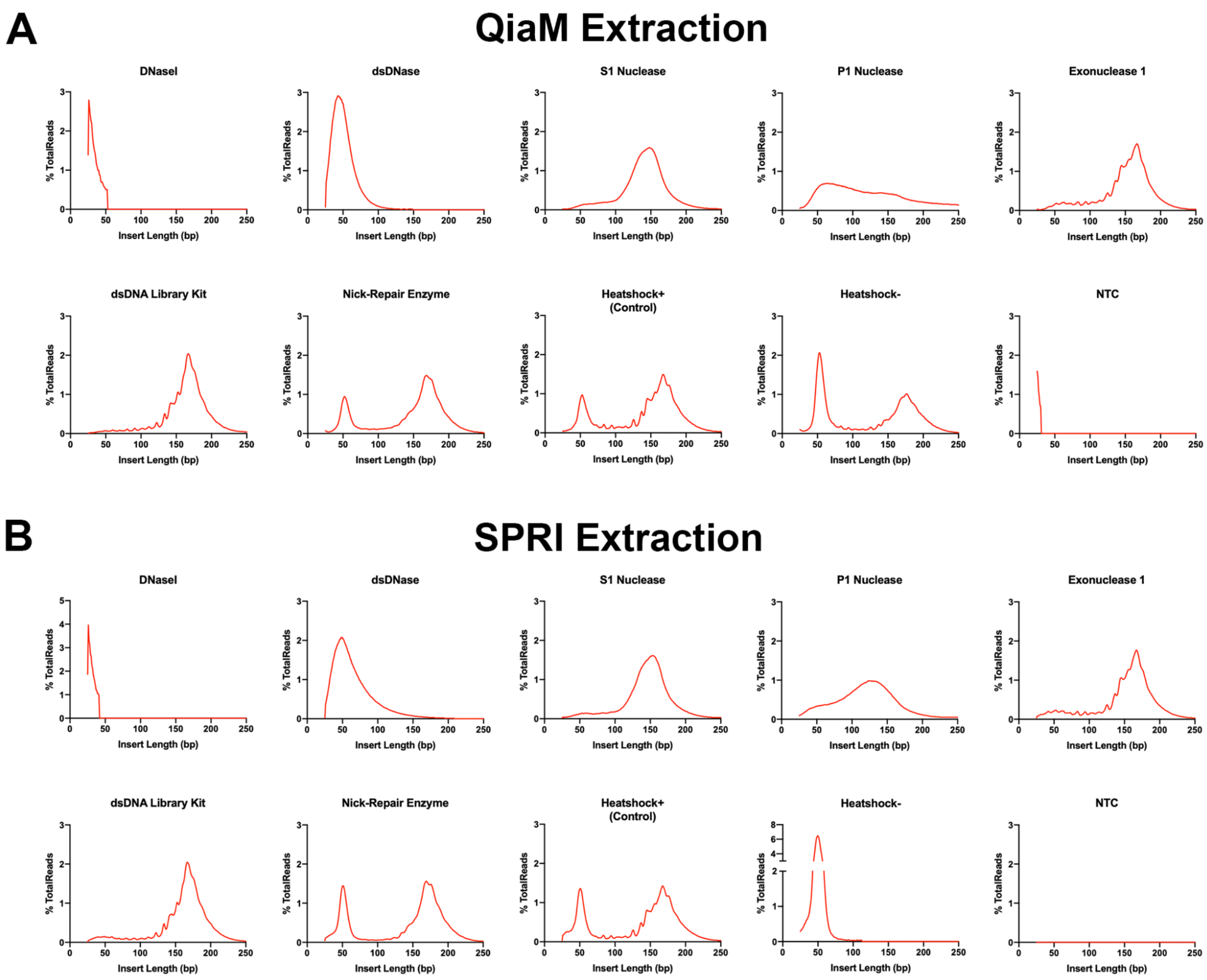

Figure 2.7 Fragment length distribution of aligned reads from samples that underwent digestions or variations in the library prep method. Alignment of sequenced libraries to human genome pretreated by digestions and library preparation variations from Donor 1 of Sup Fig 3 extracted by QiaM (A) and SPRI (B). Reads with insert sizes under 25bp and above 250bp were excluded from the plots.

2.4 Discussion

In this chapter, we demonstrate that the BRcfDNA-Seq pipeline reveals a novel population of ultrashort single-stranded cell-free DNA in plasma in addition to mononucleosomal cell-free DNA. BRcfDNA-Seq couples high isopropanol DNA extraction and single-stranded DNA library reparation to visualize the multiple conformations of DNA in the plasma sample. Both the extraction and library preparation are crucial aspects of BRcfDNA-Seq and, if modified, may change the downstream outcomes.

Both the QiaM and SPRI extraction methods appear to be essential steps in retaining the low-molecular-weight cfDNA in the final pool of extracted DNA. Although their conformations of capture are slightly different (SPRI using beads vs silica membranes for QiaM), both methods rely on the high volumes of isopropanol (He et al., 2022) to capture uscfDNA (Figure 2.4A and C). Further optimization can likely be made to improve the extraction step. Increasing isopropanol is essential, but perhaps an even greater volume would improve yields. Comparing and contrasting isopropanol with ethanol which also lowers the dielectric constant shield around DNA molecules, could also be another direction for improving yield(He et al., 2022). These alcohols predominantly differ in their inherent volatility, affecting their ability to be completely removed or their tendency to coprecipitate DNA with other salts (He et al., 2022).

Additionally, during the time frame in which this observation was made, three other research groups reported the presence of ultrashort single-strand fragments in plasma using distinctly different methods for extraction (Cheng et al., 2022; Hisano et al., 2021; Hudecova et al., 2021). One group used a conventional phenol-chloroform-based extraction method (Hisano et al., 2021), whereas another group also used magnetic beads with a commercial nucleic acid extraction kit (Hudecova et al., 2021). It is unclear whether increased isopropanol was used in the magnetic bead methodology. The third method reported used 10nt biotinylated capture probes with randomized base pairs to directly capture random single-stranded cell-free DNA in plasma (Cheng et al., 2022). Based on the ratio of uscfDNA and mncfDNA, the phenol-chloroform method apparently recovers uscfDNA at similar ratios observed in the QiaM and SPRI extraction methods used in this chapter. We observe that the peak at 50nt is higher than 167bp of the mncfDNA population. The direct capture method results in a very high uscfDNA: mncfDNA ratio (Cheng et al., 2022). This could be explained by the nature of the method in which it is biased towards shot single-stranded targets and would lower affinity to double-stranded mncfDNA. The magnetic bead protocol (Hudecova et al., 2021) appears to demonstrate a proportion where the peak of uscfDNA is slightly lower than the mncfDNA.

Evaluating the efficacy of the extraction between the five methods (3 other groups plus QiaM and SPRI) would be an important next step. The difficulty with this task is that specifically quantifying uscfDNA from the pool of purified DNA is challenging. A total DNA measurement would not provide any ratio relationships between uscfDNA and mncfDNA. Sequencing provides this relationship but requires careful spike-in experiments will be needed to clarify the recovered concentration compared to the spiked-in amount. In one uscfDNA study, spike-in with oligos of various sizes as a reference suggested that the uscfDNA had a concentration of 2.0 ng/ml (Cheng et al., 2022). Total cell-free DNA has been reported to range from 0 to 2000ng/ml. The concentration of 2.0ng/ml for uscfDNA can be viewed as both a very large or low concentration (Bryzgunova et al., 2021). Hence, it is difficult to assess the actual uscfDNA concentration without more sophisticated strategies.

Between all four groups, single-stranded library preparation was performed. This commonality in protocol demonstrates the importance of matching the method with the intended characteristics of the target. In addition to uscfDNA, heat-denaturing prior to adapter ligation will incorporate a wider variety of DNA conformations, such as jagged dsDNA, nicked

DNA, and blunt-ended dsDNA. Therefore, this justifies the "broad range' attribute of our NGS pipeline.

The other early reports on uscfDNA also performed similar experiments to validate the strandedness of uscfDNA. These included a comparison with dsDNA library kits and the omittance of heat-denature prior to single-stranded DNA preparation (Cheng et al., 2022; Hisano et al., 2021; Hudecova et al., 2021). Strand-specific digestions were also performed, although not as extensive in the variety of the nuclease we used in this chapter. These reproducible observations provide greater confidence in the strandedness characteristics of uscfDNA.

In summary, we show that the BRcfDNA-Seq NGS pipeline can successfully capture and demonstrate the presence of ultrashort cell-free DNA from healthy human plasma. Further analysis of the bioinformatic information regarding the genomic characteristics of uscfDNA will be discussed in Chapter 3.

2.5 References

Aboul-Maaty, N. A.-F., & Oraby, H. A.-S. (2019). Extraction of high-quality genomic DNA from different plant orders applying a modified CTAB-based method. Bulletin of the National Research Centre, 43(1), 25. https://doi.org/10.1186/s42269-019-0066-1

Ali, N., Rampazzo, R. de C. P., Costa, A. D. T., & Krieger, M. A. (2017). Current Nucleic Acid Extraction Methods and Their Implications to Point-of-Care Diagnostics. BioMed Research International, 2017, 9306564. https://doi.org/10.1155/2017/9306564

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., Willerslev, E., Zhang, G., Scofield, R. P., Holdaway, R. N., & Bunce, M. (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. Proceedings of the Royal Society B: Biological Sciences, 279(1748), 4724–4733. https://doi.org/10.1098/rspb.2012.1745

Avgeris, M., Marmarinos, A., Gourgiotis, D., & Scorilas, A. (2021). Jagged Ends of Cell-Free DNA: Rebranding Fragmentomics in Modern Liquid Biopsy Diagnostics. Clinical Chemistry, 67(4), 576–578. https://doi.org/10.1093/clinchem/hvab036

Avison, M. (2006). Measuring Gene Expression. Taylor & Francis. https://doi.org/10.4324/9780203889879

Barlow, A., Fortes, G. G., Dalén, L., Pinhasi, R., Gasparyan, B., Rabeder, G., Frischauf, C., Paijmans, J. L. A., & Hofreiter, M. (2016). Massive influence of DNA isolation and library preparation approaches on palaeogenomic sequencing data (p. 075911). bioRxiv. https://doi.org/10.1101/075911

Birnboim, H. C., & Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Research, 7(6), 1513–1523. https://doi.org/10.1093/nar/7.6.1513

Bronkhorst, A. J., Ungerer, V., & Holdenrieder, S. (2019). Comparison of methods for the quantification of cell-free DNA isolated from cell culture supernatant. Tumor Biology, 41(8), 1010428319866369. https://doi.org/10.1177/1010428319866369

Bryzgunova, O. E., Konoshenko, M. Y., & Laktionov, P. P. (2021). Concentration of cell-free DNA in different tumor types. Expert Review of Molecular Diagnostics, 21(1), 63–75. https://doi.org/10.1080/14737159.2020.1860021

Budelier, K., & Schorr, J. (2001). Purification of DNA by anion-exchange chromatography. Current Protocols in Molecular Biology, Chapter 2, Unit2.1B. https://doi.org/10.1002/0471142727.mb0201bs42

Burnham, P., Kim, M. S., Agbor-Enoh, S., Luikart, H., Valantine, H. A., Khush, K. K., & De Vlaminck, I. (2016). Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Scientific Reports, 6. https://doi.org/10.1038/srep27859

Chen, Y., Gong, Y., Dou, L., Zhou, X., & Zhang, Y. (2022). Bioinformatics analysis methods for cell-free DNA. Computers in Biology and Medicine, 143, 105283. https://doi.org/10.1016/j.compbiomed.2022.105283

Cheng, L. Y., Dai, P., Wu, L. R., Patel, A. A., & Zhang, D. Y. (2022). Direct capture and sequencing reveal ultra-short single-stranded DNA in biofluids. IScience, 25(10), 105046. https://doi.org/10.1016/j.isci.2022.105046

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. Developmental Biology, 278(2), 274–288. https://doi.org/10.1016/j.ydbio.2004.11.028

DeAngelis, M. M., Wang, D. G., & Hawkins, T. L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Research, 23(22), 4742–4743. https://doi.org/10.1093/nar/23.22.4742

Diefenbach, R. J., Lee, J. H., Kefford, R. F., & Rizos, H. (2018). Evaluation of commercial kits for purification of circulating free DNA. Cancer Genetics, 228–229, 21–27. https://doi.org/10.1016/j.cancergen.2018.08.005

Ellegren, H., & Låås, T. (1989). Size-exclusion chromatography of DNA restriction fragments: Fragment length determinations and a comparison with the behaviour of proteins in size-exclusion chromatography. Journal of Chromatography A, 467, 217–226. https://doi.org/10.1016/S0021-9673(01)93966-4

Farrah, S. R., Shah, D. O., & Ingram, L. O. (1981). Effects of chaotropic and antichaotropic agents on elution of poliovirus adsorbed on membrane filters. Proceedings of the National Academy of Sciences of the United States of America, 78(2), 1229–1232.

Fleischhacker, M., & Schmidt, B. (2007). Circulating nucleic acids (CNAs) and cancer—A survey. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 1775(1), 181–232. https://doi.org/10.1016/j.bbcan.2006.10.001

Gansauge, M.-T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nature Protocols, 8(4), 737–748. https://doi.org/10.1038/nprot.2013.038

Green, M. R., & Sambrook, J. (2017). Precipitation of DNA with Isopropanol. Cold Spring Harbor Protocols, 2017(8), pdb.prot093385. https://doi.org/10.1101/pdb.prot093385

Hawkins, T. L., O'Connor-Morin, T., Roy, A., & Santillan, C. (1994). DNA purification and isolation using a solid-phase. Nucleic Acids Research, 22(21), 4543–4544.

He, S., Cao, B., Yi, Y., Huang, S., Chen, X., Luo, S., Mou, X., Guo, T., Wang, Y., Wang, Y., & Yang, G. (2022). DNA precipitation revisited: A quantitative analysis. Nano Select, 3(3), 617–626. https://doi.org/10.1002/nano.202100152

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. BioTechniques, 56(2), 61–77. https://doi.org/10.2144/000114133

Hess, J. F., Kohl, T. A., Kotrová, M., Rönsch, K., Paprotka, T., Mohr, V., Hutzenlaub, T., Brüggemann, M., Zengerle, R., Niemann, S., & Paust, N. (2020). Library preparation for next generation sequencing: A review of automation strategies. Biotechnology Advances, 41, 107537. https://doi.org/10.1016/j.biotechadv.2020.107537

Hisano, O., Ito, T., & Miura, F. (2021). Short single-stranded DNAs with putative non-canonical structures comprise a new class of plasma cell-free DNA. BMC Biology, 19(1), 225. https://doi.org/10.1186/s12915-021-01160-8

Hudecova, I., Smith, C. G., Hänsel-Hertsch, R., Chilamakuri, C. S., Morris, J. A., Vijayaraghavan, A., Heider, K., Chandrananda, D., Cooper, W. N., Gale, D., Garcia-Corbacho, J., Pacey, S., Baird, R. D., Rosenfeld, N., & Mouliere, F. (2021). Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. Genome Research. https://doi.org/10.1101/gr.275691.121

Kirby, K. S. (1956). A new method for the isolation of ribonucleic acids from mammalian tissues. Biochemical Journal, 64(3), 405–408.

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research, 40(1), e3. https://doi.org/10.1093/nar/gkr771

Lampignano, R., Neumann, M. H. D., Weber, S., Kloten, V., Herdean, A., Voss, T., Groelz, D., Babayan, A., Tibbesma, M., Schlumpberger, M., Chemi, F., Rothwell, D. G., Wikman, H., Galizzi, J.-P., Riise Bergheim, I., Russnes, H., Mussolin, B., Bonin, S., Voigt, C., … Heitzer, E. (2020). Multicenter Evaluation of Circulating Cell-Free DNA Extraction and Downstream Analyses for the Development of Standardized (Pre)analytical Work Flows. Clinical Chemistry, 66(1), 149–160. https://doi.org/10.1373/clinchem.2019.306837

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing

Consortium. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860–921. https://doi.org/10.1038/35057062

McCormick, R. M. (1989). A solid-phase extraction procedure for DNA purification. Analytical Biochemistry, 181(1), 66–74. https://doi.org/10.1016/0003-2697(89)90394-1

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science (New York, N.Y.), 338(6104), 222–226. https://doi.org/10.1126/science.1224344

Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Research, 16(3), 1215.

Mouliere, F., El Messaoudi, S., Pang, D., Dritschilo, A., & Thierry, A. R. (2014). Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. Molecular Oncology, 8(5), 927–941. https://doi.org/10.1016/j.molonc.2014.02.005

Nakayama, Y., Yamaguchi, H., Einaga, N., & Esumi, M. (2016). Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions. PLoS ONE, 11(3), e0150528. https://doi.org/10.1371/journal.pone.0150528

Perkins, G., Yap, T. A., Pope, L., Cassidy, A. M., Dukes, J. P., Riisnaes, R., Massard, C., Cassier, P. A., Miranda, S., Clark, J., Denholm, K. A., Thway, K., Gonzalez De Castro, D., Attard, G., Molife, L. R., Kaye, S. B., Banerji, U., & de Bono, J. S. (2012). Multi-Purpose Utility of Circulating Plasma DNA Testing in Patients with Advanced Cancers. PLoS ONE, 7(11). https://doi.org/10.1371/journal.pone.0047020

Reis-Filho, J. S. (2009). Next-generation sequencing. Breast Cancer Research, 11(3), S12. https://doi.org/10.1186/bcr2431

Sambrook, J., & Russell, D. W. (2006). Purification of nucleic acids by extraction with phenol:chloroform. CSH Protocols, 2006(1), pdb.prot4455. https://doi.org/10.1101/pdb.prot4455

Sanchez, C., Roch, B., Mazard, T., Blache, P., Dache, Z. A. A., Pastor, B., Pisareva, E., Tanos, R., & Thierry, A. R. (2021). Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics. JCI Insight, 6(7), 144561. https://doi.org/10.1172/jci.insight.144561

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463–5467.

Shi, R., Lewis, R. S., & Panthee, D. R. (2018). Filter paper-based spin column method for cost-efficient DNA or RNA purification. PLOS ONE, 13(12), e0203011. https://doi.org/10.1371/journal.pone.0203011

Sidransky, D. (1997). Nucleic acid-based methods for the detection of cancer. Science (New York, N.Y.), 278(5340), 1054–1059. https://doi.org/10.1126/science.278.5340.1054

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. (2016). Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell, 164(0), 57–68. https://doi.org/10.1016/j.cell.2015.11.050

Tshepelevitsh, S., Hernits, K., Jenčo, J., Hawkins, J. M., Muteki, K., Solich, P., & Leito, I. (2017). Systematic Optimization of Liquid–Liquid Extraction for Isolation of Unidentified Components. ACS Omega, 2(11), 7772–7776. https://doi.org/10.1021/acsomega.7b01445

van der Pol, Y., Moldovan, N., Verkuijlen, S., Ramaker, J., Boers, D., Onstenk, W., de Rooij, J., Bahce, I., Pegtel, D. M., & Mouliere, F. (2022). The Effect of Preanalytical and Physiological Variables on Cell-Free DNA Fragmentation. Clinical Chemistry, 68(6), 803–813. https://doi.org/10.1093/clinchem/hvac029

Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature, 171(4356), Article 4356. https://doi.org/10.1038/171737a0

3

BIOINFORMATIC ASSESSMENT OF ULTRASHORT SINGLE-STRANDED CELL-FREE DNA AND ITS POTENTIAL AS A BIOMARKER FOR NON-SMALL CELL LUNG CARCINOMA

3.1 Abstract

In the previous chapter, using BRcfDNA-Seq, we identified a population of ultrashort single-stranded cell-free DNA in human plasma. This species has a modal size of 50nt and is distinctly separate from mono-nucleosomal cell-free DNA. In this chapter, we review the integral steps in bioinformatic pipelines for DNA analysis to examine the genomic characteristics of the sequenced cfDNA data between uscfDNA and mncfDNA. We hypothesize that uscfDNA possesses unique characteristics useful for disease detection. To this end, we used BRcfDNA-Seq to process the uscfDNA and mncfDNA within plasma samples from 18 non-cancer controls and 14 late-stage non-small cell lung carcinoma (NSCLC) patients with histologically classified adenocarcinoma. In comparison to mncfDNA, we assessed if analysis of functional element (FE) peaks, fragmentomics, End-Motifs, and G-Quadruplex (G-Quad)

signature prevalence which illustrate different biological processes, could be valuable features of uscfDNA for NSCLC determination. In non-cancer subjects, compared to mncfDNA, uscfDNA fragments had a 45.2-fold increased tendency to form FE peaks, were enriched in promoter, intronic, and exonic regions, demonstrated a distinct end-motif frequency profile, and presented with a 4.9-fold increase in G-Quad signature sequences. Within NSCLC subjects, only the uscfDNA population had discoverable FE peak-specific biomarker candidates and demonstrated different end-motif frequency candidates distinct from mncfDNA. Although both cDNA populations showed increased fragmentation in NSCLC, the G-Quad signatures were more discriminatory in uscfDNA. Compilation of the significant cfDNA features using principal component analysis reveals that the first five principal components of uscfDNA and mncfDNA had a cumulative explained variance of >80%. These observations further illustrate the dissimilarity between uscfDNA and mncfDNA, which justifies the exploration of uscfDNA as a promising new class of biomarker to augment pre-existing liquid biopsy approaches.

3.2 Introduction

3.2.1 Introduction to Non-somatic Mutation cfDNA Biomarker Features

Traditional cfDNA liquid biopsy leverages the observation that if a tumor is present, the circulating cfDNA may contain a proportion of mutated sequences (Ignatiadis et al., 2021). Certain cancer types, however, are not associated with any pathognomonic driver mutations, while a subset of cancers present with low concentrations of circulating tumor DNA (ctDNA) (Mouliere et al., 2018). Therefore, other strategies will need to be explored to improve current liquid biopsy sensitivity.

Alternatively, low-depth non-targeted sequencing (whole-genome sequencing of cfDNA samples) can yield useful information. This hinges on the hypothesis that the tumor microenvironment and other changes in the body (such as cancer-induced inflammatory states) may affect cfDNA presentation (van der Pol & Mouliere, 2019). Sequenced data requires a bioinformatic workflow to preprocess and align the data to the human genome. Once processed, the data can be investigated for specific genomic-related questions, which can become useful biomarkers to differentiate between non-cancer and cancer cohorts.

3.2.2 Next-Generation Sequencing Technologies for cfDNA

Although there are various methods to analyze cfDNA, few technologies provide as much granularity as next-generation sequencing since it can report individual bases of each fragment to provide a canvas for many types of analysis. The breakthrough in sequencing technology began with the development of Sanger sequencing technology which is efficient for sequencing short DNA fragments(Sanger et al., 1977). However, this technology was limited to the number of different sequences of fragments it could output. Next-generation sequencing (NGS) refers to a collection of technologies that use parallel sequencing arrays to produce millions of short-read sequences (Morey et al., 2013). In brief, during sequencing, the small DNA fragments generated during the library preparation adhere to a flow cell by hybridizing to complementary sequences compatible with the universal adapters. Bridge amplification occurs to generate clonal clusters of single-strand molecules (Voelkerding et al., 2009). Later, sequencing primers bind to the adapters replicating the single-strand sequences. As each base is added, a fluorescent color is revealed, which is captured by cameras to report which base is in the sequence. This process allows for massively parallel sequencing of all the fragments simultaneously.

There are several types of different NGS platforms. Illumina has dominated the market due to its early entry into the sector, aggressive patenting, and constant effort to reduce costs (Eisenstein, 2023; Phillips & Douglas, 2018). Alternatively, ThermoFisher's Ion Torrent technology does not use fluorescence but instead uses semiconductor technology. When nucleotides are synthesized during DNA polymerization, hydrogen ions are released, which are detected by pH meters. The technology provides shorter sequencing runs due to simpler chemistry compared to Illumina. Ion Torrent also has concerns with accuracy and sequencing errors with long homopolymers (Besser et al., 2018).

Newer technologies have emerged called "Third Generation Sequencing" platforms (Petersen et al., 2019). These sequencing technologies substantially improve on sequencing longer fragments, reducing the need to fragment genomic DNA first (van Dijk et al., 2018). This methodology is particularly useful for the assembly of genome, which requires high-depth sequencing and time-consuming analyses to stitch sequences back to their original orientation. In cfDNA biology, the ability for ultra-long read sequencing may not be necessary because they are already very fragmented, although there is growing interest for large molecular weight cfDNA (Choy et al., 2022; Yu et al., 2021).

Pacific Bioscience offers SMRT (Single Molecule Real Time) Sequencing and can generate reads as long as 20kbs. The PacBio RSII platform, however, is known to be associated with high error rates. Oxford Nanopore Technology is another technology capable of short or ultra-long reads (Reuter et al., 2015). The Oxford Nanopore provides the ability to automate the library preparation prior to sequencing as well as portable sequencers. In general, different sequencers are optimal for different situations. In this thesis, all sequencing was performed on the Illumina platform due to its availability and our focus on short reads.

3.2.3 Basic Bioinformatic Workflow

During sequencing, the Illumina binary base call (.bcl) sequence raw file is generated. These files measure the intensity of the light channel and identify the most likely base at a given position of a sequencing read. The real-time analysis software will write this base and quality to the .bcl file and are stored in a binary format (Cacho et al., 2016). The .bcl file needs to be converted to a .fastq file. Fastq files are text-based sequences that contain both the sequence and the quality metrics, and for paired-end sequencing, two .fastq files will be generated for read direction. Sequencing can be multiplexed since all samples can have their own unique index. Sequenced reads will need to be demultiplexed before the final analysis. There is a potential for index hopping, which can occur when indexes from different samples get swapped, causing reads to be demultiplexed incorrectly (Ros-Freixedes et al., 2018). To circumvent this, a completely independent index can be used for each sample.

The initial step in all bioinformatic pipelines is to clean up the reads with a quality control tool. Preprocessing reads will remove adapter sequences and poor-quality sequences. Quality control tools can be combined with trimming tools such as Trimmomatic (Bolger et al., 2014) to remove these adapters and low-quality sequences. Fastp is a combination tool combining these quality control steps (S. Chen et al., 2018).

After the reads have undergone quality control, they can be aligned to a reference genome. The human reference genome is constantly improving, and it is best to use the newest assembly. The currently most used assembly is the GRCh38.p14 (Church, 2022) which predominantly covers the euchromatic regions of the genome. Despite having millions of unknown bases, it was still useful as an alignment reference. Recently, a new reference that adds five full chromosome arms and more sequences than earlier references was released by utilizing the long-read capability of Oxford Nanopore and PacBio called the CHM13 T2T assembly (Nurk et al., 2022). Since the newest reference had not been updated with all annotations, the data in this chapter was aligned using the prior gold-standard reference.

Alignment to references can be aligners such as BWA-MEM (Li & Durbin, 2009), Bowtie2 (Langmead & Salzberg, 2012), CUSHAW3 (Y. Liu et al., 2014), or MOSIAK (Lee et al., 2014). During alignment, the bioinformatic tools will take the individual sequences and attempt to find which place in the genome the sequences best match. These aligners vary in their computational requirements and compatibility with the original sequencing platforms.

In common NGS assays, several steps can introduce technical biases and limit the accuracy of quantification. One of the considerations during NGS sequencing is that duplicates are formed during the final index PCR. PCR duplicates may skew the proportion of molecules seen in the final library. One method of overcoming this is to deduplicate based on sequence (Ebbert et al., 2016). This method will identify identical sequences with the same start and end point and only retain one copy. Deduplication is particularly relevant when determining the allele frequency of somatic mutations. If the duplicate wild type sequences are not removed, it may mask rare sequences like base pair mutations. However, the deduplication method may overcorrect if there are biological sequences with the same sequence length. Unique Molecular Identifiers (UMI) address this issue by tagging individual molecules (Chung et al., 2019; Kennedy et al., 2014). Having the molecules tracked individually can differentiate which, if there are two molecules (or more molecules) with the same sequence and length, should be preserved or removed. This allows for a more accurate portrayal of the profile of reads in the library.

One of the final ways to look at the NGS analysis is through visual summary methods. These visual methods summarize the NGS data with essential information, including mapping percentage, fragment length, overall map coverage, and coverage of each individual chromosome. One popular summarizer is Qualimap (García-Alcalde et al., 2012). The data encapsulated by this data is useful for determining some of the quality metrics of the sequencing run.

Generally, this is the typical workflow for bioinformatic preprocessing for NGS data. Next other bioinformatic strategies can be used for analyzing NGS data to answer biological questions of interest which are embedded in the sequencing of the data. In this chapter, we investigate genomic characteristics of the uscfDNA and mncfDNA regarding four domains. These domains are functional element peaks, fragmentomics, end-motif, and G-Quadruplex abundance.

3.2.4 Functional Elements Peaks

Since fragmentation of cell-free DNA is a non-random process, there is an opportunity to examine if specific genomic element identities coded by the cell-free DNA sequences demonstrate alterations in cancer states. The non-targeted nature of BRcfDNA-Seq will reveal the biological biases of both populations of cell-free DNA. The first foray into this analysis is identifying if uscfDNA fragments congregate in genomic positions forming "peaks" (Figure 3.1). If uscfDNA is abundant in peaks, depending on if the peaks are in proximity to regulatory regions, it may reflect that accessibility to nucleases is due to the epigenetic structural formations of the DNA-histone interactions (Oruba et al., 2020). Examining the functional element peaks based on genomic element categories only provides a bird's eye view of the approximate locations where functional element peaks land. In the same vein as cell-free miRNA analysis (Lawrie et al., 2008) or cell-free DMR analysis (Benelli et al., 2021), further examination of the specific identities of the fragments and their appearance patterns may give rise to new biomarkers between non-cancer and NSCLC samples.

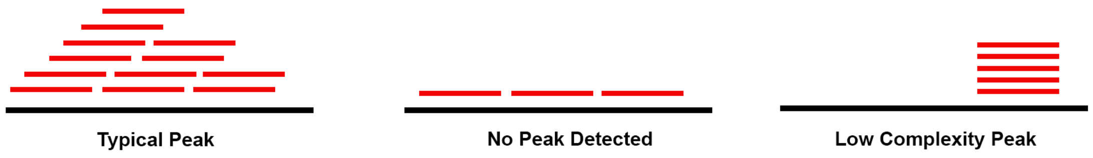

Figure 3.1 Schematic of MACS2 peak calling. MACS2 predicts peaks based on the conformation of aligned reads. Regions with low coverage reads or high coverage but low complexity are unlikely to be called as peaks (Based on (Landt et al., 2012).

3.2.5 Fragmentomics

Initial interest in fragmentomic analysis of cell-free DNA derived from the observation that the modal peak of fetal cell-free DNA was shorter than maternal DNA molecules (Lo et al., 2010). The size distribution of fetal cell-free DNA displayed binary peaks at 167bp (nucleosome with linker DNA) and 143bp (nucleosome without linker DNA), contrasting with maternal DNA, which only had a modal peak of 167bp. This revealed that the fragmentation patterns shown in cell-free DNA did not reflect a random process but instead were related to the structural biology of the genome from apoptotic cells (Hu et al., 2022). Developing this idea further, researchers applied the analysis of fragment size for cancer detection to see if any trends could be useful for elevating the diagnostic ability. Analysis of cell-free DNA size as a predictor of cancer state has not reached a consensus, with evidence suggesting both increases (Chan et al., 2008; Umetani et al., 2006; Wang et al., 2003) and decreases (Mouliere et al., 2011, 2018; Snyder et al., 2016; Underhill et al., 2016). Further studies have observed that cfDNA with tumor sequences tend to be highly fragmented (Mouliere et al., 2018) and enriching these fragments can improve mutated DNA signal (Figure 3.2A). Other non-somatic attributes, such as the nucleosome position patterns (Snyder et al., 2016; Ulz et al., 2016), and patterns near transcription sites (Esfahani et al., 2022; Ivanov et al., 2015) may be altered in cancer and be viable biomarkers.

“Fragmentomics” leverages analysis of the ratios of different cfDNA fragment lengths (Figure 3.2B) to generate risk and detection profiles for cancer patients. Previous studies, however, utilize NGS workflows that only assess fragments >100bp (Cristiano et al., 2019; Foda et al., 2022; Vessies et al., 2022). In this chapter, we can explore if the fragmentomic analysis of datasets generated by BRcfDNA-Seq, which reveal fragments of both uscfDNA and mncfDNA fragment region size, could provide different perspective fragment characteristics of cfDNA in cancer samples.

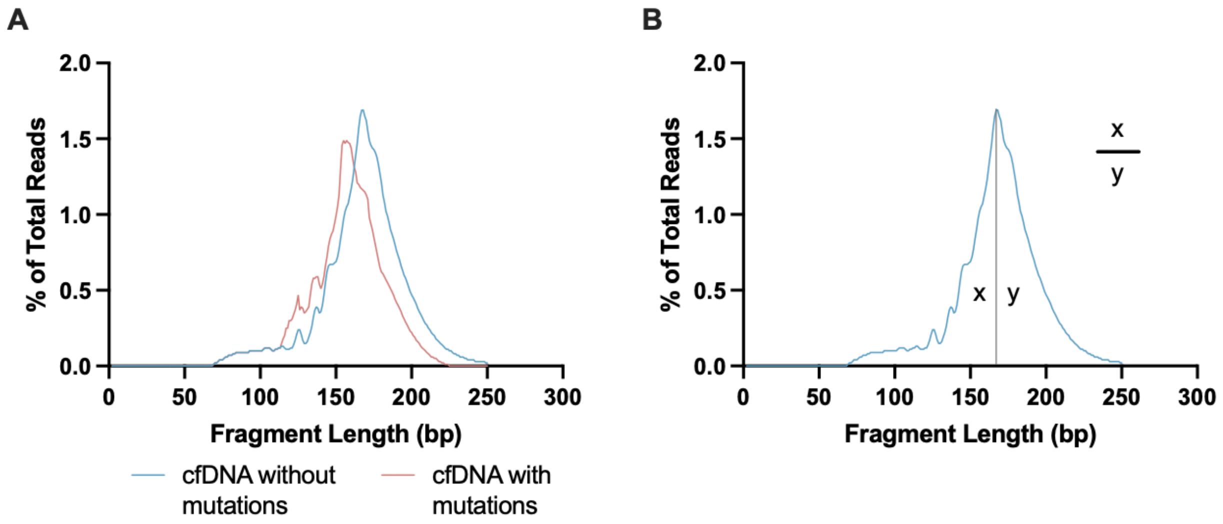

Figure 3.2 Principles of fragmentomic analysis. A) Cell-free DNA containing mutant sequences is reported to have a shorter fragment length size distribution compared to wildtype-derived cell-free DNA (Mouliere et al., 2018). B) Methodology for calculating fragmentation score: total fragments within region x is divided by total fragments within region y.

3.2.6 Cell-Free DNA End-motif Analysis

Beyond simply looking at the size distributions, the fragmentation nature of cell-free DNA provides insights into the original nucleosomal relationships and open chromatin domains (Snyder et al., 2016). These DNA fragments would be cleaved at accessible locations of the genome, and the resulting ending sites of the cell-free DNA are also not random. The ends contain over-represented DNA sequences, referred to as "preferred ends" (Jiang et al., 2018). This was initially demonstrated in noninvasive fetal genome analysis comparing fetal DNA to maternal DNA, showing that cfDNA of fetal DNA had a certain combination of base pairs at the ends of DNA fragments (Chan et al., 2016) which would be created after cleavage from specific enzymes. These fragment end sequences exhibited tissue specificity, which reflected both the structural orientation around the nucleosome and nuclease activity of the cell-free DNA when it was in its original form as intact genomic DNA.

Pre-existing end-motif studies showed that the 5′ end-motifs of plasma cell-free DNA preferentially begin with cytosine nucleotides (Chandrananda et al., 2015). The molecular basis is DNase1, and DNase1-like 3 (DNAse1L3) have been shown to be two major nucleases that provide effective clearance of DNA when released from dying cells (Napirei et al., 2009). Building on these facts, studies on the plasma from knock-out mice model studies revealed that end-motif patterns of cell-free DNA are indeed influenced by DNA nucleases (Serpas et al., 2019). In wild type mouse plasma, the top six 5-mer end-motifs (out of 256 possibilities) were "CCCA, CCTG, CCAG, CCAA, CCAT, and CCTC." In this study, when DNase-1 was knocked out, the mncfDNA fragment pattern nor did the motif ratios appear to change. DNase-1 nuclease, however, is designed to cut exposed and naked DNA, so the associated circulating nucleosome may be preventative. Instead, when another nuclease, DNAse1L3, which when knocked out, produced an elevation of fragments at 120bp but also dramatically influenced the proportion of "CC" ends in the plasma DNA end-motifs seen. Thus, DNase1L3 appears crucial in cfDNA fragmentation biology.

These findings catapulted the curiosity of end-motifs in cancer disease processes since many cancers demonstrate dysregulated DNase1l3 expression (J. Chen et al., 2021; Deng et al., 2021; Xiao et al., 2022). Other nucleases may behave differently during tumor states(Hernandez et al., 2021). A study examining the end-motifs in cancer patients showed hepatocellular carcinoma (HCC) subjects showed a preferential 4-mer end motif compared to those without HCC (Jiang et al., 2020). Furthermore, aberrations in end-motif profiles were observed in colorectal cancer, lung cancer, nasopharyngeal carcinoma, and head and neck squamous cell carcinoma (Jiang et al., 2020). Since then, other groups have also examined the cfDNA end-motifs (Jin et al., 2021; Zhitnyuk et al., 2022).

Analysis of end-motifs profiles of uscfDNA which could provide another biomarker source for cancer detection (Figure 3.3). The prior published analysis, however, was only focused and performed only for the mncfDNA fragments, and due to their use of a double-stranded library kit, it can only accurately examine the 5′ end. BRcfDNA-Seq uses a single-strand library preparation and does not require end-polishing; thus, both 5′ and 3′ ends are preserved. Therefore, this cfDNA feature was also chosen to be analyzed in this cancer cohort.

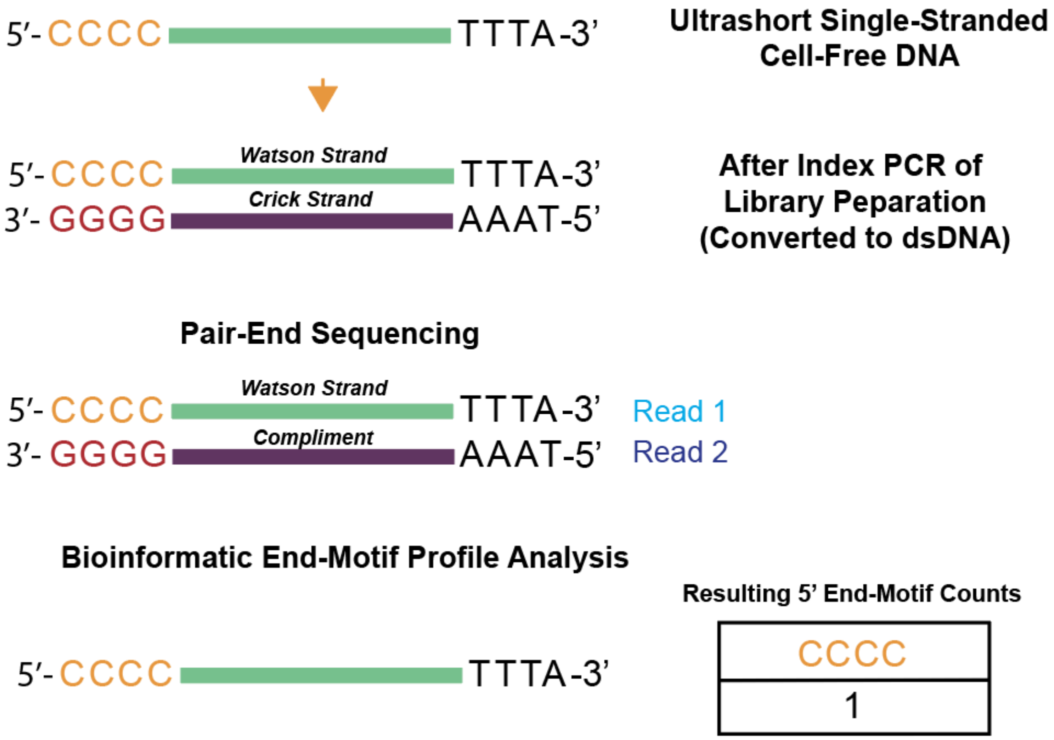

Figure 3.3 End-motif analysis. The first four base pairs from the 5′ end is considered when tabulating end-motif frequencies.

3.2.7 G-Quadruplex Signature Patterns

As demonstrated by the extensive field of fragmentomics and end-motifs, the sequence composition and topology of cfDNA is unlikely to be random. This non-randomness applies to the uscfDNA population as well. Other research groups have reported that uscfDNA population samples of healthy individuals had a higher nucleotide GC% in their uscfDNA than cancer subjects (Hudecova et al., 2021). Further examination of the sequences of uscfDNA revealed that the sequences contained G-Quadruplex signatures which were predictable since in vitro (Figure 3.4A), single-stranded DNA sequences containing a high density of guanine nucleotides are prone to forming secondary structures called G-Quadruplexes (G-Quad) (Figure 3.4B and C) (Hänsel-Hertsch et al., 2017; Varshney et al., 2020). G-Quadruplex structures are observable inaccessible chromatin regions of the genome, and certain subsets have been highly correlated to expression levels of genes in cancer cells and tumor tissue (Hänsel-Hertsch et al., 2016, 2020). Due to their involvement in genome instability, G-Quad structures are being explored as a target for therapy(Kosiol et al., 2021). Therefore, exploring this metric could be a useful metric in this clinical context.

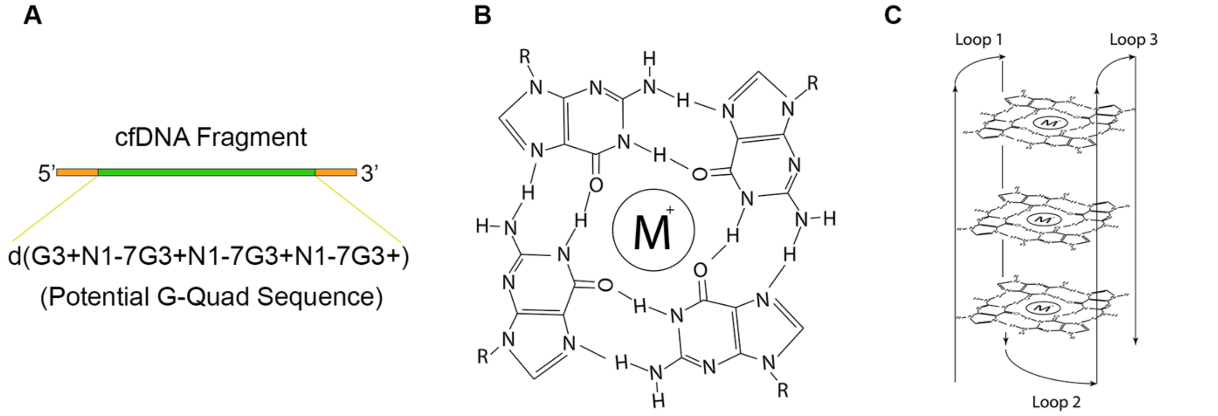

Figure 3.4 G-Quadruplex abundance analysis. A) Formula for sequences contained within cfDNA that suggest potential G-Quad structure (Todd et al., 2005). B) G-tetrad formation. C) Secondary G-Quad complex structure that can potentially form within a cfDNA molecule (Capra et al., 2010)

3.2.8 Chapter Goals

In this chapter, we first bioinformatically analyzed uscfDNA and mncfDNA populations generated from BRcfDNA-Seq to determine any major differences in their genomic characteristics. Next, we compared the non-cancer samples with the cfDNA from plasma from a cohort of late-stage non-small cell lung carcinoma (NSCLC) patients. Our goal was to determine if this uscfDNA-based analysis could reveal significant differences in patterns in relation to functional element peaks, fragmentomics, end-motif sequences, or G-Quadruplex secondary structures between the cfDNA of these two cohorts (Figure 3.5B).

3.3 Results

3.3.1 Characteristics of uscfDNA are Distinct from mncfDNA

Building off Chapter 2, we examined the differences between the uscfDNA and mncfDNA populations in non-cancer subjects using the BRcfDNA-Seq NGS pipeline (Figure 3.5A and B). Karyograms of the normalized coverage of uscfDNA and mncfDNA populations showed significantly different coverage patterns in 941 genomic bins (Figure 3.5 C and D) (q = 0.0004 to 0.01) (bins with increased coverage density are redder while lower coverage density is bluer). uscfDNA were mapped to more hotspots within the body of chromosomes and telomeres than the mncfDNA(Figure 3.5 C). Analysis of the ratio of mapped peaks to total reads using MACS2 (Y. Zhang et al., 2008) reveals that uscfDNA reads have a 45.2-fold increase in aligned peaks than mncfDNA (Figure 3.5E). Determination of the categories of genomic loci associated with the peaks indicated that uscfDNA was highly enriched in the promoter, introns, and exons (Figure 3.5F).

We examined the first four nucleotides at the 5' end of reads and measured motif frequency differences between uscfDNA and mncfDNA of the 256 possible combinations (Jiang et al., 2020). Multiple paired t-test comparisons revealed that 211/256 end-motifs had significantly different frequencies (0.000001 to 0.009) between uscfDNA and mncfDNA populations (Figure 3.5G). For mncfDNA, we observed 4 out of the top 6 matched the top 6 most prevalent motifs reported in the literature (Serpas et al., 2019) (CCCA, CCTG, CCAG, and CCTC (Table 3.1). For the uscfDNA population, only 2 out of the top 6 (CCCA and CCAG) matched the top 6 motifs previously reported.

Lastly, we examined the prevalence of G-Quad signatures and observed that uscfDNA fragments had a 4.9-fold greater abundance than mncfDNA (Figure 3.5H).

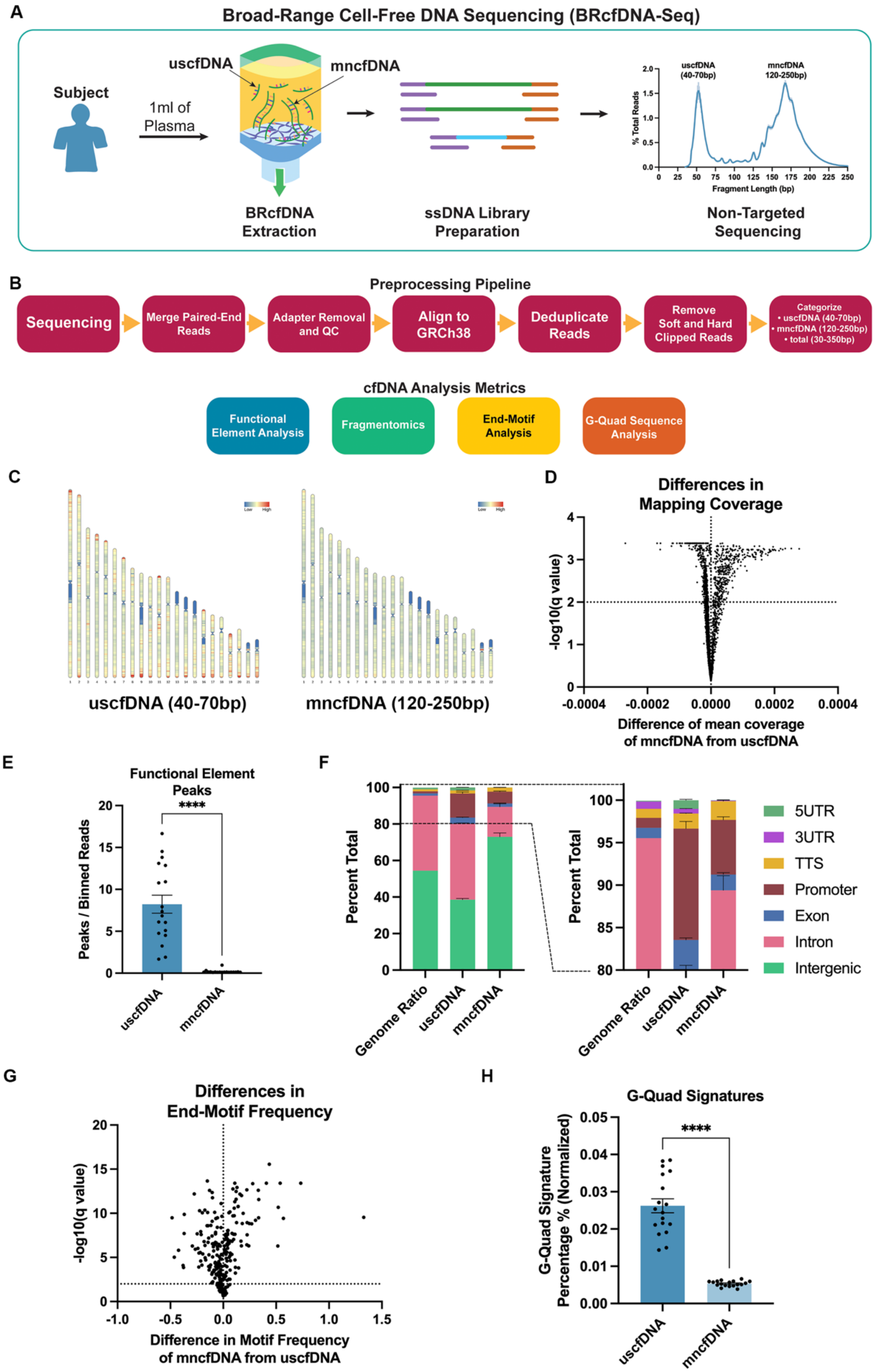
A
Broad-Range Cell-Free DNA Sequencing (BRcfDNA-Seq)
Subject
1ml of Plasma
uscfDNA
mncfDNA
BRcfDNA Extraction
ssDNA Library Preparation
uscfDNA (40-70bp)
mncfDNA 120-250bp)
% Total Reads
2.0
1.5
1.0
0.5
0.0
0 25 50 75 100 125 150 175 200 225 250
Fragment Length (bp)
Non-Targeted Sequencing
B
Preprocessing Pipeline
Sequencing
Merge Paired-End Reads
Adapter Removal and QC
Align to GRCh38
Deduplicate Reads
Remove Soft and Hard Clipped Reads
Categorize
• uscfDNA (40-70bp)
• mncfDNA (120-250bp)
• total (30-350bp)
cfDNA Analysis Metrics
Functional Element Analysis
Fragmentomics
End-Motif Analysis
G-Quad Sequence Analysis
C
Low High
uscfDNA (40-70bp)
mncfDNA (120-250bp)
D
Differences in Mapping Coverage
-log10(q value)
4
3
2
1
0
-0.0004
-0.0002
0.0000
0.0002
0.0004
Difference of mean coverage of mncfDNA from uscfDNA
E
Functional Element Peaks
****
Peaks / Binned Reads
20
15
10
5
0
uscfDNA
mncfDNA
F
Percent Total
100
80
60
40
20
0
Genome Ratio
uscfDNA
mncfDNA
Percent Total
100
95
90
85
80
Genome Ratio
uscfDNA
mncfDNA
5UTR
3UTR
TTS
Promoter
Exon
Intron
Intergenic
G
Differences in End-Motif Frequency
-log10(q value)
20
15
10
5
0
-1.0
-0.5
0.0
0.5
1.0
1.5
Difference in Motif Frequency of mncfDNA from uscfDNA
H
G-Quad Signatures
****
G-Quad Signature Percentage % (Normalized)
0.05
0.04
0.03
0.02
0.01
0.00
uscfDNA
mncfDNA

Figure 3.5 Characteristics of uscfDNA differentiate it from mncfDNA. A) Schematic of BRcfDNA-Seq reveals uscfDNA (40-70bp) in conjunction with mncfDNA (120-250bp peak) in plasma. Plasma extracted from non-cancer individuals using size-agnostic extraction coupled with single-stranded library preparation. B) Bioinformatic workflow preprocessing prior to analyzing four cfDNA features. C) Karyograms of averaged normalized coverage plots showing differences in mapping for uscfDNA and mncfDNA populations along every 1 million bp bin across the genome (bins with increased coverage density are redder while lower coverage density is bluer). Karyograms are self-normalized so that the legend reflects the intrasample dynamic range. D) Volcano plot showing that 941 genomic bins had significantly different coverage between uscfDNA and mncfDNA populations. E) Ratio of functional peaks determined by MACS2 per total reads reveal that uscfDNA reads inherently have more peaks than mncfDNA. F) Proportion of functional elements categories of the peaks are different between uscfDNA and mncfDNA. G) Volcano plot showing 211 5’-end-motifs demonstrated significant differences in frequency between uscfDNA and mncfDNA populations. H) G-Quad signatures are greatly enriched in the uscfDNA population. Student’s t-test was performed with Welch’s correction. Multiple paired t-tests were performed with a false discovery rate of 1% using the two-stage step-up method of Benjamini, Krieger, and Yekutieli. Error bars represent SEM. Stars indicate adjusted q-values are presented with * p <0.05, ** p <0.01, *** p< 0.001, and **** p <0.0001.

Table 3.1 Comparison of the six most abundant 5’ end 4-mer motifs between published reports and the uscfDNA and mncfDNA of non-cancer and NSCLC.

	Most Abundant 5’End-motifs			
Serpas et al 2018 (mncfDNA)	uscfDNA (Non-Cancer)	mncfDNA (Non-Cancer	uscfDNA (NSCLC)	mncfDNA (NSCLC)
CCCA	CCCC	CCCA	CCCC	CCCA
CCTG	CCAG	CCTT	CCAG	CCTG
CCAG	CAAA	CCTG	CCCA	CCAG
CCAA	CCCA	CCCT	CCCT	CCCT
CCAT	AAAA	CCAG	CAAA	CCAA
CCTC	CCCT	CCTC	CCAA	TGGA

3.3.2 uscfDNA and mncfDNA Fragments Map to Different Positions

We hypothesized that these unique characteristics of uscfDNA could be useful as biomarkers for cancer detection. Thus, we measured these features in the NSCLC cohort (Figure 3.6). Compared to non-cancer, the NSCLC uscfDNA population presented a coverage pattern with more hotspots (Figure 3.6A), resulting in 1764 significantly enriched bins. For the mncfDNA bins, no significantly different bins were found (Figure 3.6B).

3.3.3 Functional Element Peak Profiles of uscfDNA are Altered in NSCLC

Since uscfDNA was associated with a high peak abundance in the non-cancer plasma (Figure 3.5E), we examined if this observation was consistent in NSCLC samples. Again, uscfDNA fragments were associated with more peaks than mncfDNA (Figure 3.6C). Interestingly, for uscfDNA, the NSCLC samples trended toward a decrease in total peaks.

We categorized the peaks into select genomic regions to observe if the expected peak profiles changed in NSCLC subjects (Figure 3.6D). For uscfDNA, there was a significant decrease in observed/expected peak count for transcription termination site (TTS), exonic, intronic, intergenic, promoter, and 5’UTR peaks. By contrast, for mncfDNA, there was only a decrease in expected peaks in promoters (Figure 3.6D).

Considering uscfDNA functional element peak profiles were altered in NSCLC samples (Figure 3.6D), we further examined which specific sequences were changing in the promoters, introns, and exons categories. The top 20 most prevalent sequences between non-cancer and NSCLC cohorts were documented (Figure 3.6 E-G). We developed a “Peak Score” to assign a relative contribution score for each peak and assembled a panel of peaks that demonstrated significant differentiation in scores between cohorts (Figure 3.6H-J). From the total list, the top functional peaks were derived from all three categories (Figure 3.6I). We observed that compared to non-cancer, NSCLC was associated with 13 candidate uscfDNA functional elements (q = <0.000001 to 0.01, non-paired t-test) that collectively increased or decreased (eg. HAR1B, SMYD3, NIKX6 (Figure 3.6J). A similar analysis was performed for the mncfDNA bin, but no significant peaks were discovered.

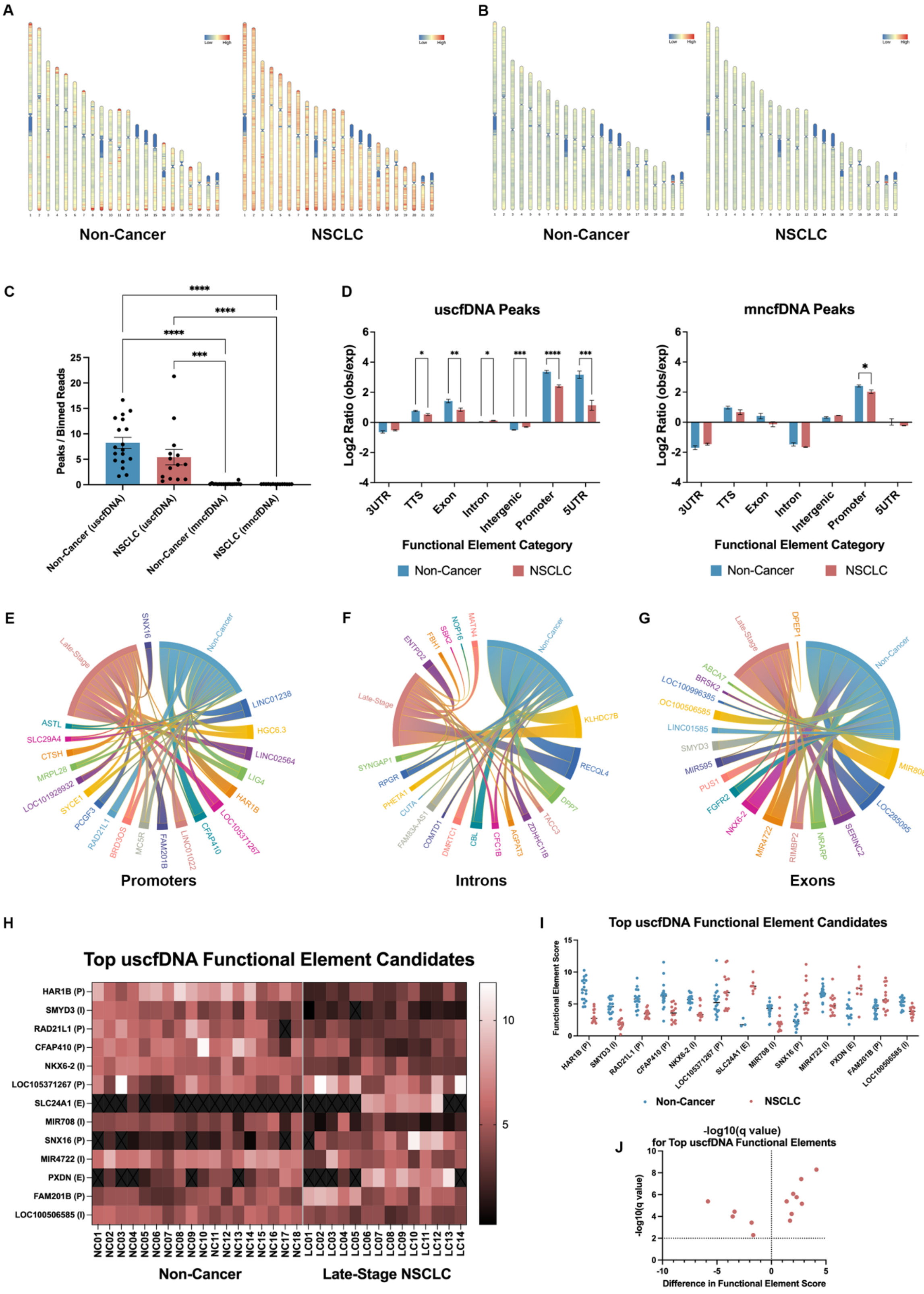
A
Low
High
Non-Cancer
NSCLC
B
Non-Cancer
NSCLC
C
Peaks / Binned Reads
0
5
10
15
20
25
****
****
****
***
Non-Cancer (uscfDNA)
NSCLC (uscfDNA)
Non-Cancer (mncfDNA)
NSCLC (mncfDNA)
D
uscfDNA Peaks
Log2 Ratio (obs/exp)
6
4
2
0
-2
-4
*
**
*
***
****
***
3UTR
TTS
Exon
Intron
Intergenic
Promoter
5UTR
Functional Element Category
Non-Cancer
NSCLC
mncfDNA Peaks
Log2 Ratio (obs/exp)
*
3UTR
TTS
Exon
Intron
Intergenic
Promoter
5UTR
Functional Element Category
Non-Cancer
NSCLC
E
Late-Stage
SNX16
Non-Cancer
LINC01238
HGC6.3
LINC02564
LIG4
HAR1B
LOC105371267
CFAP410
LINC01022
FAM201B
MC5R
BRD3OS
RAD21L1
PCGF3
SYCE1
LOC101928932
MRPL28
CTSH
SLC29A4
ASTL
Promoters
F
ENTPD2
FBH1
SBK2
NOP16
MATN4
Non-Cancer
Late-Stage
KLHDC7B
RECQL4
DPP7
TACC3
ZDHHC11B
AGPAT3
CFC1B
CBL
DMRTC1
COMTD1
FAM83A-AS1
CUTA
PHETA1
RPGR
SYNGAP1
Introns
G
Late-Stage
DPEP1
Non-Cancer
ABCA7
BRSK2
LOC100996385
OC100506585
LINC01585
SMYD3
MIR595
PUS1
FGFR2
NKX6-2
MIR4722
RIMBP2
NRARP
SERINC2
LOC285095
MIR8089
Exons
H
Top uscfDNA Functional Element Candidates
HAR1B (P)
SMYD3 (I)
RAD21L1 (P)
CFAP410 (P)
NKX6-2 (I)
LOC105371267 (P)
SLC24A1 (E)
MIR708 (I)
SNX16 (P)
MIR4722 (I)
PXDN (E)
FAM201B (P)
LOC100506585 (I)
10
5
NC01
NC02
NC03
NC04
NC05
NC06
NC07
NC08
NC09
NC10
NC11
NC12
NC13
NC14
NC15
NC16
NC17
NC18
LC01
LC02
LC03
LC04
LC05
LC06
LC07
LC08
LC09
LC10
LC11
LC12
LC13
LC14
Non-Cancer
Late-Stage NSCLC
I
Top uscfDNA Functional Element Candidates
Functional Element Score
15
10
5
0
HAR1B (P)
SMYD3 (I)
RAD21L1 (P)
CFAP410 (P)
NKX6-2 (I)
LOC105371267 (P)
SLC24A1 (E)
MIR708 (I)
SNX16 (P)
MIR4722 (I)
PXDN (E)
FAM201B (P)
LOC100506585 (I)
Non-Cancer
NSCLC
J
-log10(q value)
for Top uscfDNA Functional Elements
-log10(q value)
10
8
6
4
2
0
-10
-5
0
5
Difference in Functional Element Score

Figure 3.6 Functional Element Peaks of uscfDNA can differentiate non-cancer and NSCLC cohorts. Averaged normalized coverage karyogram plots showing mapping positions for representing one million bp bins across the genome between non-cancer and NSCLC for the uscfDNA (A) and mncfDNA (B) populations (bins with increased coverage density are redder while lower coverage density bluer). Karyograms are self-normalized so that the legend reflects the intrasample dynamic range. (C) The ratio of functional peaks determined by MACS2 per total reads revealed in the uscfDNA population there is an observable decrease in peaks in NSCLC subjects compared to non-cancer ones. D) Log2 ratio of observed vs. expected number of peaks of various functional peak categories for uscfDNA and mncfDNA show alterations in NSCLC state. Non-paired multiple t-tests with Holm-Šidák correction with alpha at 0.05 was used. Mutual and non-mutual peak identities between NSCLC and non-cancer cohorts for promoter (E), intronic (F), and exonic (G) elements are collated in chord diagrams. H), A heat map of top differentiating functional peak candidates between the non-cancer and NSCLC samples. The peak score is plotted for each sample and element and was discovered using non-paired multiple t-tests for each function element peak with a false discovery rate of 1% using the two-stage step-up method of Benjamini, Krieger, and Yekutieli. I) Individual peak score of top differentiating functional peaks between non-cancer and NSCLC (all represent q value of <0.05. J) Volcano plot demonstrating positive or negative changes in peak score of NSCLC compared to non-cancer samples. Letters in parentheses beside identities represent each element type (P: promoter | I: Intron | E: exon). Analysis was conducted with non-cancer n = 18 and NSCLC n= 14. Error bars represent SEM. Stars indicate adjusted q-values are presented with * q <0.05, ** q 0.01, *** q< 0.001, and **** q <0.0001.

3.3.4 NSCLC cfDNA has Increased Fragmentation

Next, we analyzed if the size distribution profiles of cfDNA appeared different between the non-cancer and NSCLC cohorts (Figure 3.7A). The uscfDNA peak (~50bp) appeared elevated in NSCLC compared to non-cancer. For the mncfDNA region, the distribution between the two cohorts was more distinct, with the NSCLC samples having a lower "shoulder" at 175 bp. The ratio between uscfDNA reads (40-70bp), and mncfDNA reads (120-250bp) was elevated in NSCLC samples (Figure 3.7B).

Fragmentation scores (method shown in 3.2B) revealed that in NSCLC subjects, the cfDNA is more fragmented (Figure 3.7C and D). Next, binning by genomic location for every 1 million reads showed that all positions were more fragmented in the NSCLC samples considering uscfDNA (Figure 3.7E) and mncfDNA (Figure 3.7F). There were specific bins where both uscfDNA and mncfDNA demonstrated highly significant differences in fragmentation

(2784/2874 uscfDNA candidates | q = 0.00005 to 0.00041 and 2784/2784 mncfDNA candidates | q = 0.00057 to 0.0077, non-paired multiple t-tests) (Figure 3.7G and H).

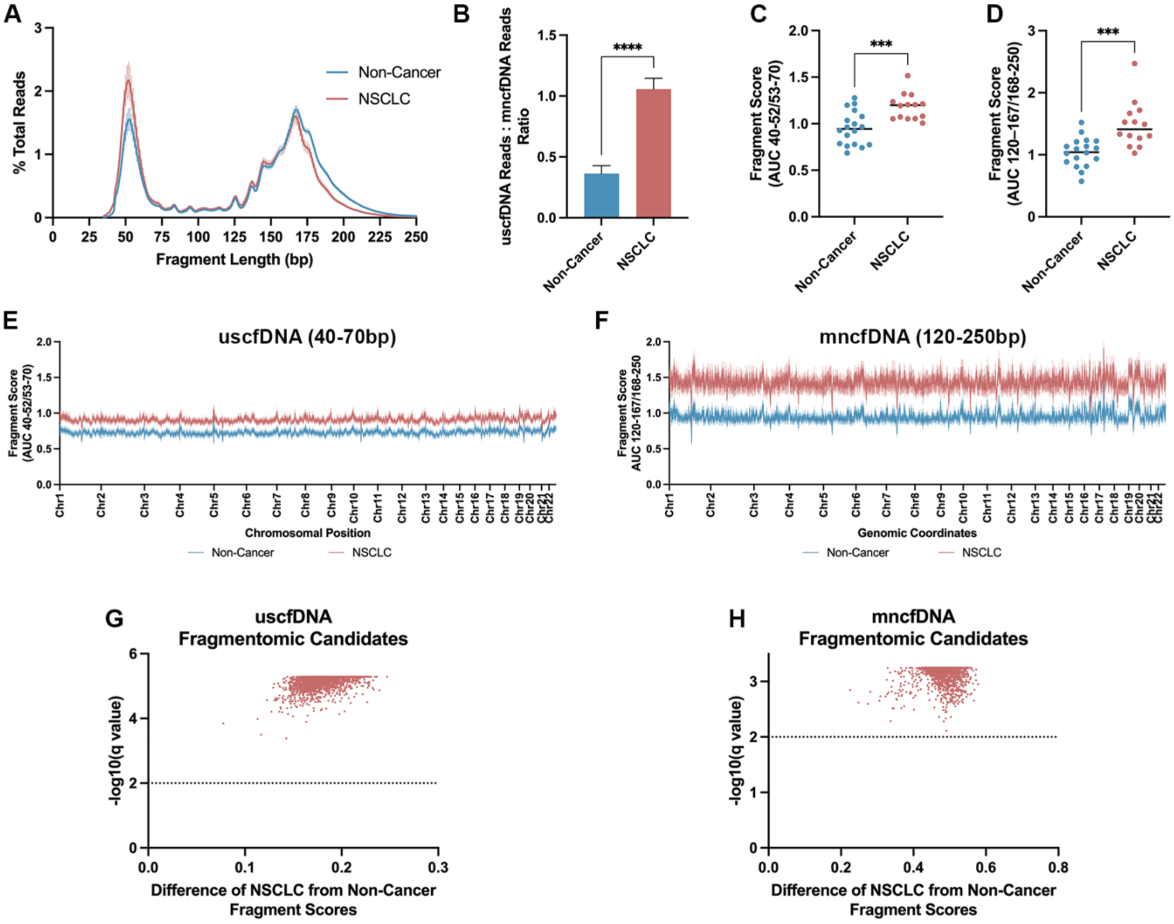

Figure 3.7 Fragmentomic Analysis of uscfDNA and mncfDNA differentiates non-Cancer from NSCLC Plasma Samples. A) Fragment size distribution profiles between non-cancer and NSCLC samples. B) the Ratio of reads/fragments within the uscfDNA bin and mncfDNA bin demonstrates in NSCLC samples is associated with an increase in uscfDNA read proportion. Higher global fragment score indicates both uscfDNA C) and mncfDNA (D) demonstrate increased fragmentation in NSCLC samples. Fragment scores were calculated and plotted for each one million bins across the genome, showing a higher resolution view that the NSCLC individuals are more fragmented than non-cancer individuals (E and F). Volcano plots summarizing multiple non-paired t-tests with a false discovery rate of 1% using the two-stage step-up method of Benjamini, Krieger, and Yekutieli for each genomic bin revealed significant regions in uscfDNA and mncfDNA populations (G and H). Data represents the average of 18 non-cancer individuals and 14 NSCLC patients. Error bars and vertical section lines represent SEM between samples. Stars indicate p-values are presented with ** p <0.01, *** p< 0.001, and **** p <0.0001 from unpaired Student t-test using Welch’s correction.

3.3.5 End-motif Profile Differs Between uscfDNA and mncfDNA

Previous reports have suggested that plasma end-motif diversity becomes more random due to the dysregulation of nucleases (Jiang et al., 2020). For both uscfDNA and mncfDNA populations, compared to non-cancer, NSCLC samples trended towards an increased Motif Diversity Score (more random), although only mncfDNA was significant (Figure 3.8 A and B).

Next, we interrogated which four base pair end-motifs were most differentiable between non-cancer and NSCLC samples. For the uscfDNA population, 127/256 (q = <0.000001 to 0.0099) end-motifs demonstrated significant distinction between the two cohorts (Figure 4A) compared with only 119/256 (q = <0.000003 to 0.0095) end-motifs candidates for mncfDNA (Figure 3.8C). Interestingly, the top six differentiating end-motifs were different from the most prevalent end-motifs previously reported (Serpas et al., 2019) and were distinct between the two cfDNA populations (Figures 3.8 B and D). For samples analyzed in this study, the most common top 6 end-motif between uscfDNA or mncfDNA of non-cancer and NSCLC was CCCT (Table 3.1).

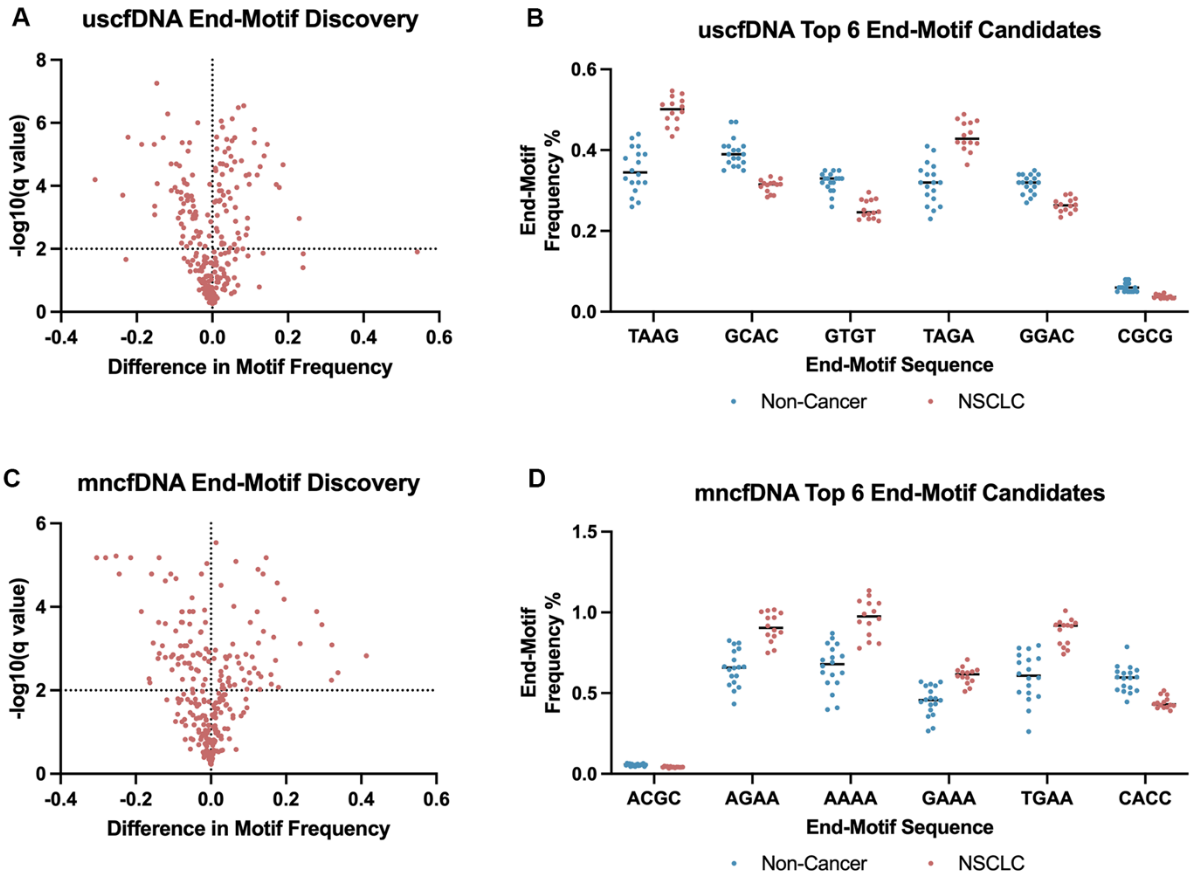

Figure 3.8 BRcfDNA-Seq reveals candidate end-motifs between non-cancer and NSCLC samples. Discovery using non-paired multiple t-tests for each 256-possible end-motif revealed significantly different end-motifs between non-cancer and NSCLC for uscfDNA (A) (127/256 with q-value >0.05) and mncfDNA (C) (119/256 with q-value >0.05) populations. Six of the most differentiable end-motifs (all are q-value >0.05) for uscfDNA (B) and mncfDNA (D) from the discovery are plotted, demonstrating their motif-frequency % changes between non-cancer and NSCLC. Non-paired multiple t-tests using a false discovery rate of 1% using the two-stage step-up of Benjamini, Krieger, and Yekutieli was used for discovery.

3.3.6 G-Quad Signatures are Decreased in NSCLC Samples

We identified the presence of G-Quad-containing signatures in both uscfDNA and mncfDNA populations aligned to exons, introns, and promoter regions in the genome (Figure 3.9A). Compared to non-cancer samples, all introns, exons, and promoter regions had a significant decrease in G-Quad signatures. Additionally, the analysis proportion of primary fragments versus theoretical complementary fragments containing G-Quad sequences appeared equal (Figure 3.9B and C).

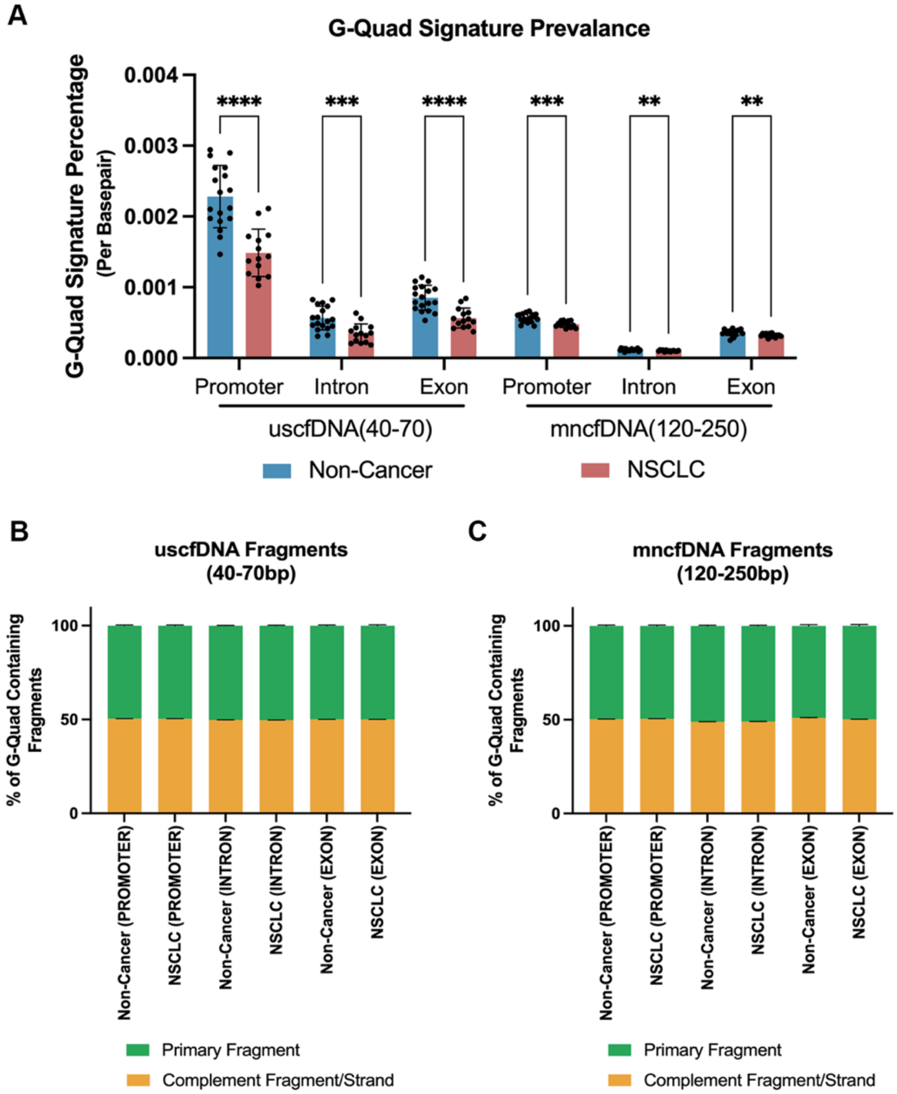

Figure 3.9 G-Quadruplex (G-Quad) signatures in the sequences of uscfDNA and mncfDNA populations are decreased in NSCLC donors compared to non-cancer individuals. A) Presence of G-Quad signatures normalized percentage (fragments with G-Quad presence / total fragments) was calculated for uscfDNA and mncfDNA fragments that aligned promoter, intronic, and exonic loci. Signature counts were normalized by dividing by the average bp (uscfDNA:50 | mncfDNA: 167). Non-paired multiple t-tests with Holm-Šidák correction with alpha at 0.05 was used. p values : **, ***,**** is <0.01, <0.001, and <0.0001. The proportion of primary uscfDNA fragment/strand to theoretical uscfDNA complement strand that contains potential G-Quad signatures for uscfDNA (B) and mncfDNA (C). Error bars indicate SEM.

3.3.7 Integration of Multiple cfDNA Biomarkers Provides Differentiation Between Non-cancer and NSCLC

We then incorporate all previously statistically significant cfDNA biomarker features from each category (Fragmentomics, Functional Element, End-Motif, and G-Quad Signature) into a principal component analysis (PCA) analysis which showed that principal components 1 and 2 (PCA1 and PCA2) could clearly separate non-cancer and NSCLC samples using both uscfDNA (Figure 3.10A) and mncfDNA (Figure 3.10B). An unsupervised clustering heatmap showed the best-performing cfDNA features which differentiate non-cancer and NSCLC plasma samples (Figure 3.10C and D). The compressed significant biomarkers into separate PCA components reveal that the first five principal components of both uscfDNA and mncfDNA have a cumulative explained variance of >80% (Figure 3.10E and F). For uscfDNA and mncfDNA, PCA1 values from non-cancer and NSCLC cohorts were significantly different (p-value <0.0001).

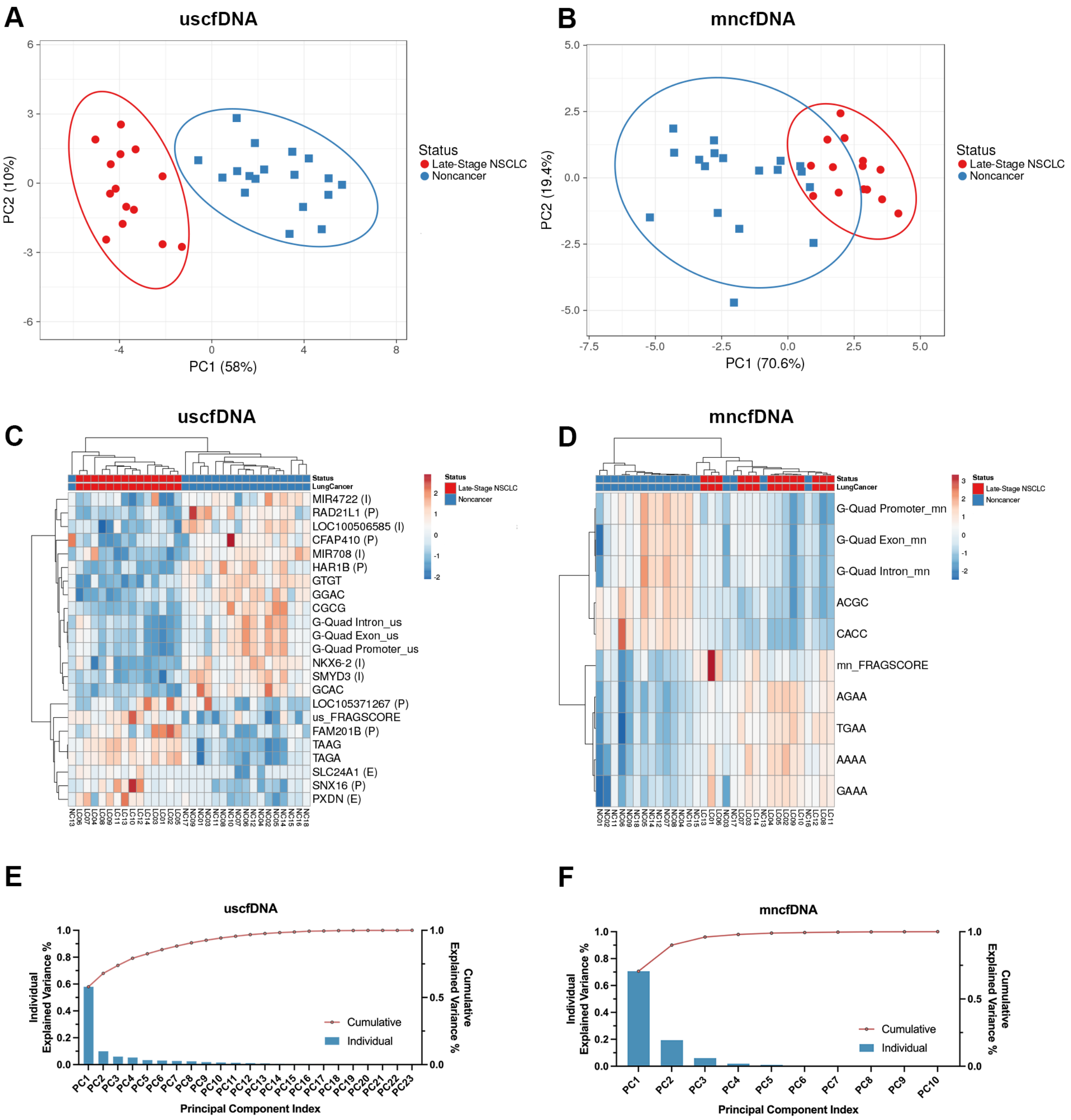
A
uscfDNA
PC2 (10%)
PC1 (58%)
6
3
0
-3
-6
-4
0
4
8
Status
Late-Stage NSCLC
Noncancer
B
mncfDNA
PC2 (19.4%)
PC1 (70.6%)
5.0
2.5
0.0
-2.5
-5.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
Status
Late-Stage NSCLC
Noncancer
C
uscfDNA
Status
LungCancer
MIR4722 (I)
RAD21L1 (P)
LOC100506585 (I)
CFAP410 (P)
MIR708 (I)
HAR1B (P)
GTGT
GGAC
CGCG
G-Quad Intron_us
G-Quad Exon_us
G-Quad Promoter_us
NKX6-2 (I)
SMYD3 (I)
GCAC
LOC105371267 (P)
us_FRAGSCORE
FAM201B (P)
TAAG
TAGA
SLC24A1 (E)
SNX16 (P)
PXDN (E)
NC13 LC06 LC07 LC04 LC08 LC09 LC11 LC13 LC10 LC12 LC14 LC03 LC01 LC02 LC05 NC17 NC09 NC01 NC03 NC11 NC08 NC10 NC07 NC06 NC12 NC04 NC02 NC05 NC14 NC15 NC16 NC18
Status
Late-Stage NSCLC
Noncancer
2
1
0
-1
-2
D
mncfDNA
Status
LungCancer
G-Quad Promoter_mn
G-Quad Exon_mn
G-Quad Intron_mn
ACGC
CACC
mn_FRAGSCORE
AGAA
TGAA
AAAA
GAAA
NC01 NC02 NC11 NC06 NC09 NC18 NC05 NC14 NC12 NC07 NC08 NC04 NC10 NC15 LC13 LC01 LC06 NC03 NC17 LC07 LC03 LC14 NC13 LC04 LC05 LC02 LC09 LC10 NC16 LC12 LC08 LC11
Status
Late-Stage NSCLC
Noncancer
3
2
1
0
-1
-2
E
uscfDNA
Individual
Explained Variance %
Cumulative
Explained Variance %
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.5
0.0
Cumulative
Individual
PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11 PC12 PC13 PC14 PC15 PC16 PC17 PC18 PC19 PC20 PC21 PC22 PC23
Principal Component Index
F
mncfDNA
Individual
Explained Variance %
Cumulative
Explained Variance %
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.5
0.0
Cumulative
Individual
PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10
Principal Component Index

Figure 3.10 Integration of select significant biomarkers under the four cfDNA domains (Fragmentomics, Functional Elements, End-Motif, and G-Quad) depicts the separation between cohorts. Non-cancer is represented as blue and NSCLC as red. Principal component analysis (PCA) scores were calculated and plotted for the most significant biomarkers for uscfDNA (A) and mncfDNA (B) populations. Unit variance scaling is applied to rows; SVD with imputation is used to calculate principal components. X and Y axis show principal component 1 and principal component 2. This analysis reveals that for uscfDNA, 58%(PC1) and 10%(PC2) and for mncfDNA, 70.6%(PC1) and 19.5% (PC2) explains the total variance. Prediction ellipses are such that with a probability of 0.95, a new observation from the same group will fall inside the ellipse (N = 32 data points). Unsupervised clustering heatmap shows biomarker categories most discriminatory for uscfDNA (C) and mncfDNA (D). Rows are centered; unit variance scaling is applied to rows. Imputation is used for missing value estimation. Both rows and columns are clustered using correlation distance and average linkage. Plots of individual vs. cumulative explained variance shows the contribution of individual PCA categories for uscfDNA (E) and mncfDNA (F).

3.4 Discussion

In this chapter, using BRcfDNA-Seq, we illustrated that functional peak formation, G-Quad signature prevalence, and end-motif frequencies inherently differ between plasma uscfDNA and mncfDNA (Figure 3.5).   Furthermore, we showcase features of plasma uscfDNA that have the potential to be used as new biomarkers for cancer detection. As a proof of concept, we examined and compared features in both uscfDNA and mncfDNA for their ability to differentiate non-cancer from late-stage NSCLC subjects. Of the four features of cfDNA that we analyzed, we observed that functional element peaks (Figure 3.5E) and G-Quad signatures (Figure 3.5H) were unique characteristics of uscfDNA that are not associated with mncfDNA. The top 6 differentiating end-motif uscfDNA candidates differed from mncfDNA (Figure 3.5G). These features hint that uscfDNA may be derived from separate biological processes and justifies its examination as an independent biomarker type.

The presence of uscfDNA introduces new potential biological insights in cfDNA biology. In literature, the functions of RNA, a prominent single-stranded entity, are well described. RNA is involved in transcription, amino-acid transfer, protein complexes, gene expression, and signal transfer via exosomes. By comparison, circulating ssDNA biology has been largely unexplored, and it is plausible that ssDNA may have more functions than initially thought. In molecular biology, there is limited technology to evaluate ssDNA. With the development of BRcfDNA-seq, future studies interested in assessing ultrashort single-stranded DNA molecules are now possible. In this regard, there is merit in exploring how uscfDNA plays a role in normal physiology and how it may change with age compared to the mncfDNA population (Teo et al., 2019).

Based on the data presented here, uscfDNA does not necessarily appear to be involved in the cell death pathways for the disposal of genomic DNA. Extensive literature has described the origins of mncfDNA as a byproduct of genomic DNA degradation (Burnham et al., 2016; Nagata et al., 2003). Based on our observations, the genomic coverage of uscfDNA does not map evenly amongst the chromosomes in the genome compared to mncfDNA (Figures 3.6 C and D). It is unclear why uscfDNA fragments inherently coalesce into specific peaks at a higher prevalence than mncfDNA (Figure 3.5E). The enriched presence in the promoter, exon, and intronic peaks may reflect changes in nucleosome positioning in regions of the genome involved in high transcriptional activity (Ivanov et al., 2015). Dependent on the nucleosome and DNA interplay, cell states regulate which genomic regions are susceptible to DNA fragmentation by nucleases. Regarding specific element identity, several of the most distinct peaks that exhibit changes in proportion between non-cancer and NSCLC have previously been described to be associated with cancer states. For example, the HAR1B promoter regulates a long non-coding RNA used as a biomarker in bone and soft-tissue sarcomas (Yamada et al., 2021). The CFAP410 gene (also known as C21orf2) encodes a ciliary protein involved in cilia formation and DNA repair (Fang et al., 2015; Shin et al., 2020). SNX16 has been described with both pro and anti-tumor activity (Shen et al., 2020; L. Zhang et al., 2013). Therefore, identifying the dynamic nature of these uscfDNA peaks may support their use as a biomarker.

The observed enrichment in the G-Quad signature of uscfDNA suggests an additional mechanism. During transcription, nucleic acid structures composed of RNA-DNA hybrids accompanying displaced single-stranded DNA (Brambati et al., 2020) are formed as R-loops. Within the R-loop complex, the transient single-stranded DNA can be configured into G-Quad secondary structures to aid strand separation. Within cells, RNA-DNA hybrids have been reported to accumulate in the cytoplasm after R-loop processing (Bhatia et al., 2014). Unscheduled or aberrant R-loop homeostasis can contribute to cancer phenotypes. Interestingly, in our data, we observe an equal proportion of primary fragments that contain G-Quad sequences to theoretical complementary fragments that contain G-Quad sequences (Figures 3.9B and C). This suggests that if uscfDNA is derived from an R-loop complex, it could either originate from the displaced strand or the DNA of the RNA-DNA hybrid. In the plasma, instead of enrichment, we observed a decrease in G-Quad signatures in the cfDNA (in particular promoter sequences matching a previous report) (Figure 3.9A) (Hudecova et al., 2021). The absence of G-Quad structures in circulation could reflect impaired R-loop processing and compromised G-Quad ejection resulting in the accumulation of G-Quad signatures in the cytoplasm of tumor cells (Brambati et al., 2020).

Along these lines, although not yet described in eukaryotes, the bacteria genome contains "retrons" sequences which code for a special type of reverse transcriptase and a non-coding RNA sequence to generate DNA/RNA hybrid called multicopy single-stranded DNA (msDNA) (Inouye & Inouye, 1993; Schubert et al., 2021). The retron ssDNA is considered part of the bacterial immune system and helps detect invading viruses (Millman et al., 2020). Some msDNA have been described to be as short as 48nt, so it is conceivable that an eukaryotic version may contribute to the uscfDNA pool in plasma where the RNA component has already degraded (Mao et al., 1997).

It has long been observed that in late-stage cancer, not only does the concentration of cell-free DNA increase, but the average fragment length can also decrease by 10-20bp (Lapin et al., 2018). Mutation containing cell-free DNA from tumors is consistently shorter than wild type DNA, and this skewed impression fragment size in late-stage cancer is likely due to the increased ratio of cancer cells undergoing apoptosis (Mouliere et al., 2018). These previous studies, however, only utilize extraction and DNA-quantification methods that consider the double-stranded mncfDNA population. Whether this observed pattern in late-stage cancer donors is mirrored by uscfDNA is not clear. Conversely, a study on cfDNA from pancreatic patient plasma using single-stranded library preparation (extracted with the equivalent of QiaC) showed that earlier stages are associated with shorter fragments (X. Liu et al., 2019). This apparent contradiction may hint that size profiles and concentrations of these two populations of cfDNA may have contrasting trajectories between the healthy, early-stage, and late-stage cancer phases.

To this end, the global analysis of uscfDNA fragmentomics (Figure 3.7) and end-motifs (Figure 3.8) could differentiate the two cohorts. The visual size-distribution changes in the proportion of uscfDNA in the fragment profile of NSCLC samples (Figure 3.7A) were reflected in the quantification by the uscfDNA: mncfDNA reads analysis (Figure 3.7B). This result contrasted with a previous report that uscfDNA abundance decreases in samples with greater ctDNA burden (Hudecova et al., 2021). The apparent direction of uscfDNA changes may be influenced by cancer types or preprocessing techniques and warrants further exploration. Mirroring previous literature, our fragment score analysis showed that both populations of cfDNA displayed increased fragmentation in NSCLC samples (Figure 3.7C and D) (Cristiano et al., 2019). Binned comparisons suggested that certain genomic coordinates display more distinct fragment scores (Figure 3.7E and F) and are candidate locations for further study. Bins of 1 million bp, however, will not provide enough granularity for specific sequence discovery. Other investigators have used targeted capture to report that in mncfDNA, the fragment pattern of active promoters of cfDNA shows greater randomness of fragmentation compared to inactive genes (Esfahani et al., 2022). Using a targeted panel or greater sequencing depth would be useful to observe if uscfDNA demonstrates a similar behavior.

DNA fragment end-motif profiles reflect a non-random process of orchestrated nuclease activity (Jiang et al., 2020). Strikingly, the ranking of the top 6 end-motifs was dissimilar between uscfDNA and mncfDNA (Table 3.1) and is not only suggestive of biological differences but also suggests that the populations should be interrogated separately. Although not significant, similar to previous reports, the observed trend in decreased Motif Diversity/Shannon’s Entropy cfDNA end-motif proportion could indicate a dysregulation in nuclease activity (Figure 3.8) (Serpas et al., 2019). Previous reports have indicated that the “CCCA, CCAG, CCTG” are C- motif significantly decreased in hepatocellular carcinoma (associated with downregulation of DNASE1L3 to create CNNN patterns). Although “CCAG” and “CCTG” appeared (CCCA was absent) differently in uscfDNA (all three were absent for mncfDNA), they ranked #54 and #97 in terms of q-values. This may suggest that in uscfDNA, it may reflect activity not only from DNase1L3 but also the involvement of other unexplored nucleases such as DNase2 or T1REX1 (Han & Lo, 2021).

In conclusion, we demonstrate that the BRcfDNA-Seq pipeline reveals a unique class of ultrashort single-stranded cell-free DNA of nuclear origin with a modal size of 50nt. The uscfDNA population is an exciting new cfDNA biomarker class with characteristics distinct from mncfDNA. In addition to fragmentomics and end-motif analysis, functional element peaks and enrichment in G-Quad signatures are inherent features that could help address cases where there are no clear pathognomonic somatic mutations (Mouliere et al., 2018). This exploration of alternative cfDNA features can produce biomarker candidates, which can eventually be integrated with conventional ctDNA liquid biopsy leading to greater sensitivity for cancer detection.

3.5 References

Benelli, M., Franceschini, G. M., Magi, A., Romagnoli, D., Biagioni, C., Migliaccio, I., Malorni, L., & Demichelis, F. (2021). Charting differentially methylated regions in cancer with Rockermeth. Communications Biology, 4(1), Article 1. https://doi.org/10.1038/s42003-021-02761-3

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., & Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases, 24(4), 335–341. https://doi.org/10.1016/j.cmi.2017.10.013

Bhatia, V., Barroso, S. I., García-Rubio, M. L., Tumini, E., Herrera-Moyano, E., & Aguilera, A. (2014). BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. Nature, 511(7509), 362–365. https://doi.org/10.1038/nature13374

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brambati, A., Zardoni, L., Nardini, E., Pellicioli, A., & Liberi, G. (2020). The dark side of RNA:DNA hybrids. Mutation Research/Reviews in Mutation Research, 784, 108300. https://doi.org/10.1016/j.mrrev.2020.108300

Burnham, P., Kim, M. S., Agbor-Enoh, S., Luikart, H., Valantine, H. A., Khush, K. K., & De Vlaminck, I. (2016). Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Scientific Reports, 6. https://doi.org/10.1038/srep27859

Cacho, A., Smirnova, E., Huzurbazar, S., & Cui, X. (2016). A Comparison of Base-calling Algorithms for Illumina Sequencing Technology. Briefings in Bioinformatics, 17(5), 786–795. https://doi.org/10.1093/bib/bbv088

Capra, J. A., Paeschke, K., Singh, M., & Zakian, V. A. (2010). G-Quadruplex DNA Sequences Are Evolutionarily Conserved and Associated with Distinct Genomic Features in Saccharomyces cerevisiae. PLoS Computational Biology, 6(7), e1000861. https://doi.org/10.1371/journal.pcbi.1000861

Chan, K. C. A., Jiang, P., Sun, K., Cheng, Y. K. Y., Tong, Y. K., Cheng, S. H., Wong, A. I. C., Hudecova, I., Leung, T. Y., Chiu, R. W. K., & Lo, Y. M. D. (2016). Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proceedings of the National Academy of Sciences of the United States of America, 113(50), E8159–E8168. https://doi.org/10.1073/pnas.1615800113

Chan, K. C. A., Leung, S.-F., Yeung, S.-W., Chan, A. T. C., & Lo, Y. M. D. (2008). Persistent Aberrations in Circulating DNA Integrity after Radiotherapy Are Associated with Poor Prognosis in Nasopharyngeal Carcinoma Patients. Clinical Cancer Research, 14(13), 4141–4145. https://doi.org/10.1158/1078-0432.CCR-08-0182

Chandrananda, D., Thorne, N. P., & Bahlo, M. (2015). High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. BMC Medical Genomics, 8, 29. https://doi.org/10.1186/s12920-015-0107-z

Chen, J., Ding, J., Huang, W., Sun, L., Chen, J., Liu, Y., Zhan, Q., Gao, G., He, X., Qiu, G., Long, P., Wei, L., Lu, Z., & Sun, Y. (2021). DNASE1L3 as a Novel Diagnostic and Prognostic Biomarker for Lung Adenocarcinoma Based on Data Mining. Frontiers in Genetics, 12, 699242. https://doi.org/10.3389/fgene.2021.699242

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England), 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Choy, L. Y. L., Peng, W., Jiang, P., Cheng, S. H., Yu, S. C. Y., Shang, H., Olivia Tse, O. Y., Wong, J., Wong, V. W. S., Wong, G. L. H., Lam, W. K. J., Chan, S. L., Chiu, R. W. K., Chan, K. C. A., & Lo, Y. M. D. (2022). Single-Molecule Sequencing Enables Long Cell-Free DNA Detection and Direct Methylation Analysis for Cancer Patients. Clinical Chemistry, 68(9), 1151–1163. https://doi.org/10.1093/clinchem/hvac086

Chung, J., Lee, K.-W., Lee, C., Shin, S.-H., Kyung, S., Jeon, H.-J., Kim, S.-Y., Cho, E., Yoo, C. E., Son, D.-S., Park, W.-Y., & Park, D. (2019). Performance evaluation of commercial library construction kits for PCR-based targeted sequencing using a unique molecular identifier. BMC Genomics, 20(1), 216. https://doi.org/10.1186/s12864-019-5583-7

Church, D. M. (2022). A next-generation human genome sequence. Science, 376(6588), 34–35. https://doi.org/10.1126/science.abo5367

Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., Jensen, S. Ø., Medina, J. E., Hruban, C., White, J. R., Palsgrove, D. N., Niknafs, N., Anagnostou, V., Forde, P., Naidoo, J., Marrone, K., Brahmer, J., Woodward, B. D., Husain, H., … Velculescu, V. E. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature, 1. https://doi.org/10.1038/s41586-019-1272-6

Deng, Z., Xiao, M., Du, D., Luo, N., Liu, D., Liu, T., Lian, D., & Peng, J. (2021). DNASE1L3 as a Prognostic Biomarker Associated with Immune Cell Infiltration in Cancer. OncoTargets and Therapy, 14, 2003–2017. https://doi.org/10.2147/OTT.S294332

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., Duce, J., Kauwe, J. S. K., Ridge, P. G., & for the Alzheimer's Disease Neuroimaging Initiative. (2016).

Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics, 17(7), 239. https://doi.org/10.1186/s12859-016-1097-3

Eisenstein, M. (2023). Innovative technologies crowd the short-read sequencing market. Nature, 614(7949), 798–800. https://doi.org/10.1038/d41586-023-00512-4

Esfahani, M. S., Hamilton, E. G., Mehrmohamadi, M., Nabet, B. Y., Alig, S. K., King, D. A., Steen, C. B., Macaulay, C. W., Schultz, A., Nesselbush, M. C., Soo, J., Schroers-Martin, J. G., Chen, B., Binkley, M. S., Stehr, H., Chabon, J. J., Sworder, B. J., Hui, A. B.-Y., Frank, M. J., … Alizadeh, A. A. (2022). Inferring gene expression from cell-free DNA fragmentation profiles. Nature Biotechnology, 40(4), 585–597. https://doi.org/10.1038/s41587-022-01222-4

Fang, X., Lin, H., Wang, X., Zuo, Q., Qin, J., & Zhang, P. (2015). The NEK1 interactor, C21ORF2, is required for efficient DNA damage repair. Acta Biochimica et Biophysica Sinica, 47(10), 834–841. https://doi.org/10.1093/abbs/gmv076

Foda, Z. H., Annapragada, A. V., Boyapati, K., Bruhm, D. C., Vulpescu, N. A., Medina, J. E., Mathios, D., Cristiano, S., Niknafs, N., Luu, H. T., Goggins, M. G., Anders, R. A., Sun, J., Mehta, S. H., Thomas, D. L., Kirk, G. D., Adleff, V., Phallen, J., Scharpf, R. B., … Velculescu, V. E. (2022). Detecting liver cancer using cell-free DNA fragmentomes. Cancer Discovery, CD-22-0659. https://doi.org/10.1158/2159-8290.CD-22-0659

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. Bioinformatics (Oxford, England), 28(20), 2678–2679. https://doi.org/10.1093/bioinformatics/bts503

Han, D. S. C., & Lo, Y. M. D. (2021). The Nexus of cfDNA and Nuclease Biology. Trends in Genetics: TIG, 37(8), 758–770. https://doi.org/10.1016/j.tig.2021.04.005

Hänsel-Hertsch, R., Beraldi, D., Lensing, S. V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M., Tannahill, D., & Balasubramanian, S. (2016). G-Quadruplex structures mark human regulatory chromatin. Nature Genetics, 48(10), 1267–1272. https://doi.org/10.1038/ng.3662

Hänsel-Hertsch, R., Di Antonio, M., & Balasubramanian, S. (2017). DNA G-Quadruplexes in the human genome: Detection, functions and therapeutic potential. Nature Reviews. Molecular Cell Biology, 18(5), 279–284. https://doi.org/10.1038/nrm.2017.3

Hänsel-Hertsch, R., Simeone, A., Shea, A., Hui, W. W. I., Zyner, K. G., Marsico, G., Rueda, O. M., Bruna, A., Martin, A., Zhang, X., Adhikari, S., Tannahill, D., Caldas, C., & Balasubramanian,

S. (2020). Landscape of G-Quadruplex DNA structural regions in breast cancer. Nature Genetics, 52(9), 878–883. https://doi.org/10.1038/s41588-020-0672-8

Hernandez, L. I., Araúzo-Bravo, M. J., Gerovska, D., Solaun, R. R., Machado, I., Balian, A., Botero, J., Jiménez, T., Zuriarrain Bergara, O., Larburu Gurruchaga, L., Urruticoechea, A., & Hernandez, F. J. (2021). Discovery and Proof-of-Concept Study of Nuclease Activity as a Novel Biomarker for Breast Cancer Tumors. Cancers, 13(2), 276. https://doi.org/10.3390/cancers13020276

Hu, X., Ding, S. C., & Jiang, P. (2022). Emerging frontiers of cell-free DNA fragmentomics. Extracellular Vesicles and Circulating Nucleic Acids, 3(4), 380–392. https://doi.org/10.20517/evcna.2022.34

Hudecova, I., Smith, C. G., Hänsel-Hertsch, R., Chilamakuri, C. S., Morris, J. A., Vijayaraghavan, A., Heider, K., Chandrananda, D., Cooper, W. N., Gale, D., Garcia-Corbacho, J., Pacey, S., Baird, R. D., Rosenfeld, N., & Mouliere, F. (2021). Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. Genome Research. https://doi.org/10.1101/gr.275691.121

Ignatiadis, M., Sledge, G. W., & Jeffrey, S. S. (2021). Liquid biopsy enters the clinic—Implementation issues and future challenges. Nature Reviews Clinical Oncology, 18(5), Article 5. https://doi.org/10.1038/s41571-020-00457-x

Inouye, S., & Inouye, M. (1993). The retron: A bacterial retroelement required for the synthesis of msDNA. Current Opinion in Genetics & Development, 3(5), 713–718. https://doi.org/10.1016/s0959-437x(05)80088-7

Ivanov, M., Baranova, A., Butler, T., Spellman, P., & Mileyko, V. (2015). Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. BMC Genomics, 16 Suppl 13(Suppl 13), S1. https://doi.org/10.1186/1471-2164-16-S13-S1

Jiang, P., Sun, K., Peng, W., Cheng, S. H., Ni, M., Yeung, P. C., Heung, M. M. S., Xie, T., Shang, H., Zhou, Z., Chan, R. W. Y., Wong, J., Wong, V. W. S., Poon, L. C., Leung, T. Y., Lam, W. K. J., Chan, J. Y. K., Chan, H. L. Y., Chan, K. C. A., … Lo, Y. M. D. (2020). Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discovery, 10(5), 664–673. https://doi.org/10.1158/2159-8290.CD-19-0622

Jiang, P., Sun, K., Tong, Y. K., Cheng, S. H., Cheng, T. H. T., Heung, M. M. S., Wong, J., Wong, V. W. S., Chan, H. L. Y., Chan, K. C. A., Lo, Y. M. D., & Chiu, R. W. K. (2018). Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proceedings of the National Academy of Sciences of the United States of America, 115(46), E10925–E10933. https://doi.org/10.1073/pnas.1814616115

Jin, C., Liu, X., Zheng, W., Su, L., Liu, Y., Guo, X., Gu, X., Li, H., Xu, B., Wang, G., Yu, J., Zhang, Q., Bao, D., Wan, S., Xu, F., Lai, X., Liu, J., & Xing, J. (2021). Characterization of fragment sizes, copy number aberrations and 4-mer end motifs in cell-free DNA of hepatocellular carcinoma for enhanced liquid biopsy-based cancer detection. Molecular Oncology, 15(9), 2377–2389. https://doi.org/10.1002/1878-0261.13041

Kennedy, S. R., Schmitt, M. W., Fox, E. J., Kohrn, B. F., Salk, J. J., Ahn, E. H., Prindle, M. J., Kuong, K. J., Shen, J.-C., Risques, R.-A., & Loeb, L. A. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. Nature Protocols, 9(11), 2586–2606. https://doi.org/10.1038/nprot.2014.170

Kosiol, N., Juranek, S., Brossart, P., Heine, A., & Paeschke, K. (2021). G-Quadruplexes: A promising target for cancer therapy. Molecular Cancer, 20(1), 40. https://doi.org/10.1186/s12943-021-01328-4

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., … Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research, 22(9), 1813–1831. https://doi.org/10.1101/gr.136184.111

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), Article 4. https://doi.org/10.1038/nmeth.1923

Lapin, M., Oltedal, S., Tjensvoll, K., Buhl, T., Smaaland, R., Garresori, H., Javle, M., Glenjen, N. I., Abelseth, B. K., Gilje, B., & Nordgård, O. (2018). Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer. Journal of Translational Medicine, 16. https://doi.org/10.1186/s12967-018-1677-2

Lawrie, C. H., Gal, S., Dunlop, H. M., Pushkaran, B., Liggins, A. P., Pulford, K., Banham, A. H., Pezzella, F., Boultwood, J., Wainscoat, J. S., Hatton, C. S. R., & Harris, A. L. (2008). Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. British Journal of Haematology, 141(5), 672–675. https://doi.org/10.1111/j.1365-2141.2008.07077.x

Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., & Marth, G. T. (2014). MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLOS ONE, 9(3), e90581. https://doi.org/10.1371/journal.pone.0090581

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liu, X., Liu, L., Ji, Y., Li, C., Wei, T., Yang, X., Zhang, Y., Cai, X., Gao, Y., Xu, W., Rao, S., Jin, D., Lou, W., Qiu, Z., & Wang, X. (2019). Enrichment of short mutant cell-free DNA fragments enhanced detection of pancreatic cancer. EBioMedicine, 41, 345–356. https://doi.org/10.1016/j.ebiom.2019.02.010

Liu, Y., Popp, B., & Schmidt, B. (2014). CUSHAW3: Sensitive and Accurate Base-Space and Color-Space Short-Read Alignment with Hybrid Seeding. PLOS ONE, 9(1), e86869. https://doi.org/10.1371/journal.pone.0086869

Lo, Y. M. D., Chan, K. C. A., Sun, H., Chen, E. Z., Jiang, P., Lun, F. M. F., Zheng, Y. W., Leung, T. Y., Lau, T. K., Cantor, C. R., & Chiu, R. W. K. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Science Translational Medicine, 2(61), 61ra91. https://doi.org/10.1126/scitranslmed.3001720

Mao, J. R., Inouye, S., & Inouye, M. (1997). MsDNA-Ec48, the smallest multicopy single-stranded DNA from Escherichia coli. Journal of Bacteriology, 179(24), 7865–7868. https://doi.org/10.1128/jb.179.24.7865-7868.1997

Millman, A., Bernheim, A., Stokar-Avihail, A., Fedorenko, T., Voichek, M., Leavitt, A., Oppenheimer-Shaanan, Y., & Sorek, R. (2020). Bacterial Retrons Function In Anti-Phage Defense. Cell, 183(6), 1551-1561.e12. https://doi.org/10.1016/j.cell.2020.09.065

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., & Cocho, J. A. (2013). A glimpse into past, present, and future DNA sequencing. Molecular Genetics and Metabolism, 110(1–2), 3–24. https://doi.org/10.1016/j.ymgme.2013.04.024

Mouliere, F., Chandrananda, D., Piskorz, A. M., Moore, E. K., Morris, J., Ahlborn, L. B., Mair, R., Goranova, T., Marass, F., Heider, K., Wan, J. C. M., Supernat, A., Hudecova, I., Gounaris, I., Ros, S., Jimenez-Linan, M., Garcia-Corbacho, J., Patel, K., Østrup, O., … Rosenfeld, N. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. Science Translational Medicine, 10(466). https://doi.org/10.1126/scitranslmed.aat4921

Mouliere, F., Robert, B., Arnau Peyrotte, E., Del Rio, M., Ychou, M., Molina, F., Gongora, C., & Thierry, A. R. (2011). High fragmentation characterizes tumour-derived circulating DNA. PloS One, 6(9), e23418. https://doi.org/10.1371/journal.pone.0023418

Nagata, S., Nagase, H., Kawane, K., Mukae, N., & Fukuyama, H. (2003). Degradation of chromosomal DNA during apoptosis. Cell Death and Differentiation, 10(1), 108–116. https://doi.org/10.1038/sj.cdd.4401161

Napirei, M., Ludwig, S., Mezrhab, J., Klöckl, T., & Mannherz, H. G. (2009). Murine serum nucleases—Contrasting effects of plasmin and heparin on the activities of DNase1 and DNase1-like 3 (DNase1l3). The FEBS Journal, 276(4), 1059–1073. https://doi.org/10.1111/j.1742-4658.2008.06849.x

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. Science, 376(6588), 44–53. https://doi.org/10.1126/science.abj6987

Oruba, A., Saccani, S., & van Essen, D. (2020). Role of cell-type specific nucleosome positioning in inducible activation of mammalian promoters. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-14950-5

Petersen, L. M., Martin, I. W., Moschetti, W. E., Kershaw, C. M., & Tsongalis, G. J. (2019). Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. Journal of Clinical Microbiology, 58(1), e01315-19. https://doi.org/10.1128/JCM.01315-19

Phillips, K. A., & Douglas, M. P. (2018). The Global Market for Next-Generation Sequencing Tests Continues Its Torrid Pace. The Journal of Precision Medicine, 4, https://www.thejournalofprecisionmedicine.com/wp-content/uploads/2018/11/Phillips-Online.pdf.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. Molecular Cell, 58(4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Ros-Freixedes, R., Battagin, M., Johnsson, M., Gorjanc, G., Mileham, A. J., Rounsley, S. D., & Hickey, J. M. (2018). Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. Genetics Selection Evolution, 50(1), 64. https://doi.org/10.1186/s12711-018-0436-4

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463–5467.

Schubert, M. G., Goodman, D. B., Wannier, T. M., Kaur, D., Farzadfard, F., Lu, T. K., Shipman, S. L., & Church, G. M. (2021). High-throughput functional variant screens via in vivo production of single-stranded DNA. Proceedings of the National Academy of Sciences, 118(18), e2018181118. https://doi.org/10.1073/pnas.2018181118

Serpas, L., Chan, R. W. Y., Jiang, P., Ni, M., Sun, K., Rashidfarrokhi, A., Soni, C., Sisirak, V., Lee, W.-S., Cheng, S. H., Peng, W., Chan, K. C. A., Chiu, R. W. K., Reizis, B., & Lo, Y. M. D. (2019). Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. Proceedings of the National Academy of Sciences, 116(2), 641–649. https://doi.org/10.1073/pnas.1815031116

Shen, Z., Li, Y., Fang, Y., Lin, M., Feng, X., Li, Z., Zhan, Y., Liu, Y., Mou, T., Lan, X., Wang, Y., Li, G., Wang, J., & Deng, H. (2020). SNX16 activates c-Myc signaling by inhibiting ubiquitin-mediated proteasomal degradation of eEF1A2 in colorectal cancer development. Molecular Oncology, 14(2), 387–406. https://doi.org/10.1002/1878-0261.12626

Shin, D. H., Kim, A. R., Woo, H. I., Jang, J.-H., Park, W.-Y., Kim, B. J., & Kim, S. J. (2020). Identification of the CFAP410 Pathogenic Variants in a Korean Patient with Autosomal Recessive Retinitis Pigmentosa and Skeletal Anomalies. Korean Journal of Ophthalmology: KJO, 34(6), 500–502. https://doi.org/10.3341/kjo.2020.0087

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. (2016). Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell, 164(0), 57–68. https://doi.org/10.1016/j.cell.2015.11.050

Teo, Y. V., Capri, M., Morsiani, C., Pizza, G., Faria, A. M. C., Franceschi, C., & Neretti, N. (2019). Cell-free DNA as a biomarker of aging. Aging Cell, 18(1), e12890. https://doi.org/10.1111/acel.12890

Todd, A. K., Johnston, M., & Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Research, 33(9), 2901–2907. https://doi.org/10.1093/nar/gki553

Ulz, P., Thallinger, G. G., Auer, M., Graf, R., Kashofer, K., Jahn, S. W., Abete, L., Pristauz, G., Petru, E., Geigl, J. B., Heitzer, E., & Speicher, M. R. (2016). Inferring expressed genes by whole-genome sequencing of plasma DNA. Nature Genetics, 48(10), 1273–1278. https://doi.org/10.1038/ng.3648

Umetani, N., Giuliano, A. E., Hiramatsu, S. H., Amersi, F., Nakagawa, T., Martino, S., & Hoon, D. S. B. (2006). Prediction of breast tumor progression by integrity of free circulating DNA in serum. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 24(26), 4270–4276. https://doi.org/10.1200/JCO.2006.05.9493

Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., Gligorich, K. M., Rostomily, R. C., Bronner, M. P., & Shendure, J. (2016). Fragment Length of Circulating Tumor DNA. PLOS Genetics, 12(7), e1006162. https://doi.org/10.1371/journal.pgen.1006162

van der Pol, Y., & Mouliere, F. (2019). Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. Cancer Cell, 36(4), 350–368. https://doi.org/10.1016/j.ccell.2019.09.003

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends in Genetics: TIG, 34(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Varshney, D., Spiegel, J., Zyner, K., Tannahill, D., & Balasubramanian, S. (2020). The regulation and functions of DNA and RNA G-Quadruplexes. Nature Reviews. Molecular Cell Biology, 21(8), 459–474. https://doi.org/10.1038/s41580-020-0236-x

Vessies, D. C. L., Schuurbiers, M. M. F., van der Noort, V., Schouten, I., Linders, T. C., Lanfermeijer, M., Ramkisoensing, K. L., Hartemink, K. J., Monkhorst, K., van den Heuvel, M. M., & van den Broek, D. (2022). Combining variant detection and fragment length analysis improves detection of minimal residual disease in postsurgery circulating tumour DNA of stage II-IIIA NSCLC patients. Molecular Oncology, 16(14), 2719–2732. https://doi.org/10.1002/1878-0261.13267

Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry, 55(4), 641–658. https://doi.org/10.1373/clinchem.2008.112789

Wang, B. G., Huang, H.-Y., Chen, Y.-C., Bristow, R. E., Kassauei, K., Cheng, C.-C., Roden, R., Sokoll, L. J., Chan, D. W., & Shih, I.-M. (2003). Increased plasma DNA integrity in cancer patients. Cancer Research, 63(14), 3966–3968.

Xiao, Y., Yang, K., Liu, P., Ma, D., Lei, P., & Liu, Q. (2022). Deoxyribonuclease 1-like 3 Inhibits Hepatocellular Carcinoma Progression by Inducing Apoptosis and Reprogramming Glucose Metabolism. International Journal of Biological Sciences, 18(1), 82–95. https://doi.org/10.7150/ijbs.57919

Yamada, H., Takahashi, M., Watanuki, M., Watanabe, M., Hiraide, S., Saijo, K., Komine, K., & Ishioka, C. (2021). LncRNA HAR1B has potential to be a predictive marker for pazopanib therapy in patients with sarcoma. Oncology Letters, 21(6), 455. https://doi.org/10.3892/ol.2021.12716

Yu, S. C. Y., Jiang, P., Peng, W., Cheng, S. H., Cheung, Y. T. T., Tse, O. Y. O., Shang, H., Poon, L. C., Leung, T. Y., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2021). Single-molecule sequencing reveals a large population of long cell-free DNA molecules in maternal plasma. Proceedings of the National Academy of Sciences, 118(50), e2114937118. https://doi.org/10.1073/pnas.2114937118

Zhang, L., Qin, D., Hao, C., Shu, X., & Pei, D. (2013). SNX16 negatively regulates the migration and tumorigenesis of MCF-7 cells. Cell Regeneration, 2(1), 2:3. https://doi.org/10.1186/2045-9769-2-3

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology, 9(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhitnyuk, Y. V., Koval, A. P., Alferov, A. A., Shtykova, Y. A., Mamedov, I. Z., Kushlinskii, N. E., Chudakov, D. M., & Shcherbo, D. S. (2022). Deep cfDNA fragment end profiling enables cancer detection. Molecular Cancer, 21(1), 26. https://doi.org/10.1186/s12943-021-01491-8

4

ASSESSING THE METHYLATION PROFILE AND THE BIOMARKER CAPABILITY OF ULTRASHORT SINGLE-STRANDED CELL-FREE DNA IN THE PLASMA OF NON-SMALL CELL LUNG CARCINOMA SUBJECTS

4.1 Abstract

Epigenetic abnormalities in the genome promote and propagate the development of oncological behavior in cells. These genomic aberrations can be reflected in the cell-free DNA and have been used as viable biomarkers for early cancer detection and treatment guidance. The epigenetic configuration of the genome reflects both the cell type and metabolic state. Correspondingly, cell-free DNA can be used to infer the tissue of origin of the cell-free DNA within the plasma. Although the methylation profile of mononucleosomal cell-free DNA (mncfDNA) has been previously described, the methylation profile of ultrashort single-stranded cell-free DNA (uscfDNA) has not been previously explored. In this chapter, we describe an optimized library preparation protocol for uscfDNA in which single-stranded 5-

methylcytosine (5mC)-premethylated adapters are ligated to cell-free DNA fragments prior to bisulfite conversion to assess methylation patterns for each size of cfDNA populations. This method improves the accuracy of downstream uscfDNA analysis by preventing bisulfite-degraded DNA from being incorporated into the final library. Using this technique, we report that compared to mncfDNA, bisulfite-converted uscfDNA is enriched in promoter and CpG islands and is globally lower by ~15%. Additionally, the fragments mapping to SINE, LINE, and simple repeat elements exhibit distinct methylation patterns based on the original fragment size (uscfDNA vs. mncfDNA). Using the CeLfie methylation deconvolution algorithm, we infer that uscfDNA may derive from eosinophils, neutrophils, and monocytes. As a proof of concept, we show that the methylation profile of uscfDNA can distinguish non-small cell lung carcinoma and non-cancer subjects through hypermethylation of regulatory and G-Quad-containing elements, methylated region (DMR) candidates, tissue-of-origin profiles, and coverage alterations in regions associated with epigenetic marks of active expression. In closing, this methodology is recommended for any methylation-based investigations involving ultrashort DNA footprints.

4.2 Introduction

4.2.1 Introduction to cfDNA Methylation as a Biomarker

Epigenetics is the study of alterations of gene activity within cells without permanent alterations to the DNA sequence. Significant types of epigenetic modifications include DNA methylation, histone modifications, and small RNA expression (Allis & Jenuwein, 2016). Unlike changes in the genetic sequence, epigenetic modifications are reversible. With the rise of accessible sequencing technologies, an emerging interest is studying the methylation patterns that constitute the epigenetic patterns in biological and disease conditions. Interestingly, despite its fragmented nature, plasma cell-free DNA derived from necrotic or apoptotic cells retains the original genomic DNA's inherent methylation and histone modification features. This previous work has been focused on mononucleosomal cell-free DNA (mncfDNA). However, the methylation pattern of ultrashort single-stranded cell-free DNA (uscfDNA) has not been previously examined. Exploring the methylation behavior of this cell-free DNA sub-population can provide insights into epigenetic biology, tissue of origin, and biomarker potential of uscfDNA.

4.2.2 Methylation Principles

Methylation is the addition of a methyl group to the 5th carbon of a cytosine residue by DNA methyltransferase enzymes (DNMTs), resulting in a 5-methylcytosine residue (5mC). In mammals, most 5-methylcytosines are preceded by guanine residues in a term called "CpG methylation." It has been observed that upstream to 70% of promoter regions in the genome, there is often the presence of dense clusters of CpG residues called CpG islands. These regulatory regions are known as CpG islands (CGIs) (Saxonov et al., 2006). CpG islands tend to have lower methylation rates, and the further away CpG dinucleotides are from CpG islands, the higher probability those residues will be methylated (Lister et al., 2009; Shimoda et al., 2014). CpG shores, which are regions typically 2 kb upstream and downstream from CpG islands, are often moderately methylated and are negatively associated with gene expression levels (Irizarry et al., 2009).

In cellular activity, methylation is involved in gene regulation, embryonic development (Okano et al., 1999), genomic imprinting (Reik & Walter, 2001), stem-cell differentiation, and transposon inactivation (Smith & Meissner, 2013). Historically, in vitro experiments in Xenopus oocytes demonstrated that methylation can also modulate gene repression (Vardimon et al., 1982).

DNA methylation can change the functional state of regulatory regions but does not affect the pairing of the cytosine and guanine restudies in the DNA. Methylation provides an epigenetic marking and contributes to stable epigenetic repression by imprinting, X inactivation, or silencing of repetitive sequences of DNA (Jones, 2012). Once methylated, downstream histone modification, nucleosome positioning, altered DNA binding proteins activity, and transcription factors lead to downregulating gene activity. Contrastingly, regulatory sequences take on an unmethylated state when active (Ziller et al., 2013).

In contrast, to CpG methylation, contexts such as CpA, CpT, and CpC are considered non-CpG methylation. Non-CpG methylation is associated with plants and fungi and previously was thought to not be involved in mammals. In mammals, non-CG methylation was initially considered an artifact of incomplete conversion during bisulfite treatment. In contrast, there is growing evidence that certain cell types within mammals have traces of non-CpG methylation. Non-CpG methylation has been reported in embryonic stem cells (ESC), somatic cell nuclear transfer stem cells (SCNT-SC), pluripotent stem cells (iPSCs), oocytes, neutrons, and glia cells (Titcombe et al., 2022). The biological role of non-CG is still unclear, but it could also play a role in gene regulation and cancer (Ramasamy et al., 2021).

4.2.3 Methylation Analysis Techniques

Integral to dissecting the role of DNA methylation in health and disease scenarios are tools that can simultaneously measure DNA methylation across large portions of the genome. For methylation analysis, most technologies involve differentiation between the presence of a 5-methylcytosine (5mC) and a regular cytosine residue.

Technologies for methylation analysis can be categorized into two methods. One type can be considered non-sequenced methylation analysis, and the other type provides a granularity of which specific cytosine is methylated in whole-genome DNA methylation profiling analysis. Enzyme-linked immunosorbent assay (ELISA) and luminometric methylation assays (LUMA) are two established strategies for non-sequenced DNA methylation analysis. The ELISA method is rapid and convenient but lacks specificity. LUMA, on the other hand, harnesses the cleaving points of HpAII or MspI catalyzation points but will not reveal where the specific cytosine residues are.

Bisulfite treatment is considered the gold standard for genome-wide methods for methylation analysis and was first described in 1992 (Frommer et al., 1992). Treatment of genomic DNA (or any DNA) with sodium bisulfite will convert any unmethylated cytosine to uracil, whereas methylated cytosines do not undergo this conversion.

In this chapter, we use the EZ DNA Methylation-Lightning Kit, which contains a ready-to-use bisulfite conversion reagent that is directly added to the DNA. Although proprietary bisulfite conversion reagent. In combination, they include an M-binding buffer composed of Tris-HCL, EDTA, and NaCl to encourage the converted DNA to bind to the silica-based matrix that allows DNA purification in a durable polypropylene construction. The L-Desulphonation Buffer desulfonates the sulphonate attached to the cytosine and uracil at the final steps. The M-Wash buffer will use to wash any impurities off the bounded DNA in the Zymo-Spin^{TM} IC Column. The elution buffer is a DNA-philic reagent likely containing water and tris-EDTA to elute the DNA off the column.

One drawback of bisulfite conversion is that it can be destructive to DNA. One report concluded that bisulfite appears to break down DNA to an average length of 600bp (Munson et al., 2007). Initial fragment size and concentration are factors that will affect the final yield of bisulfite-treated cfDNA that can be used for downstream analysis. A false positive result can occur if there is incomplete deamination. In contrast, methylated cytosine reacts at a much lower rate, and most residues will appear unchanged(Werner et al., 2019). One study demonstrated that shorter fragments of cell-free DNA, such as the 167-bp mncfDNA, undergo lower amounts of fragmentation than larger molecular weight cfDNA (Werner et al., 2019). However, this suggests that if larger molecular weight cfDNA or genomic DNA disintegrate, their non-native fragmented sizes can occlude smaller footprint cfDNA.

An alternative to the non-bisulfite method is methyl-DNA immunoprecipitation in combination with next-generation sequencing (MeDIP-seq) (S. Y. Shen et al., 2018). This genome-wide approach enriches and analyzes methylated DNA by capturing 5-mC molecules with an antibody. This method requires a certain level of DNA concentration from the sample to successfully pull down enough material for analysis.

Although bisulfite-based whole genome sequencing can provide a complete map of the ~28 million CpG sites in the human genome, it requires sufficient sequencing depth for adequate coverage. One alternative is to only enrich the signal from specific methylation targets assessed using a for certain genome regions. This methodology works by having individual beads along a multiplexed surface with probe DNA for specific sequences. One established manufacturer is Illumina Infinium BeadChips. In this system, each bead has a 23-base pair address (to determine location) and a 50-base pair probe for a genomic region. These probe sequences complement a 50-base pair region after bisulfite conversion. Although easy to use, these arrays are designed for cellular experiments where large amounts of DNA (250-750ng) are readily available. In contrast, an attempt to use this array for cell-free DNA applications requires the pooling of DNA from the plasma of multiple subjects to accumulate enough DNA. Since cfDNA cannot be amplified (unlike intact genomic DNA), it is not feasible to amplify before or after bisulfite conversion to gather enough DNA for a bead array (Moss et al., 2018).

4.2.4 Methylation in Lung Cancer Cells

Cancer cells exhibit epigenetic abnormalities in the methylation profile, associated with oncogenic cell transformation and genomic instability. In general, it has been observed that genome-wide hypomethylation is present in cancer tissue samples compared to those of healthy (Jones & Baylin, 2002). Hypomethylation appears to occur in gene-poor regions such as repetitive DNA regions and several coding and intronic regions. These hypomethylation patterns lead to chromosome instability, transposon activation, loss of imprinting, and mitotic recombination (Esteller, 2008). Additionally, this can also activate silenced oncogenes and retrotransposon elements. Opposingly specific regional hypermethylation is also associated with cancer cells.   CpG islands which are typically hypomethylated, become hypermethylated during cancer which can silence tumor suppressor genes (Weber et al., 2005).

More specifically, because CpG sites are regulated by the interplay between DNMTs and DNA demethylase activity, the dysregulation of this homeostasis can lead to cancer transformation events. Several methylation issues that occur include DNMT dysregulation, TET (Ten-eleven translocation) enzyme dysregulation, hypomethylation, and hypermethylation. For DNMT, there are five known DNMTs in humans: DNMT1, DNMT2, DNMT3A, DNMT3B, and DNMT3L. These subtypes demonstrate variable activity and behavior toward different methylation contexts. DNMT1 acts in hemimethylated CpG sites and is responsible for the maintenance of the newly synthesized DNA during replication (Jurkowska et al., 2008) DNMT3A and DNMT3B methylated CpG sites during germ cell development an early embryonic stage. DNMT2 methylates tRNAs (Tuorto et al., 2015). DNMT3L does not appear to catalyze methylation reactions on its own but instead enhances DNMT3A and DNMT3B's activity (Jurkowska et al., 2008). Overexpression of DNMT has been observed in lung cancer with the upregulation of DNMT1, which is associated with a poor prognosis (H. Kim et al., 2006; Lin et al., 2007, 2010). Opposingly, in lung cancer cell models, depletion of DNMT1 and DNMT3B resulted in growth arrest, apoptosis, and reactivation of tumor suppressor genes. DNMT dysregulation can disrupt the cell type (Espada et al., 2007). Overexpression of DNMT increases cell proliferation in lung cancers as they are involved in ribosome synthesis and may be involved maturation of the rRNA (Tang et al., 2009). DNMT1 suppresses promoters such as hMLH1 and hmSH2, which normally suppress the cell cycle. Therefore, when these promoters are suppressed, the cells begin to proliferate (Wu et al., 2020).

Dysregulation of TET enzymes is also associated with cancer promotion (Rasmussen & Helin, 2016). TET oxidizes 5mC and reverses its suppressive activity. The family of TET includes TET1, TET2, and TET3, which are all capable of oxidation but differ in their molecular architecture (Kohli & Zhang, 2013). Compared to TET1 and TET3, TET2 has a higher frequency of somatic mutations. There have been mixed observations for TET1 behavior in primary tumor models. Recent studies suggest that TET1 is downregulated by DNA promoter methylation, which prevents its ability to correct the methylation dysregulation shifting toward CpG methylation tumor suppressor genes (Yang et al., 2013). Despite these observations, no global rule encompasses all the complex aberrations that can occur in the epigenetic systems in cancer (Filipczak et al., 2019).

Lung cancer mirrors other cancers in other anatomical sites in that global hypomethylation occurs in repetitive regions such as Short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), and subtelomere repeats (Rauch et al., 2008). Hypomethylation of LINE-1 is associated with a worse prognosis in advanced stages.

Typically, regulatory regions are hypermethylated, but hypomethylation can also result in cancer oncogene activation in some situations. For example, demethylation of CpG sites has been observed in synuclein gamma (SNCG) (H. Liu et al., 2005) which is involved in cancer migration and invasion (Shao et al., 2018). Another example are melanoma-associated antigen (MAGE) genes, which are upregulated in 70-85% of NSCLC tumors, and their upregulation is correlated with hypomethylation(Jang et al., 2001). Hypomethylation also contributes to genomic instability by reactivating retrotransposons. Hypomethylation at the 3' tandem repeat region of HRAS has been shown to contribute to gene loss (Vachtenheim et al., 1994) and hypomethylation of retro transposable elements such as LINE-1 and Alu can enhance transcription and instability in NSCLC (Daskalos et al., 2009).

One paradigm shift theory regarding hypomethylation is that increased hypomethylation could instead result in rearrangement in an attempt to slow down cancer growth rather than a cause of cancer. Thus, global hypomethylation may be a tumor-suppressive pattern rather than a tumor-promoting trait(Johnstone et al., 2020).

In contrast to hypomethylation, there are abundant studies on targets of hypermethylation. Various common CGislands of tumor suppressor genes have been described to become hypermethylated in lung cancers. These genes are critical in cellular functions and are also seen to be dysregulated in cancer. Apoptosis genes (CASP8, DAPK, TNFRSF6, DR4, DR5) (Hopkins-Donaldson et al., 2003; D. H. Kim, Nelson, Wiencke, Christiani, et al., 2001; Shivapurkar et al., 2002), cell cycle regulation genes (CDK2A,p16, PTEN, RASSF1A). ((Baylin et al., 2001; D. H. Kim, Nelson, Wiencke, Zheng, et al., 2001; Merlo et al., 1995), DNA repair genes (MGMT, MLH1, MSH2) (Brabender et al., 2003; Gomes et al., 2014) signal pathway genes (APC, RARB-2, RUNX3, SHOX2) (Grote et al., 2004; D. H. Kim, Nelson, Wiencke, Christiani, et al., 2001; Schmidt et al., 2010), cell adhesion and invasion genes (CDH1, CDH13, TSLC1) (Heller et al., 2006; D. S. Kim et al., 2007) impact cellular function when their regular methylation is disrupted. Modifying epigenetic patterns does not necessarily lead to gene inactivation, but their changes may be critical markers indicative of an arising cancer state.

In summary, these epigenetic cellular hypo- and hypermethylation changes could be observed in cell-free DNA and be viable indicators of tumor activity from distant sites.

4.2.5 Methylation for Lung Cancer Liquid Biopsy

Examining methylation of cell-free DNA has many clinical promises if applied to a liquid biopsy application (Xu et al., 2019). The genome has 28 million CpG sites (Babenko et al., 2017) which can be differentially methylated to affect gene expression. An analysis of the methylation patterns of cell-free DNA can provide another aspect of cfDNA in addition to tracking the whole-genome sequence alone.

Identifying the 5mC modifications is one approach for differentiating cancer from healthy (Chan et al., 2013; Lehmann-Werman et al., 2016; Sun et al., 2015). Since methylation profiles are unique to each cell type, not only in dormancy but also status, one advantage of examining methylation is that it can infer the tissue of origin from which the cell-free DNA originated. Ergo, this has allowed cell-free methylation analysis to inform changes for other non-cancer diseases. For example, in other diseases models, cfDNA methylation patterns are useful in identifying pancreatic B-cell death for type-1 diabetes, or islet-graft recipients, oligodendrocyte DNA in multiple sclerosis, and neuronal DNA in traumatic brain injury or cardiac arrest, and exocrine pancreatic DNA in pancreatic cancer or pancreatitis (Lehmann-Werman et al., 2016).

Early search into cfDNA methylation viability showed that SHOX2 (a gene involved in signal transduction) methylation status could be detected in the plasma and was reported to have a sensitivity of 60% and specificity of 90% to differentiate lung cancer from non-cancer controls(Kneip et al., 2011). Using quantitative methylation specific-PCR demonstrated that the RASSF1A and RARB gene methylation levels were increased (Ponomaryova et al., 2013). Examining the circulating signal of genes in the plasma revealed that the methylation values of

B3GAT2, BCAR1, HOPX, HOXD11, MIR1203, MYL9, SLC9A3R2, SYT5, VTRNA1-3, and HLF genes were different in lung cancer for healthy control samples (Xu et al., 2019).

Examining cell-free presentation of promoter methylation aberrations has also been considered. A study looking at the cfDNA DCLK1 promoter methylation using methylation-specific PCR concluded that out of 65 lung cancer patients, 49.2% of them showed DCLK1 promoter methylation with a 42.9% sensitivity and 91.% specificity (Powrózek et al., 2016).

Alterations in multiple methylation targets are another strategy for cfDNA biomarkers. In one study, a panel of six genes: CDO1, HOXA9, AJAP1, PTGDR, UNCX, and MARCH1 could collectively detect lung cancer with a sensitivity of 90% for early-stage lung cancer (IA) (Ooki et al., 2017). Another study showed that RTEL1 and PCDHGB6 promoter methylation status could differentiate NSCLC I-III with an AUC of 0.75, a sensitivity of 64.6%, and a specificity of 90.0% (Olsson et al., 2016). In another study, the investigators used a biomarker set of MIR129-2, LINC01158, CCDC181, PRKCB, TBR1, ZNF781, MARCH11, VWC2, SLC9A3, and HOXA7 (Vrba et al., 2020) and claimed an AUC of 0.956, sensitivity of 83% and specificity of 95%.

Aside from delineating between different disease cohorts, cfDNA methylation has demonstrated the ability to monitor the disease status of the subjects. One study examined the cfDNA promoter methylation status of BRMS1, showing that those with methylated BRMS1 had lower overall survival and progression-free survival time (Balgkouranidou et al., 2014). In another example, NSCLC patients with methylated cell-free DNA KMT2C had lower overall survival in operable and metastatic NSCLC patients (Mastoraki et al., 2021). SOX17 is another promoter region in its highly methylated form negatively correlated to the NSCLC survival prognosis (Balgkouranidou et al., 2016). These studies are suggestive that cfDNA methylation has clinical merit and should be further explored.

4.2.6 Tissue of Origin Analysis

Considering that cell-free DNA is composed of DNA originating from distant cell types, the key challenge of liquid biopsy analysis is to infer which cell type the cfDNA is derived from. Determining the source of the primary cancer is essential for early diagnosis. Different tissue groups for cancer do not necessarily have specific pathognomonic mutations. Iconic somatic mutations such as TP53, KRAS, and ERBB2 are involved in various cancers of different tissue types (Cosmic Database, 2018). If the tissue of origin associated with this mutation can be determined, it may be useful for pinpointing the site of urgency.

Epigenetic features can be useful for determining the tissue of origin since each cell type has a particular methylation status guiding its genome architecture. Methylation of nucleotides has been used to reversibly identify genomic DNA and is a fundamental mark of cell identity (Dor & Cedar, 2018). Several reports have shown that cell-free DNA from specific genomic loci tissue-specific methylation locations can identify the cell of origin (Gai et al., 2018; Gala-Lopez et al., 2018; Lam et al., 2017; Lehmann-Werman et al., 2016, 2018; Zemmour et al., 2018). Other investigators have taken a whole-genome sequencing approach using whole-genome bisulfite sequencing to infer the tissue of origin of four tissues (Sun et al., 2015). In healthy donors, analysis through methylation markers of cfDNA indicates that they mainly originate from immune blood cells such as neutrophils, monocytes, macrophages, or megakaryocytes, but other tissue types such as skeletal and endothelial contribute as well (Lui et al., 2002; Moss et al., 2018; Sun et al., 2015).

Another approach is reduced representative bisulfite sequencing. This technique uses a specific restriction enzyme (commonly MsPI) to enrich CpG content areas and regions with potential CpG methylation. This reduces the amount of nucleotides required to 1% of the genome (Meissner et al., 2005). Leveraging this technique, another group was able to identify the origin of two cancer types (S. Guo et al., 2017). On the back end, other groups have optimized the probabilistic approach for cancer detection using pre-existing data and some newer samples to deconvolute based on the methylation pattern (Kang et al., 2017; W. Li et al., 2018).

Alternatively, since cell-free DNA presents in a fragmented form, several creative strategies have been employed to attempt to deconvolute the tissue of origins of cell-free plasma. Snyder et al. attempted to use fragment sequencing as a surrogate for nucleosome positioning, which reflects expression signatures of specific cells (Snyder et al., 2016). This concept was further explored by using the available whole-genome sequencing fragment sequences to infer the gene expression profile in the plasma for the cell of origin (Ulz et al., 2016).

4.2.7 Deconvolution Algorithms

Deconvolution based on DNA methylation is useful when inferring the different cell types within a sample (Da et al., 2010; Titus et al., 2017). Historically, one of the earliest demonstrations of successful deconvolution was in leukocyte subtypes in the blood, which could be parsed out using methylation markers in contrast to conventional histological or cytometric assessments (Houseman et al., 2012). Deconvolution can be based on reference data sets or trained on the samples (Houseman et al., 2012; Teschendorff et al., 2017).

Methylation-based deconvolution relies on epigenome-wide association studies (EWAS), which have shown that changes in that cell proportion can change in different disease states. Initial EWAS data was collected using targeted panels.

In intact cells, DNA-methylation deconvolution relies on establishing a cell-specific differentiated methylation region profile (DMR) using a sorted purified cell population. Ideally, cell-type DMR patterns should be consistent amongst reference samples, and since methylation retains the hereditary nature, the methylation profiles tend to be very robust and consistent. Although it has been shown to change in fetal/newborn blood samples meaning that a separate reference is required (de Goede et al., 2015; Gala-Lopez et al., 2018)

There are two classes of cell-type deconvolution approaches: reference-based and without reference. Houseman developed the original reference-based algorithm with sample DNA as a weighted combination of the individual methylation profiles from underlying cell types (Houseman et al., 2012). In contrast, reference-free methods have been developed which use machine learning to estimate the cellular proportions of unsupervised deconvolution methods (Teschendorff et al., 2017). Advantages of reference-based deconvolution include quantifying cell types at a single-sample level, detecting alterations in cell types, and the absence of assumptions. Disadvantages are that it requires knowledge of the cell types, good quality DNA-methylation profiles, and the inability to account for cell-cell interactions. For reference-free algorithms, the advantages are that it does not require knowledge of cell types or pre-existing reference material. It applies to any tissue type and can account for cell-cell interactions. The disadvantages are that these models require assumptions and cannot provide details on individual samples since it relies on batch analysis.

4.2.8 Chapter Goals

In this chapter, we attempted to determine the methylation state of uscfDNA by developing a method for determining the methylation profile of uscfDNA with the hope of revealing insights into its epigenetic characteristics. We aimed to circumvent degradation of bisulfite-induced degradation by developing a new method where permethylated single-stranded adapters (5mCAdpBS-Seq) are ligated to the cfDNA fragments prior to bisulfite conversion and subsequent library amplification. As a proof of concept, we used this method to evaluate if the methylation features of uscfDNA can be used as a new novel biomarker for cancer detection applications by analyzing a small cohort of late-stage NSCLC plasma samples compared to non-cancer controls.

4.3 Results

4.3.1 Merging Paired-End Reads Prior to Alignment Impacts Fragment Length Profile of BS-Treated cfDNA Libraries

Since uscfDNA and mncfDNA are fragment sizes that are <300bp, conventional 2x150bp sequencing will result in sequencing some sections of the cfDNA fragment twice. We observed that both paired reads demonstrated a degree of overlapping sequences ranging from complete overlap (uscfDNA – 50bp) to partial overlap (mncfDNA – 167bp). There are also certain circumstances when two paired-end reads do not have consensus sequences that justify their exclusion (Figure 4.1A and B). We tested the approach of merging paired reads prior to alignment on a cohort of sequenced bisulfite-conversion prior to adapter attachment libraries (BS-Seq) compared to non-bisulfite converted BRcfDNA-Seq libraries (Figure 4.1C-F). In the unmerged analysis, the BRcfDNA-Seq library demonstrated a two-peak profile indicating the presence of uscfDNA (40-70bp) with a peak at ~52bp and mncfDNA peak at ~167bp (Figure 4.1C). After BS-Seq, the two major peaks at the uscfDNA and mncfDNA region also remained (Figure 4.1D). For the uscfDNA peak, however, it appeared to have shifted from 52bp to 55bp compared to the BRcfDNA-Seq libraries. Interestingly, in the mncfDNA region, two peaks were present (150bp and 167bp). A substantial proportion of aligned reads with a length between 75 to 130bp was not present in the BRcfDNA-Seq libraries. With the merged-reads bioinformatic processing, the BRcfDNA-Seq libraries demonstrated a similar profile as with the unmerged (Figure 4.1C). For the BS-Seq libraries, after merging, there were two major observations: Firstly, the uscfDNA peak shifted back to 52bp, and secondly, the mncfDNA peak, it now appeared as a single 167bp peak (Figure 4.1D).

When calculating MAPQ scores for every bin size, the BRcfDNA-Seq libraries had comparable MAPQ scores, with the merged-reads protocol being slightly lower (Figure 4.1E and F). For the BS-Seq, both default and merged processing had a lower MAPQ score for bins from 30-39bp but stabilized for the bins >40bp. The merged reads were slightly lower compared to the paired-end processing.

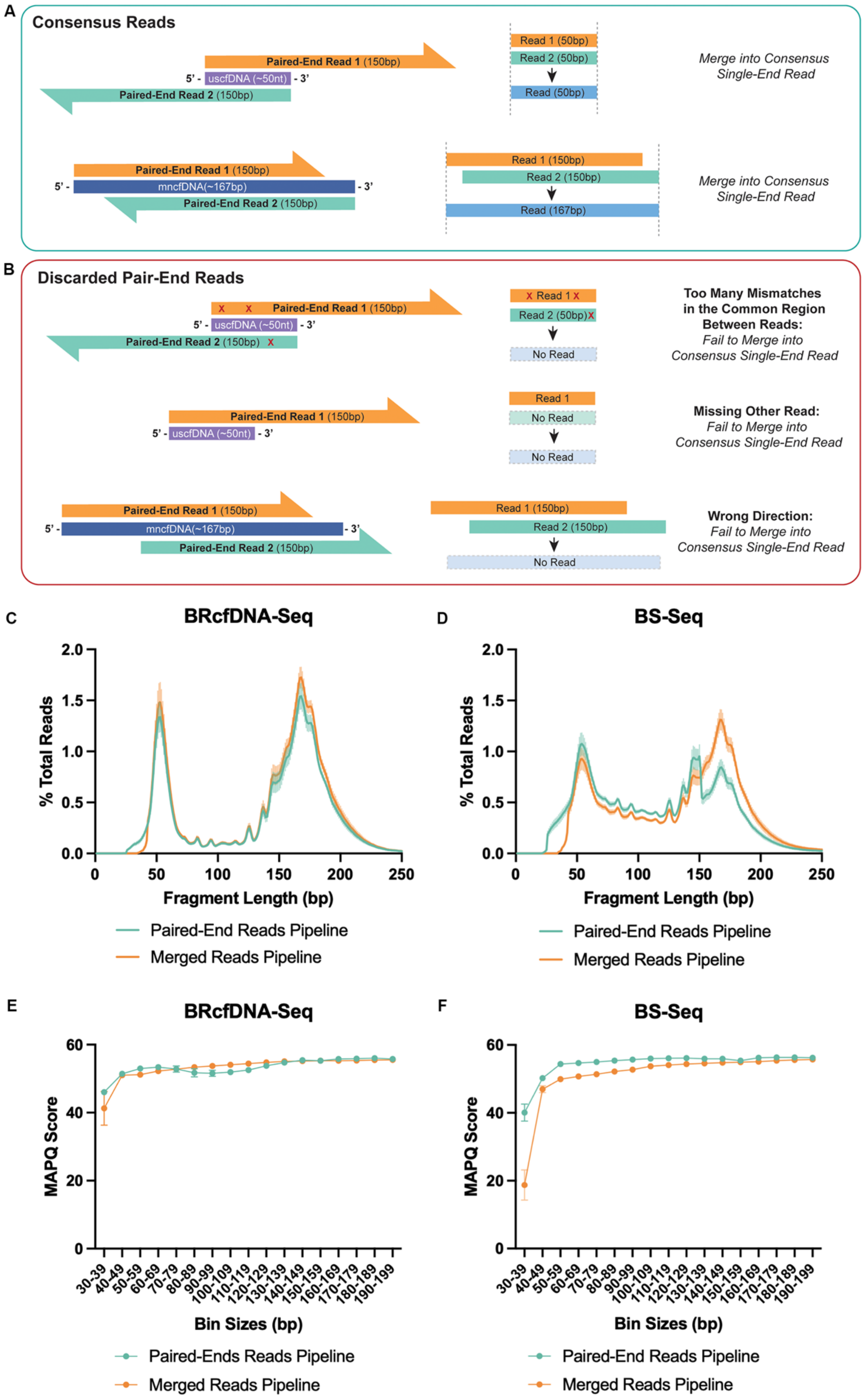
A
Consensus Reads
Paired-End Read 1 (150bp)
5' - uscfDNA (~50nt) - 3'
Paired-End Read 2 (150bp)
Read 1 (50bp)
Read 2 (50bp)
Read (50bp)
Merge into Consensus Single-End Read
Paired-End Read 1 (150bp)
5' - mncfDNA(~167bp) - 3'
Paired-End Read 2 (150bp)
Read 1 (150bp)
Read 2 (150bp)
Read (167bp)
Merge into Consensus Single-End Read
B
Discarded Pair-End Reads
X X Paired-End Read 1 (150bp)
5' - uscfDNA (~50nt) - 3'
Paired-End Read 2 (150bp) X
X Read 1 X
Read 2 (50bp) X
No Read
Too Many Mismatches in the Common Region Between Reads:
Fail to Merge into Consensus Single-End Read
Paired-End Read 1 (150bp)
5' - uscfDNA (~50nt) - 3'
Read 1
No Read
No Read
Missing Other Read:
Fail to Merge into Consensus Single-End Read
Paired-End Read 1 (150bp)
5' - mncfDNA(~167bp) - 3'
Paired-End Read 2 (150bp)
Read 1 (150bp)
Read 2 (150bp)
No Read
Wrong Direction:
Fail to Merge into Consensus Single-End Read
C
BRcfDNA-Seq
% Total Reads
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
250
Fragment Length (bp)
Paired-End Reads Pipeline
Merged Reads Pipeline
D
BS-Seq
% Total Reads
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
250
Fragment Length (bp)
Paired-End Reads Pipeline
Merged Reads Pipeline
E
BRcfDNA-Seq
MAPQ Score
60
40
20
0
30-39
40-49
50-59
60-69
70-79
80-89
90-99
100-109
110-119
120-129
130-139
140-149
150-159
160-169
170-179
180-189
190-199
Bin Sizes (bp)
Paired-Ends Reads Pipeline
Merged Reads Pipeline
F
BS-Seq
MAPQ Score
60
40
20
0
30-39
40-49
50-59
60-69
70-79
80-89
90-99
100-109
110-119
120-129
130-139
140-149
150-159
160-169
170-179
180-189
190-199
Bin Sizes (bp)
Paired-End Reads Pipeline
Merged Reads Pipeline

Figure 4.1 Merging paired-end reads demonstrates a more similar profile to untreated non-targeted sequencing. Schematic pre-merging pipeline paired-end reads prior to alignment. A) Situations where reads are accepted for downstream analysis. In the uscfDNA (~50bp) scenario consensus sequence of read 1 and 2 should have 100% overlap, whereas for mncfDNA (~167bp), there will be a 150bp perfect overlap of read 1 read 2 with 17 bp with no overlap. These reads will still be accepted. B) Potential Scenarios where reads fail to merge and are discarded from downstream analysis. C) BRcfDNA-Seq libraries show little difference in pattern with and without merging per processing pipeline. The lines represent five samples with the dashes lines as standard error. D) BS-Seq libraries show a difference in the pattern when reads are merged prior to alignment (dip at 150bp). MAPQ scores for binned reads of 10bp for BRcfDNA-Seq (E) and BS-Seq (F) libraries for both paired ends and merged bioinformatic preprocessing. Vertical lines in (C&D) indicate SEM from the mean of 5 subjects. Some error bars may not be observable in C and D due to their length being smaller than the size of the data point.

We examined the change in the percentage of total reads as each sequence library underwent each bioinformatic pipeline step (Figure 4.2A,D). A large proportion of reads (72.6% ± 3.2 reads remaining) were observed to be eliminated during the merging step (Figure 4.2B,E). When compared to the unmerged analysis, merged processing universally resulted in lower remaining reads (51.9 ± 4.7% vs. 46.6 ± 3.6%) (Figure 4.2C,F). Despite the more significant read loss, all sequenced libraries were processed with both reads merged prior to alignment.

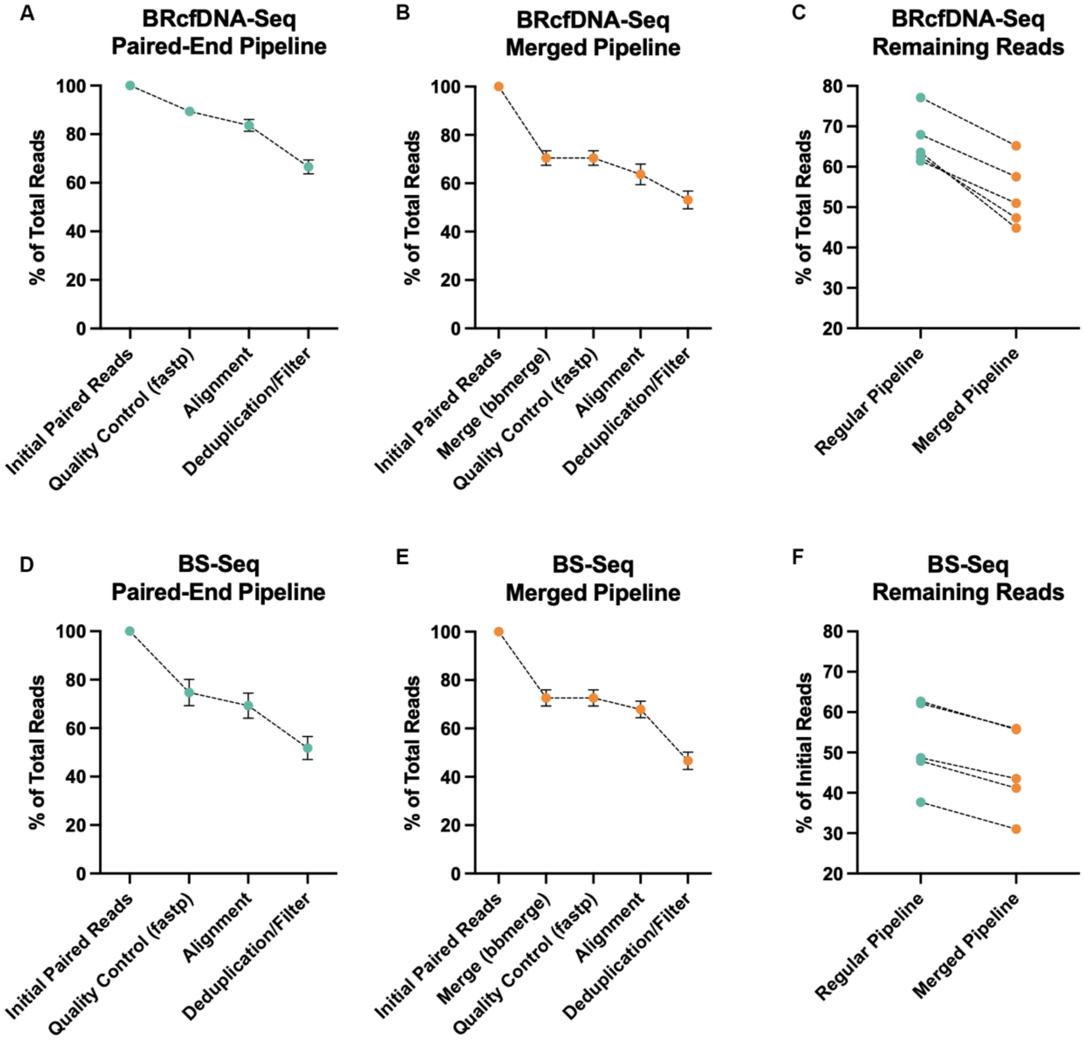

Figure 4.2 The majority of reads are excluded during the initial merging of reads. The percent of total reads is shown after each bioinformatic preprocessing step comparing BRcfDNA-Seq libraries processed using both the Paired-End Reads (A) and Merged Reads (B) protocol. BS-Seq libraries are shown comparing Paired-End (D) and Merged Reads (E) processing. Comparison of final read count of individual samples between Pair-End Reads vs. Merged Read pipelines for BRcfDNA-Seq (C) and BS-Seq (F). Error bars indicate SEM.

4.3.2 The 5mcAdpBS-Seq Protocol Reduces the Inclusion of Degraded DNA into the uscfDNA Region of the Final Library

Since the initial BS-Seq experiments indicated differences in the size-distribution shape compared to non-BS BRcfDNA-Seq, we hypothesized that the apparent elevated 70-130bp region originated from genomic DNA degradation during the bisulfite conversion process(Figure 4.3A). To this end, we tested if attaching single-stranded 5mC protected adapters prior to bisulfite treatment (5mCAdpBS-Seq Protocol) would reduce the incorporation of degraded DNA (Figure 4.3B).

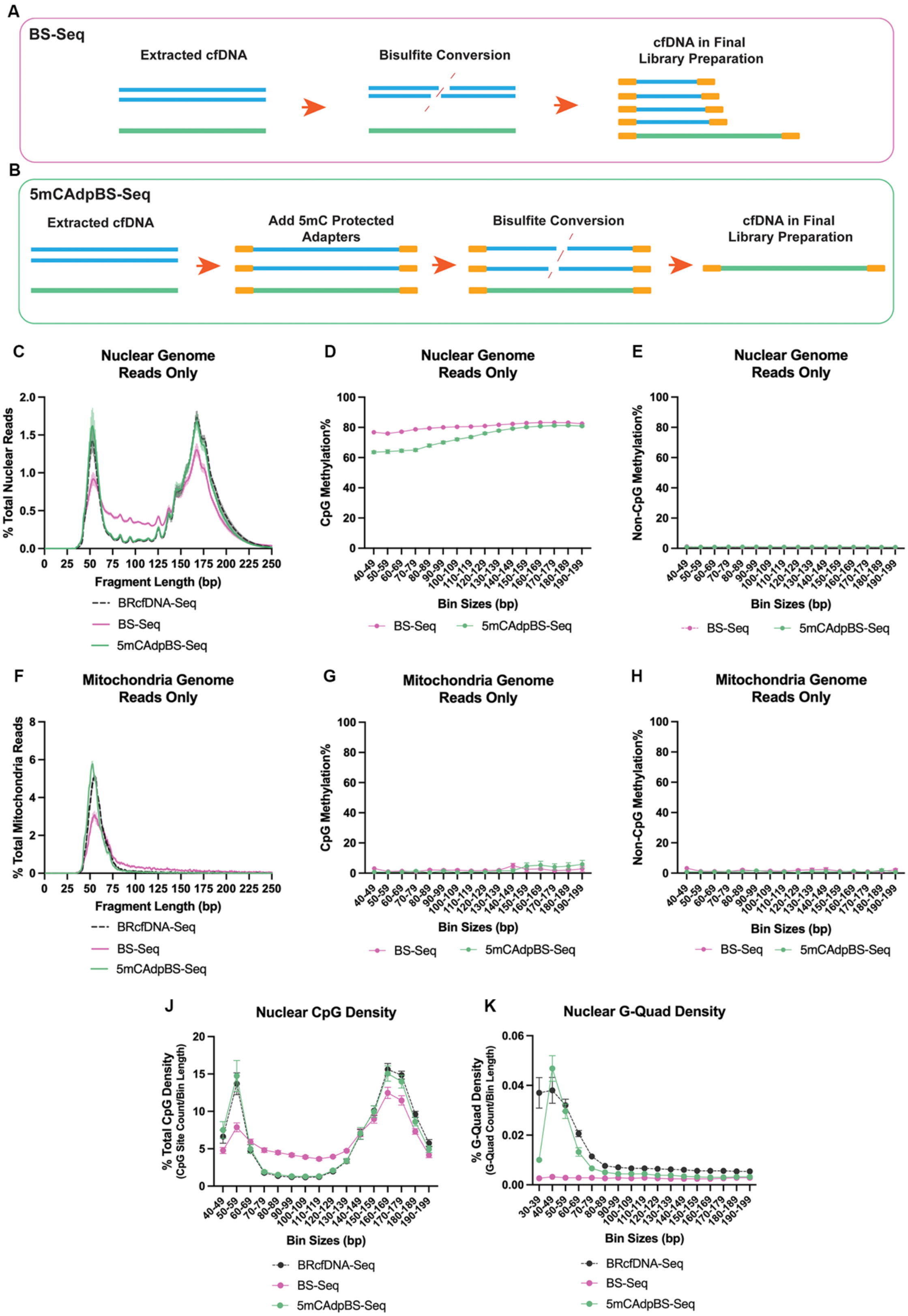
A
BS-Seq
Extracted cfDNA
Bisulfite Conversion
cfDNA in Final Library Preparation
B
5mCAdpBS-Seq
Extracted cfDNA
Add 5mC Protected Adapters
Bisulfite Conversion
cfDNA in Final Library Preparation
C
Nuclear Genome Reads Only
% Total Nuclear Reads
2.0
1.5
1.0
0.5
0.0
0 25 50 75 100 125 150 175 200 225 250
Fragment Length (bp)
BRcfDNA-Seq
BS-Seq
5mCAdpBS-Seq
D
Nuclear Genome Reads Only
CpG Methylation%
100
80
60
40
20
0
40-49 50-59 60-69 70-79 80-89 90-99 100-109 110-119 120-129 130-139 140-149 150-159 160-169 170-179 180-189 190-199
Bin Sizes (bp)
BS-Seq
5mCAdpBS-Seq
E
Nuclear Genome Reads Only
Non-CpG Methylation%
100
80
60
40
20
0
40-49 50-59 60-69 70-79 80-89 90-99 100-109 110-119 120-129 130-139 140-149 150-159 160-169 170-179 180-189 190-199
Bin Sizes (bp)
BS-Seq
5mCAdpBS-Seq
F
Mitochondria Genome Reads Only
% Total Mitochondria Reads
8
6
4
2
0
0 25 50 75 100 125 150 175 200 225 250
Fragment Length (bp)
BRcfDNA-Seq
BS-Seq
5mCAdpBS-Seq
G
Mitochondria Genome Reads Only
CpG Methylation%
100
80
60
40
20
0
40-49 50-59 60-69 70-79 80-89 90-99 100-109 110-119 120-129 130-139 140-149 150-159 160-169 170-179 180-189 190-199
Bin Sizes (bp)
BS-Seq
5mCAdpBS-Seq
H
Mitochondria Genome Reads Only
Non-CpG Methylation%
100
80
60
40
20
0
40-49 50-59 60-69 70-79 80-89 90-99 100-109 110-119 120-129 130-139 140-149 150-159 160-169 170-179 180-189 190-199
Bin Sizes (bp)
BS-Seq
5mCAdpBS-Seq
J
Nuclear CpG Density
% Total CpG Density (CpG Site Count/Bin Length)
20
15
10
5
0
40-49 50-59 60-69 70-79 80-89 90-99 100-109 110-119 120-129 130-139 140-149 150-159 160-169 170-179 180-189 190-199
Bin Sizes (bp)
BRcfDNA-Seq
BS-Seq
5mCAdpBS-Seq
K
Nuclear G-Quad Density
% G-Quad Density (G-Quad Count/Bin Length)
0.06
0.04
0.02
0.00
30-39 40-49 50-59 60-69 70-79 80-89 90-99 100-109 110-119 120-129 130-139 140-149 150-159 160-169 170-179 180-189 190-199
Bin Sizes (bp)
BRcfDNA-Seq
BS-Seq
5mCAdpBS-Seq

Figure 4.3 5mCAdpBS-Seq Protocol Reduces the inclusion of DNA Degradation into the ultrashort region of the Final Library. A) Schematic of routine BS-Seq workflow incorporates degraded cell-free DNA or genomic DNA, which enters the library, potentially masking the uscfDNA methylation signal. B) 5mCAdpBS-Seq protocol in which 5mC premethylated adapters are attached before bisulfite conversion preventing degraded DNA from entering the final library. BRcfDNA-Seq, BS-Seq, and 5mCAdpBS-Seq protocols generate different fragment profiles and CpG, and non-CpG methylation profiles fragments that align to the nuclear genome (C-E) and mitochondria (F-H). CpG and non-CG methylation % profiles were calculated for BS-Seq and 5mCAdpBS-Seq for every 10bp binned reads from 40-200bp. CpG Density (J) and G-Quad% (K) of 5mCAdpBS-Seq of nuclear-aligned reads resemble the BRcfDNA-Seq Profile Compared to BS-Seq Protocol. These plots are calculated from the average of 5 samples undergoing all three protocols. Error bars indicate SEM.

Bioinformatically, we compared the read attrition between the BS-Seq and 5mCAdpBS-Seq. Compared to the BS-Seq, the 5mCAdpBS-Seq protocol, there was a greater read loss in most steps (Merging, Quality Control, and Alignment) (67.8 ± 3.4 % vs. 54.6 ± 5.3% reads remaining) protocol but post-deduplication, the remaining reads between both protocols were comparable (46.6 ± 3.6% vs 45 ± 4.7% reads remaining) (Figure 4.4A). In some individual cases, the % of remaining reads for the 5mCAdpBS-Seq Protocol was higher than BS-Seq protocol (Figure 4.4B)

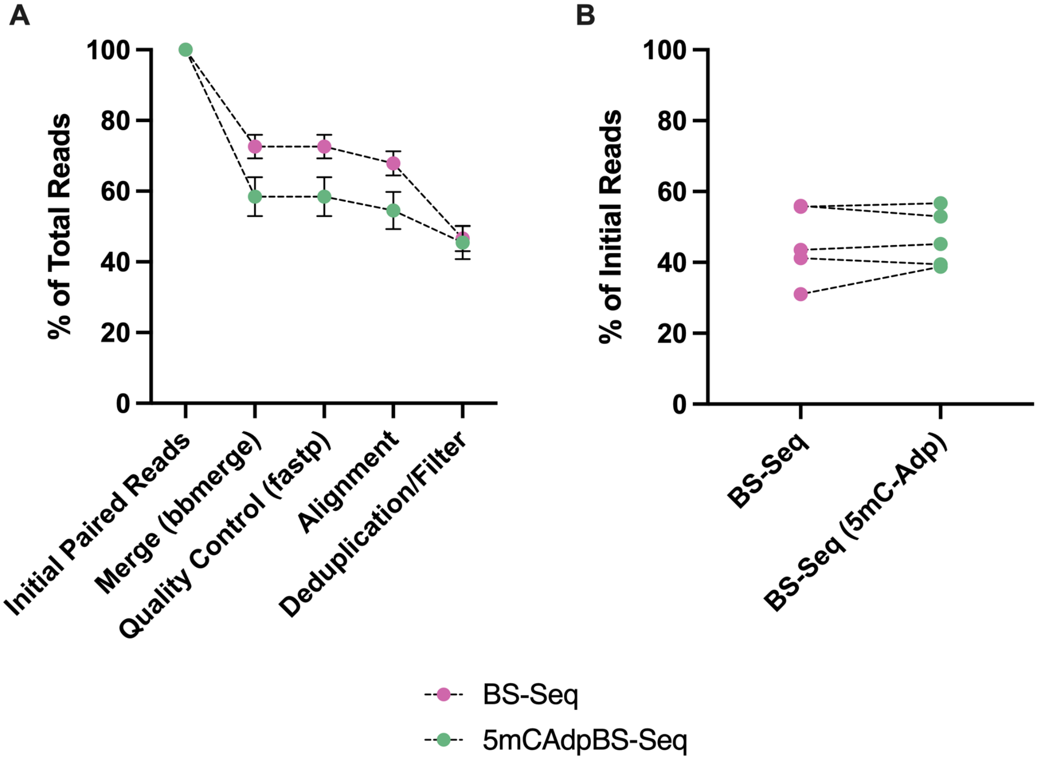

Figure 4.4 Comparative read attrition between BS-Seq and 5mCAdpBS-Seq during bioinformatic processing. A) Comparison of BS-Seq vs. 5mCAdpBS-Seq protocols in their read attrition during each step of the preprocessing pipeline prior to downstream analysis. B) Final remaining reads for individual plasma samples (n=5) that underwent each protocol.

As a negative control, enzymatically sheared Lambda phage DNA was spiked into plasma undergoing both the BS-Seq and 5mCAdpBS-Seq protocol to determine the efficiency of the bisulfite conversion (Figure 4.5). The fragment profile differed amongst the two protocols, with the peak of BS-Seq at ~80bp, whereas the 5mCAdpBS-Seq had a peak at ~60bp. The mean CpG methylation % for both the BS-Seq protocol and 5mCAdpBS-Seq protocols was <1% for CpG% and <1.5% for non-CpG% methylation (Figure 4.5B and C).

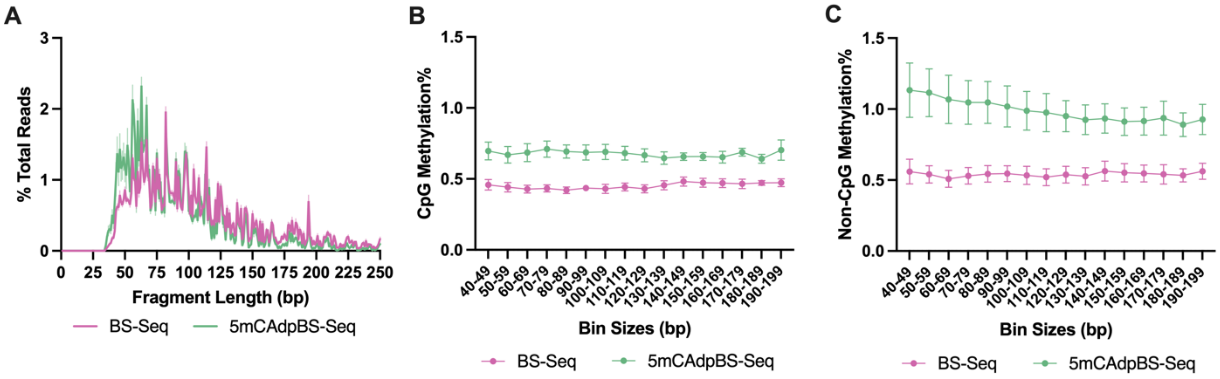

Figure 4.5 Lambda spike-in control indicates the inherent noise of bisulfite conversion methodology. A) Contrasting fragment size profiles of sheared non-methylated lambda DNA processed with BS-Seq or 5mCAdpBS-Seq protocols. CpG (B) and Non-CpG (C) methylation % analysis for 10bp bins ranging from 40-200bp show that the 5mCAdpBS-Seq protocol has a slightly higher methylation % noise. Samples are from five paired samples undergoing both protocols. Vertical lines and error bars indicate SEM from the mean of five subjects.

Reads aligning to nuclear DNA showed substantial differences in the fragment profile between the BS-Seq and 5mCAdpBS-Seq protocols (Figure 4.3C). The fragment profile from the 5mCAdpBS-Seq protocol closely resembled that of the BRcfDNA-Seq protocol. In particular, the region from 70bp to 130bp is largely absent from the DNA degradation apparent in the BS-Seq protocol profiles (Figure 4.3C). The bins amongst the uscfDNA region (40-70bp) demonstrated lower mean CpG methylation% in the 5mCAdpBS-Seq profiles compared to the BS-Seq protocol (63.6-64.6% vs 76.8-77.1%) (Figure 4.3D). In contrast, the bins overlapping the mncfDNA (120-200bp) had a similar CpG methylation% between both protocols (80.2-80.9% vs 80.5-82.5%). In both protocols, the nuclear non-CpG methylation was below 1.5% (Figure 4.3E).

4.3.3 Mitochondria cell-free DNA is Hypomethylated

Alongside nuclear DNA, the mitochondria genome (mitDNA) also contributes to the pool of cell-free DNA in circulation(An et al. 2019). Reads aligning to the mitochondria genome(Figure 4.3 F-H) can be used as a biological control for methylation efficiency since the genome of mitochondria has been described as nearly absent in CpG methylation (B. Liu et al., 2016; Mechta et al., 2017). To this end, we conducted a similar analysis for the reads aligning to the mitochondria DNA. We observed that the fragment patterns of 5mCAdpBS-Seq closely resembled the BRcfDNA-Seq pattern with a slight peak shift to the left (Figure 4.3F). Comparatively, the BS-Seq mitDNA profile had a peak at 57bp, with most fragments occupying the 40 to 75bp region. Compared to the nuclear uscfDNA, the mitDNA fragment curve is not symmetrical, with a larger shoulder from 60 to 75bp. For both protocols, the CpG and the non-CpG methylation were below <5%, with fluctuations beginning for bins >130bp (Figure 4.3G&H). As per (Figure 4.3F), there were minimal reads beyond 150bp).

4.3.4 Genomic Characteristics of the 5mCAdpBS-Seq Protocol Closely Resembles BRcfDNA-Seq

We examined the pattern of CpG density at each binned size and observed that the 5mCAdpBS-Seq protocol followed the same pattern as the BRcfDNA-Seq (Figure 4.3J), whereas the BS-Seq protocol had a lower peak at 50nt but an elevated CpG density from 70-130bp. This pattern resembled that of the fragment size distribution (Figure 4.3A).

Previous reports have shown that the uscfDNA are enriched in G-rich sequences that have the potential to form G-Quad secondary structures (Hudecova et al., 2021). G-Quad signatures were enriched in the BRcfDNA-Seq and 5mCAdpBS-Seq protocol with an observed peak at the 40-49bp bin (Figure 4.3K). However, in the same samples processed with BS-Seq, this G-Quad enrichment was absent in the ultrashort region. Based on these, the closer resemblance to the characteristics of BRcfDNA-Seq, subsequent analysis was performed using the 5mCAdpBS-Seq Protocol.

4.3.5 uscfDNA Map to Different Regions Compared to mncfDNA

Karyograms for samples that underwent the 5mCAdpBS-Seq protocol indicated differences in CpG density chromosome regions for uscfDNA and mncfDNA (Figure 4.6A). When comparing CpG site positions, 41.4 ± 5% of uscfDNA sites were common with mncfDNA (Figure 4.6B). We determined what profile of genomic element categories were associated with the fragments containing CpG sites in 5mCAdpBS-Seq (Figure 4.6C). Both mncfDNA and uscfDNA reads had a major proportion of CpG-site containing reads mapping to intron and intergenic regions. Comparatively, uscfDNA fragments appear to be significantly enriched in promoters, exons, and CpG island locations, whereas mncfDNA seem more enriched in SINE and intergenic regions.

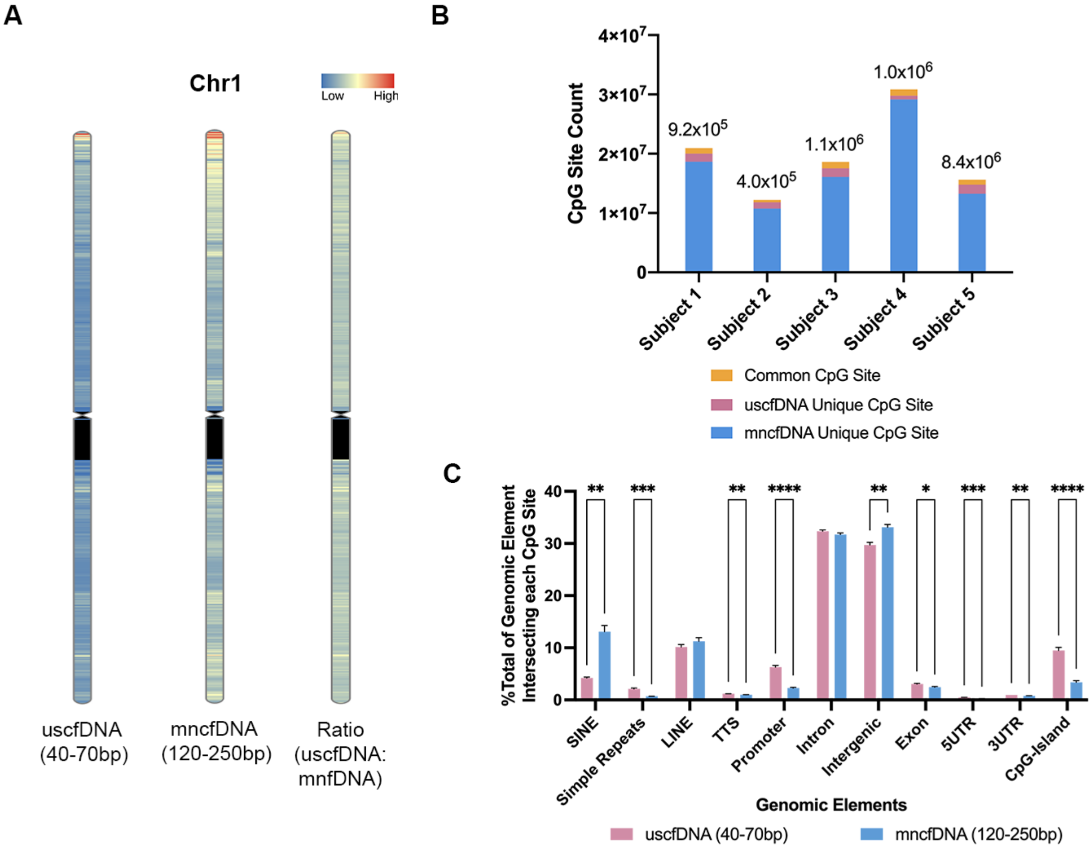

Figure 4.6 CpG positions of uscfDNA differ from mncfDNA. A) Karyograms averaged from 5 non-cancer subjects of normalized CpG density plots for 1 million bp-sized bins across chromosome 1 for the uscfDNA and mncfDNA. The ratio is calculated by dividing the mean uscfDNA %coverage by mncfDNA %coverage. B) Intra-sample count of common and unique CpG site counts between cfDNA populations. Values above bars indicate the count of common CpG sites. C) Composition of different genomic elements where CpG-sites intersected were compared between those with an uscfDNA (40-70bp) or mncfDNA (120-25bp) fragment size. SINE: short interspersed nuclear element, LINE: long interspersed nuclear element, TTS: transcription termination site, 5'UTR: 5' untranslated region, 3'UTR: 3' untranslated region. Data represents SEM and the mean of 5 paired non-cancer subjects processed with 5mcAdpBS-Seq. Stars indicate unadjusted p-values with * p <0.05, ** p <0.01 , *** p< 0.001, and **** p <0.0001.

4.3.6 uscfDNA Map to Regions Associated with Active Gene Activity Compared to mncfDNA

Since uscfDNA regions were hypomethylated by ~15% compared to mncfDNA (Figure 4.3D), we hypothesized the locations where uscfDNA mapped to were enriched in genomic regions associated with increased gene activity (Figure 4.7A) (Zhang et al., 2015). Epigenetic marks, including histone modifications, hypomethylation, and hypermethylation regions from publicly available chromatin immunohistoprecipitation (ChIP-Seq) and whole genome bisulfite sequencing datasets in select blood-related cells (monocyte, macrophage, eosinophil, and neutrophils) were intersected with the mapping regions of uscfDNA and mncfDNA generated by 5mCAdpBS-Seq to assess overlap (Figure 4.7B). We observed that uscfDNA demonstrated higher intersection % (versus a matched shuffled position control) with active gene epigenetic marks (H3K4m1, H3K4m3, H3K27ac modifications, and the hypomethylated regions), whereas mncfDNA showed an opposite trend. In contrast, for H3K27me, both uscfDNA and mncfDNA showed an increased intersection % in H3K27me, H3K9me3, and hypermethylated regions (Figure 4.7C).

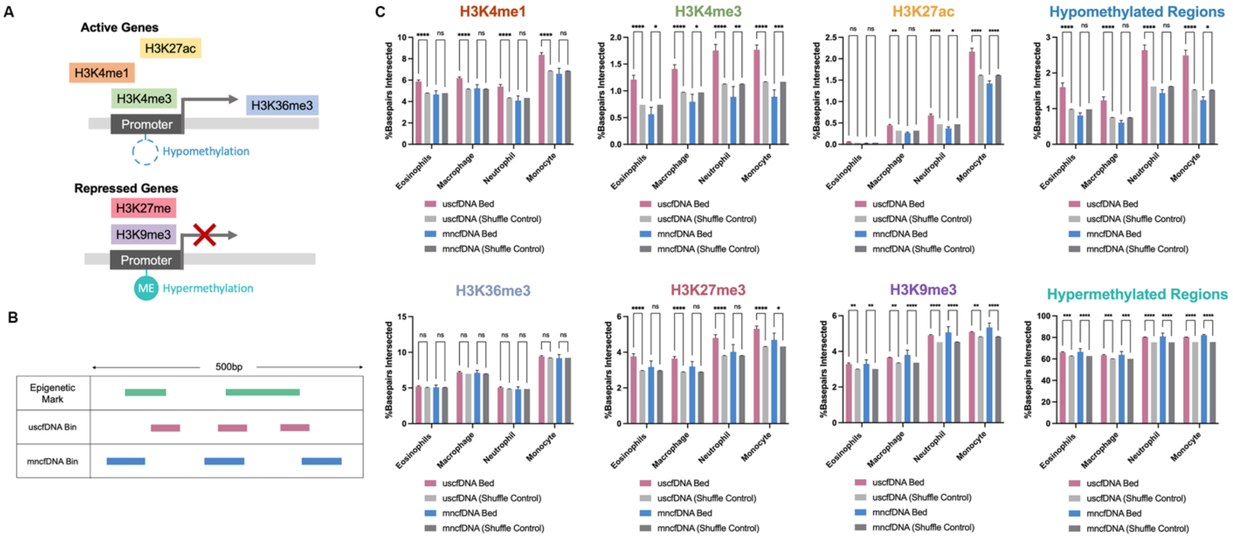

Figure 4.7 Compared to mncfDNA, a higher proportion of uscfDNA fragments intersect with epigenetic marks related to active gene activity. A) Genomic regions derived from epigenetic marks, including methylation patterns and histone modifications, are associated with active or repressed gene activity. B) Schematic of intersection methodology to determine where uscfDNA or mncfDNA bps overlap with epigenetic mark regions from reference .bed of bisulfite conversion or CHIP-seq experiment files. C) % of intersecting bps for each epigenetic mark for uscfDNA and mncfDNA bins. Randomly shuffled bed files were generated for each sample to act as a control for intersection locations. Errors bars represent SEM from the mean of 5 paired plasma samples that underwent 5mCAdpBS-Seq protocol. Non-paired multiple paired t-tests were performed after two-way ANOVA. Data represents SEM and the mean of 5 paired non-cancer and 4 NSCLC plasma subjects. Stars indicate unadjusted p-values with * p <0.05, ** p <0.01 , *** p< 0.001, and **** p <0.0001. Only %bp intersected comparisons between uscfDNA and mncfDNA are presented.

4.3.7 CpG Methylation Levels in uscfDNA are Lower Compared to mncfDNA with Differing Patterns for Genomic Elements

To further examine the methylation behavior of each genomic element, the average CpG methylation % profile for each genome element was plotted from 5000bp upstream from the center of the element to 5000bp downstream from the center of the element (Figure 4.8). In general, the CpG methylation % of uscfDNA fragments was 10-20% lower than that for mncfDNA over the same regions reflecting the observations genome-wide (Figure 4.3D). The general patterns of the CpG% methylation distribution were similar, although uscfDNA demonstrated more dynamic behavior, most likely caused by reduced coverage. For the promoter element, the uscfDNA demonstrated a wider U-shape compared to the V-shape for the mncfDNA. The three most distinct methylation patterns between the two cfDNA populations were those for Simple Repeats, LINE, Intergenic, and Exons.

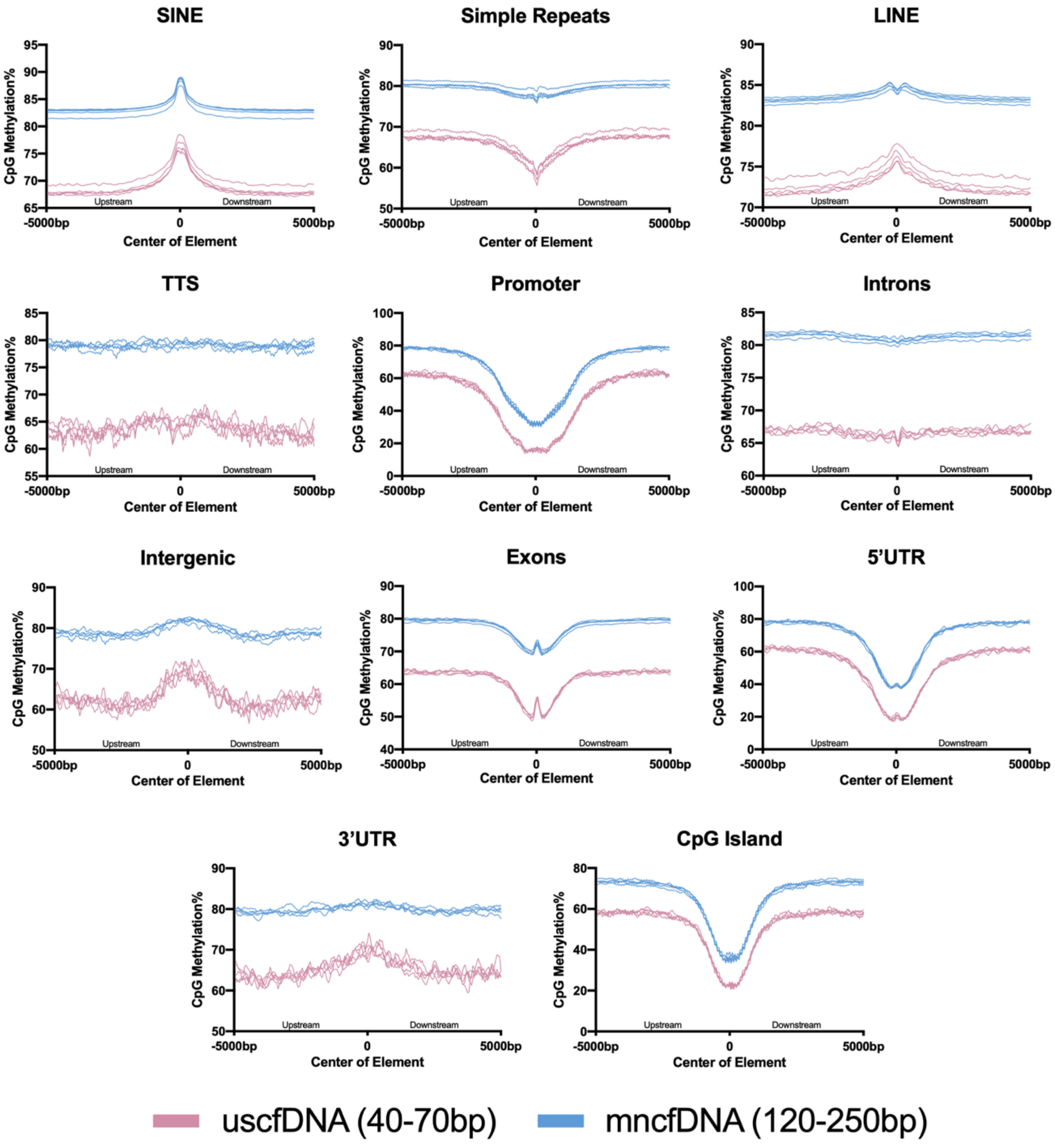

Figure 4.8 CpG Methylation patterns differ between uscfDNA and mncfDNA fragments. The average CpG methylation % patterns from 5000bp upstream and 5000bp downstream from the center of the element for uscfDNA and mncfDNA sized reads. Lines show five separate non-cancer samples processed with the 5mCAdpBS-Seq protocol.

Since there was a minor overlap between uscfDNA and mncfDNA CpG sites, we aggregated the .bam files from the uscfDNA and mncfDNA from the five subjects and analyzed the two cfDNA populations for differentially methylated regions (Figure 4.9A). Sixty-eight significant DMRs were found, where the majority were hypomethylated in uscfDNA compared to mncfDNA.

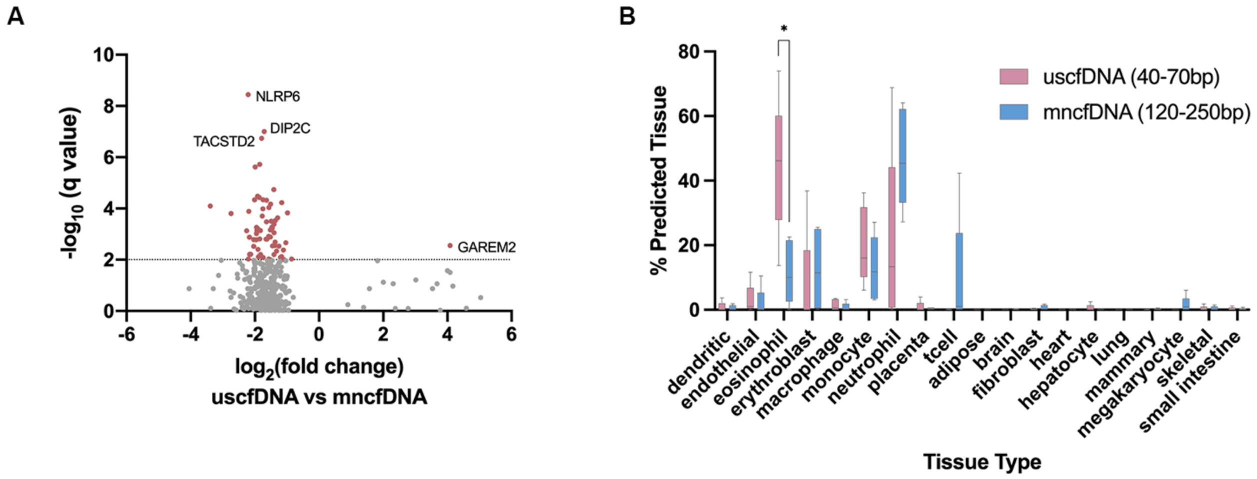

Figure 4.9 Differentially methylated regions show differences in genes and cell of origin. A) Differentially methylated region analysis between merged uscfDNA and mncfDNA .bam files from five samples show 68 significant DMRs (q-value <0.01) and the nearest downstream gene. Only candidates with a q-value <1.0 are shown. B) Box and whisker plots of CelFie deconvolution prediction of blood cell tissue of origin signal from the methylation patterns in the uscfDNA and mncfDNA. Prediction reveals that uscfDNA and mncfDNA are derived from tissues of blood cell origin. Non-paired multiple paired t-tests were performed comparing the % contribution of cell type between uscfDNA and mncfDNA. Stars represent unadjusted p-values with * p <0.05. Errors bars show min and max from five non-cancer samples that underwent 5mCAdpBS-Seq protocol.

4.3.8 Deconvolution Suggests that uscfDNA Mainly Derives from Peripheral Blood Cells

We attempted to deconvolute the fragments from the uscfDNA and mncfDNA populations into their cell/tissue-of-origin using the CpG methylation patterns (Figure 4.9B). Using the CelFie algorithm (Caggiano et al., 2021), which was designed to deconvolute signal from low input cfDNA samples, we confirmed that the major tissue of origin for both uscfDNA and mncfDNA is blood, as expected. Evidence of blood cell contribution included eosinophils, erythroblasts, monocytes, neutrophils, and T-cells. Comparatively, CelFie indicated that uscfDNA has elevated eosinophiles composition compared to the mncfDNA.

4.3.9 uscfDNA CpG Mapping Patterns and Methylation Characteristics Discriminates Non-cancer Subjects From in Late-stage NSCLC

As a proof of concept, we examined if the methylation profile of uscfDNA would be an effective biomarker for cancer detection. To that end, we processed four late-stage NSCLC samples with the 5mCAdpBS-Seq protocol and compared them with non-cancer samples. The global fragment patterns showed an elevated uscfDNA peak in the NSCLC samples and a lower rightward shoulder in the mncfDNA regions of 175 to 200bp (Figure 4.10A). For reads mapping to the nuclear genome, in the bins below 140bp, it appeared that the NSCLC samples had a 4-6% increase in CpG% methylation compared to the non-cancer samples in sizes below 140bp (Figure 4.10B).

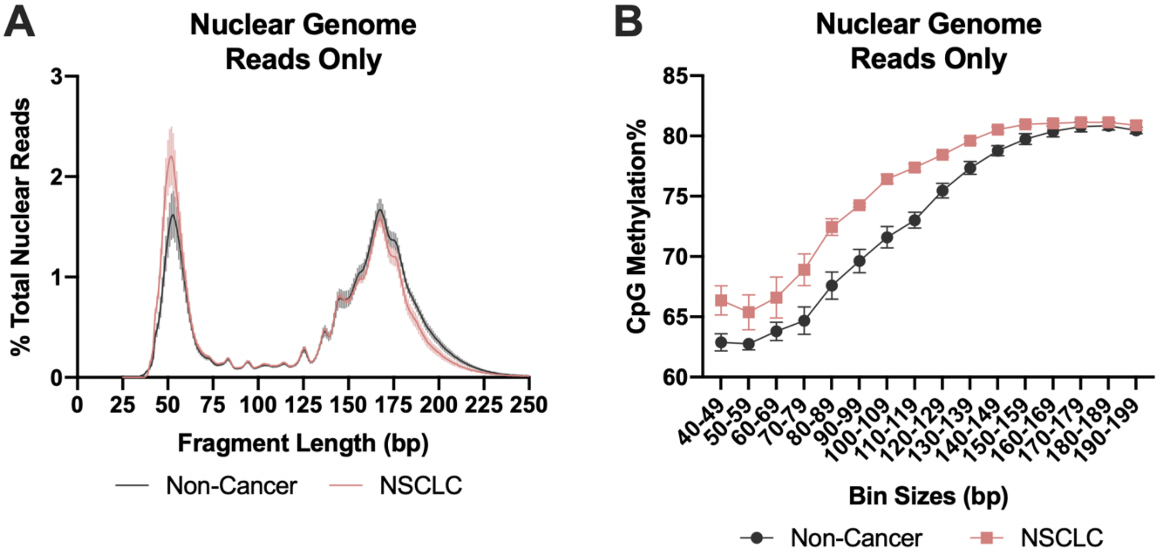

Figure 4.10 Genomic and Methylation Profiles Differ Between Non-Cancer and NSCLC Samples Processed by 5mCAdpBS-Seq. A) Fragment size distribution profile comparing non-caner and NSCLC cohorts. B) CpG Methylation of % of uscfDNA region of NSCLC samples are elevated compared to non-cancer Samples. Comparative CpG % profiles were calculated for every 10bp binned reads from 40-200bp for the nuclear genome. Plots represent the mean from 5 paired non-cancer plasma and 4 NSCLC, which underwent 5mCAdpBS-Seq protocol.

In NSCLC samples versus non-cancer, we observed that the composition profile of genomic elements intersecting with CpG-containing fragments demonstrated more elements with significantly altered proportions in the uscfDNA (Figure 4.11A) bins compared to the mncfDNA (Figure 7.11B) bins (8 vs 4). In the uscfDNA bin, there were significant changes in the proportion of SINE, Simple Repeats, promoters, introns, intergenic, 5'UTR, and CpG-Islands. In the mncfDNA, however, promoters, exons, 5'UTR, and CpG-island proportion appeared statistically different.

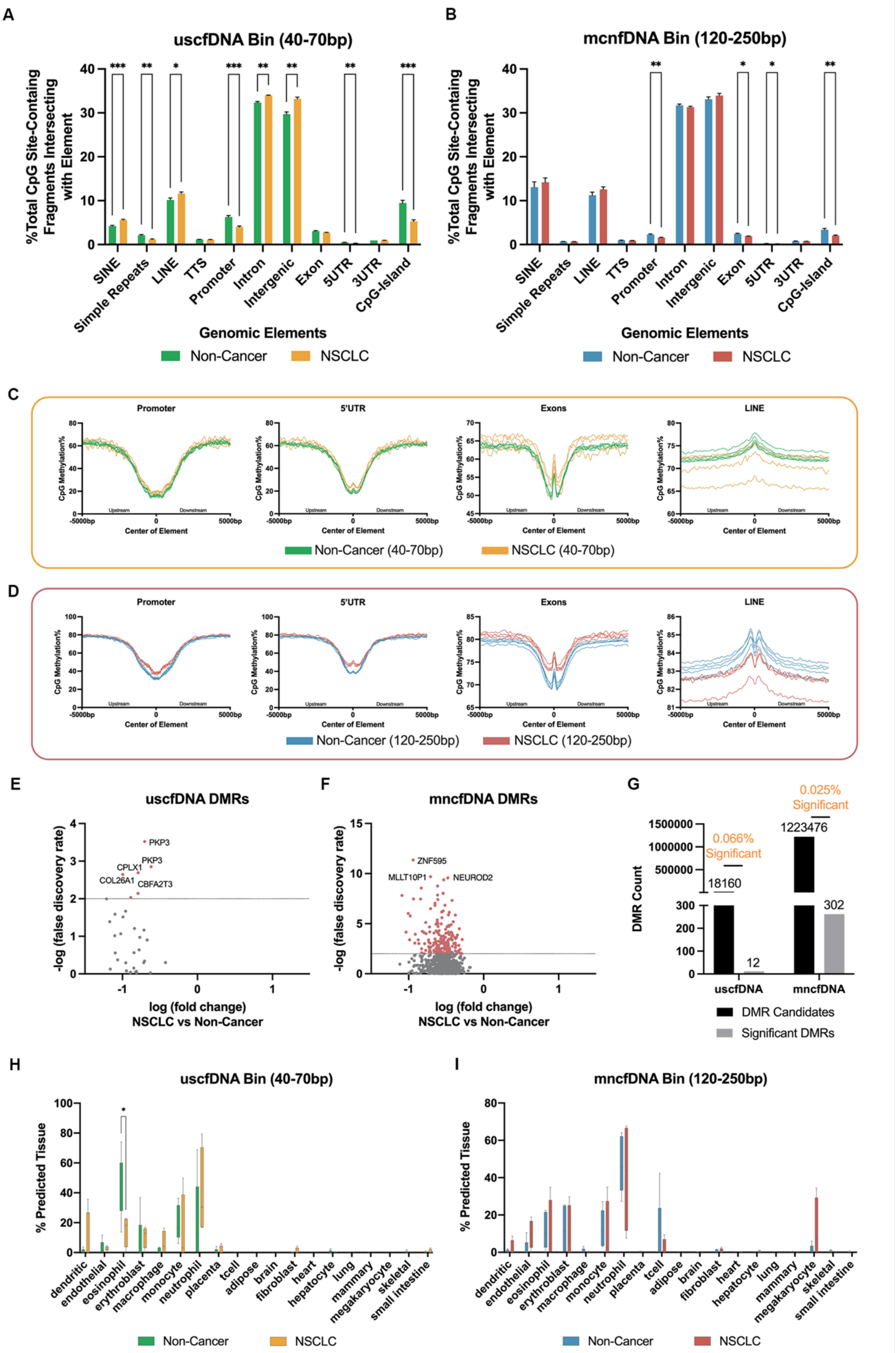
A
uscfDNA Bin (40-70bp)
%Total CpG Site-Containg Fragments Intersecting with Element
0
10
20
30
40
***
**
*
***
**
**
**
***
SINE
Simple Repeats
LINE
TTS
Promoter
Intron
Intergenic
Exon
5UTR
3UTR
CpG-Island
Genomic Elements
Non-Cancer
NSCLC
B
mcnfDNA Bin (120-250bp)
%Total CpG Site-Containg Fragments Intersecting with Element
0
10
20
30
40
**
*
*
**
SINE
Simple Repeats
LINE
TTS
Promoter
Intron
Intergenic
Exon
5UTR
3UTR
CpG-Island
Genomic Elements
Non-Cancer
NSCLC
C
Promoter
5'UTR
Exons
LINE
CpG Methylation%
Upstream
Downstream
-5000bp
0
5000bp
Center of Element
Non-Cancer (40-70bp)
NSCLC (40-70bp)
D
Promoter
5'UTR
Exons
LINE
CpG Methylation%
Upstream
Downstream
-5000bp
0
5000bp
Center of Element
Non-Cancer (120-250bp)
NSCLC (120-250bp)
E
uscfDNA DMRs
-log (false discovery rate)
PKP3
PKP3
CPLX1
COL26A1
CBFA2T3
log (fold change)
NSCLC vs Non-Cancer
F
mncfDNA DMRs
-log (false discovery rate)
ZNF595
MLLT10P1
NEUROD2
log (fold change)
NSCLC vs Non-Cancer
G
DMR Count
0.066%
Significant
0.025%
Significant
18160
12
1223476
302
uscfDNA
mncfDNA
DMR Candidates
Significant DMRs
H
uscfDNA Bin (40-70bp)
% Predicted Tissue
*
dendritic
endothelial
eosinophil
erythroblast
macrophage
monocyte
neutrophil
placenta
tcell
adipose
brain
fibroblast
heart
hepatocyte
lung
mammary
megakaryocyte
skeletal
small intestine
Non-Cancer
NSCLC
I
mncfDNA Bin (120-250bp)
% Predicted Tissue
dendritic
endothelial
eosinophil
erythroblast
macrophage
monocyte
neutrophil
placenta
tcell
adipose
brain
fibroblast
heart
hepatocyte
lung
mammary
megakaryocyte
skeletal
small intestine
Non-Cancer
NSCLC

Figure 4.11 CpG coverage and methylation patterns differ between non-cancer and NSCLC samples. The composition of different genomic element category locations where CpG-site containing reads aligned were compared between the NSCLC and Non-Cancer samples for uscfDNA (A) and mncfDNA (B). The average CpG methylation % patterns from 5000bp upstream and 5000bp downstream the center of the element are plotted for Promoters, 5UTR, exons, and LINE for uscfDNA (C) and mncfDNA (D) sized reads. Differentially methylated region analysis between merged uscfDNA and mncfDNA .bam files from 5 non-cancer subjects and 4 NSCLC subjects reveal significant DMRs in the uscfDNA(E) and mncfDNA(F) bin (q-value <0.01, only candidates with q-value <1.0 are shown). G) uscfDNA has a higher proportion of significant DMRs compared to the mncfDNA. Box and whiskers plot of CelFie deconvolution algorithm suggests changes in cell type composition between non-cancer and NSCLC samples of the uscfDNA (H) and the mncfDNA (I) sized bins. For the CelFIE deconvolution, error bars represent min and max positions with individual samples. Non-paired t-test was performed. Data represents SEM and the mean of 5 paired non-cancer and 4 NSCLC plasma subjects. Stars indicate unadjusted p-values are presented with * p <0.05, ** p <0.01, *** p< 0.001, and **** p <0.0001.

4.3.10 Methylation Pattern of Genomic Elements are Altered in NSCLC

When we examined the CpG methylation % patterns for uscfDNA (Figure 4.11C) and mncfDNA bins (Figure 4.11D), the NSCLC samples showed greater methylation variability in the cancer samples, whereas the non-cancer samples were more uniform. For promoters, 5'UTR, exons, and CpG Islands, the methylation towards the center of the element (position 0) were hypermethylated in NSCLC samples compared to the non-cancer, but this observation was more evident in the mncfDNA bins. For simple repeats, the NSCLC showed a flattened curve compared to the more V-shape of the non-cancer samples, and it was more distinct in the uscfDNA bins. For the LINE elements, the NSCLC samples appeared more hypomethylated compared to the non-cancer subjects, but this was more apparent in the mncfDNA bin. Introns, 3'UTR, and TTS elements were globally more hypermethylated in mncfDNA. In the uscfDNA, the methylation profile of the non-cancer samples was more uniform compared to the highly variable NSCLC traces. The remaining SINE and intergenic elements did not demonstrate any remarkable pattern for NSCLC versus non-cancer.

4.3.11 Differentially Methylated Regions and Deconvolution Are Potential Biomarkers for NSCLC Detection

DMR analysis between the CpG% methylation of NSCLC and non-cancer samples revealed that both the uscfDNA bin and the mncfDNA fragment bin had discoverable DMR candidates (Figure 4.11E and F). The uscfDNA bin had 12 significant DMRs out of 18160 tested regions (0.066% significant) compared to mncfDNA, which had 302 significant DMRs out of 1223476 tested regions (0.025% significant) (Figure 4.11G). For both uscfDNA and mncfDNA, significant DMRs demonstrated a decrease in methylation fold-change in NSCLC compared to non-cancer subjects (Figure 4.11E and F). For uscfDNA DMR candidates, the nearest gene was plakophilin3 (PKP3), complexin 1 (CPLX1), and collagen type XXVI alpha 1 chain (COL26A1). For mncfDNA, the top candidates were zinc finger protein 595 (ZNF595), myeloid/lymphoid or mixed-lineage leukemia translocated to pseudogene 1 (MLLT10P1), and neuronal differentiation 2 (NEUROD2).

Using the CelFie deconvolution prediction algorithm suggested differences in the tissue of origin profiles between the two cohorts (Figure 4.11H&I). In the uscfDNA fragment bin, the eosinophil signal appeared significantly decreased in NSCLC samples. Whereas in mncfDNA, there was an increase in megakaryocyte signal in some NSCLC samples (not significant).

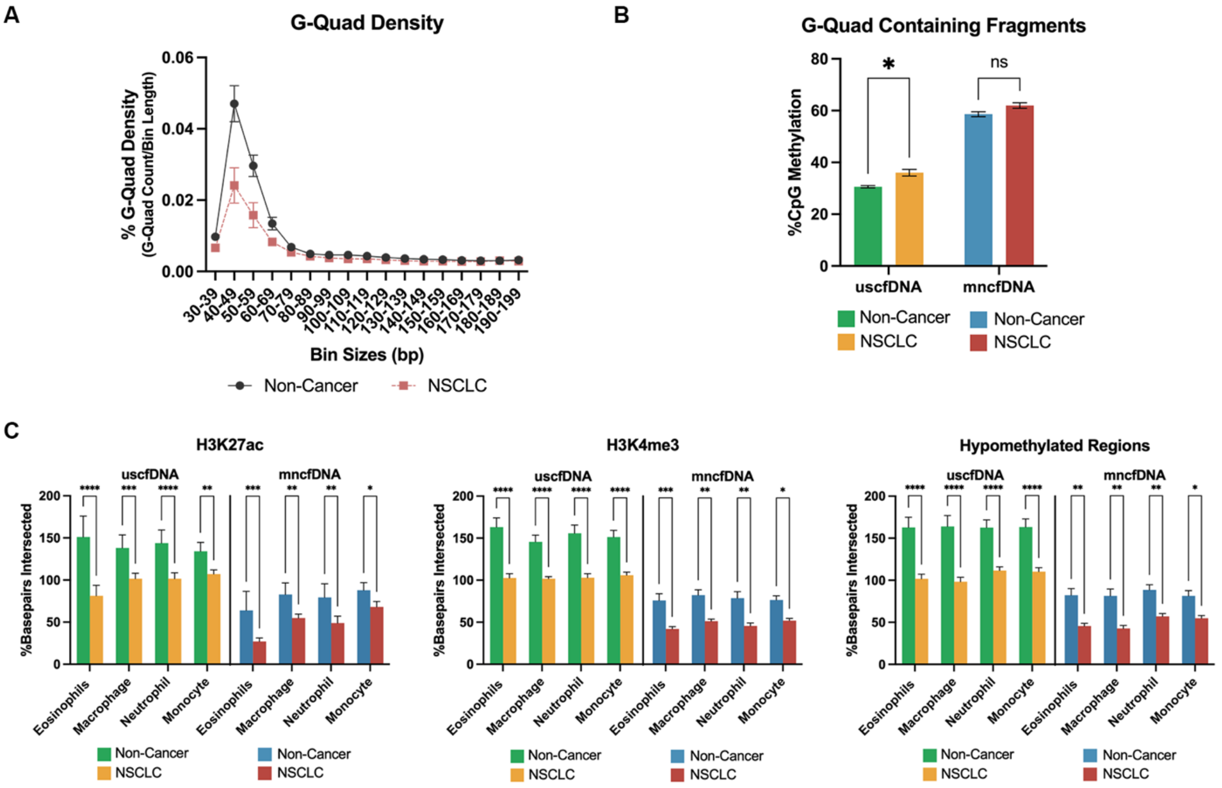

Figure 4.12 G-Quad methylation% and epigenetic mark overlap% are potential NSCLC biomarkers. A) G-Quad density is decreased in the uscfDNA regions (40-70bp) in NSCLC. B) CpG methylation % significantly increased in G-Quad-containing fragments in uscfDNA. C) Normalized % of intersecting bps for three epigenetic marks (H3K2ac, H3k4me3, and hypomethylated regions) are decreased in NSCLC samples in both uscfDNA and mncfDNA bins. % intersection was normalized to control shuffled bed files. Data represents SEM and the mean of 5 paired non-cancer and 4 NSCLC plasma subjects. Non-paired t-test was performed after ANOVA. Stars indicate unadjusted p-values with * p <0.05, ** p <0.01, and *** p< 0.001. Stars indicate unadjusted p-values are presented with * p <0.05, ** p <0.01, *** p< 0.001, and **** p <0.0001.

4.3.12 G-Quad Containing uscfDNA Fragments Shows an Increased CpG methylation % in NSCLC Samples

NSCLC samples demonstrated a decreased G-Quad signature % in the uscfDNA region compared to non-cancer samples (Figure 4.12A). When the G-Quad-containing fragments were filtered out and analyzed for CpG methylation %, NSCLS samples were observed to be hypermethylated compared to non-cancer samples (Figure 4.12B).

4.3.13 uscfDNA Overlapping Patterns with Cell Type-Specific Epigenetic Marks is Altered in NSCLC

We next examined if the normalized % of intersecting base pairs for epigenetic marks were altered in NSCLC. Three epigenetic marks (H3K27ac, H3K4me3, and hypomethylated regions) significantly decreased in NSCLC samples in uscfDNA and mncfDNA fractions (Figure 4.3.11C).

4.4 Discussion

In this chapter, we describe an optimized library preparation protocol for cfDNA in which single-stranded 5mC premethylated adapters are ligated to heat-denatured DNA fragments prior to bisulfite conversion and sequencing (5mCAdpBS-Seq). This method improves the accuracy of downstream analysis by preventing bisulfite conversion degraded DNA from being incorporated into the final library and masking the methylation signal of uscfDNA. For the first time, we observe using the 5mCAdpBS-Seq protocol the CpG methylation% of uscfDNA is approximately 60% compared to the 70-80% of mncfDNA. The unique methylated patterns are suggestive that uscfDNA could originate through a different mechanism compared to mncfDNA, which is worth further exploring from both a biological or biomarker perspective.

For uscfDNA, the use of a single-stranded DNA library generates two sequenced reads from an inherently single-stranded template (either ssDNA or denatured dsDNA). Bioinformatically, merging the forward and reverse reads retains the 5' to 3' orientation (the R1), representing the characteristics of the original fragment (Troll et al., 2019). The merging protocol used requires a minimum of 9bp overlap between two reads which translates to the inclusion of only merged reads which are shorter than 291bp from 2x150bp sequencing. Since our interest is primarily in mncfDNA and shorter fragments, this methodology is justified (Sanchez et al., 2018). Discarding reads larger than 291bp (dinucleosomal or trinucleosomal cfDNA, which cannot find sufficient overlap) would logically lead to a reduction in total reads for both BRcfDNA-Seq (66.5 ± 6.4% vs. 51.8 ± 10.6% remaining reads) and BS-Seq (53.2 ± 8.2% vs. 45.5 ± 10.5% remaining reads) (Figure 4.2).

For BS-Seq libraries, the pair-end protocol generated a two-peaked mononucleosomal profile that became absent when the merged reads protocol was implemented (Figure 4.1D). One explanation is during the paired-end preprocessing protocol, there are situations where only one read was generated during sequencing. An accumulation of full-length orphan reads sized at 150bp could explain the spike at 150bp (Figure 4.1D). In the merged-reads protocol, this preliminary stringent approach would filter for fragments of high confidence since both paired reads must match to proceed with alignment.

Although the merged reads protocol “repaired” the double-peak distribution in the mncfDNA region, bisulfite conversion still inadvertently causes DNA degradation (K. Tanaka & Okamoto, 2007). Longer reaction times and high temperatures are associated with greater DNA degradation, while low temperatures and short incubation times could lead to incomplete conversion (Grunau et al., 2001; Raizis et al., 1995). Hence optimization of the temperature and time is required to fully capture the cfDNA methylation profile. Although an earlier report claimed that cfDNA undergoes minimal degradation during bisulfite conversion(Werner et al., 2019), those investigators likely did not realize there visualize fragments shorter than 100bp.

Therefore, without the optimized purification or visualizing methods for uscfDNA, they would not have been able to observe the accumulation of DNA decay at 70-120bp, as reported in this current study.

Other alternatives to bisulfite conversion include enzymatic conversion, which promises lower DNA degradation and improved methylation yield. It is, however, time-consuming, and reports have shown that the conversion efficiency does not compare to bisulfite conversion (Zheng et al., 2022). We conducted preliminary experiments with the enzymatic conversion protocols, but they did not generate libraries of sufficient quality to proceed with sequencing (Figure 4.13). Enzymatic conversion kits are also not optimized for ssDNA (Vaisvila et al., 2021) since the initial ten-eleven translocation2 (TET2) oxidation step has a preference for dsDNA compared to ssDNA and RNA (Leddin & Cisneros, 2019). In order to evaluate its applicability to uscfDNA, this conversion method would need further optimization (DeNizio et al., 2019).

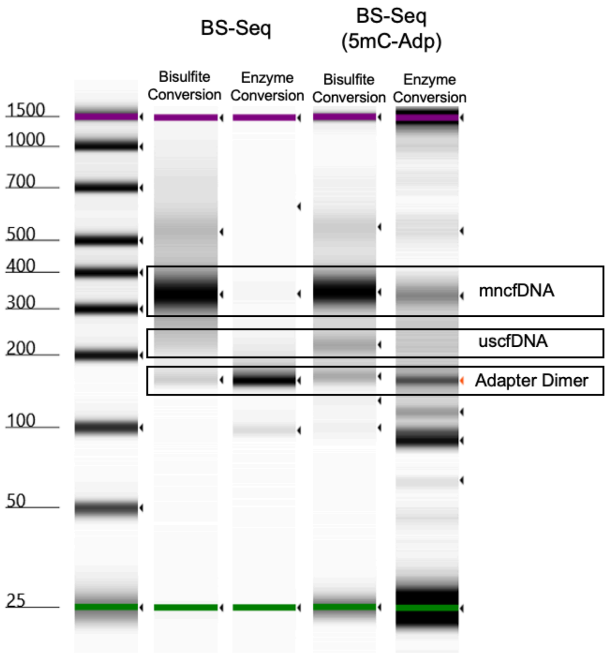

Figure 4.13 Enzymatic conversion of 5mC did not generate sufficient libraries. Electrophoresis gel of the comparison of bisulfite and enzyme conversion protocols for extracted cell-free DNA from 2mL of non-cancer plasma after single-stranded library preparation. BS-Seq and 5mCAdpBS-Seq protocols are shown. The enzyme conversion protocol generates libraries with only adapter dimers or cell-free DNA-sized bands with low concentrations.

Higher molecular weight cfDNA has been documented to be more susceptible to bisulfite degradation compared to mncfDNA (Werner et al., 2019). Therefore, during the BS-Seq protocol, the observed degraded DNA likely originated from these larger pieces of cfDNA. The CpG residues of genomic DNA are reportedly 70-80% hypermethylated (Strichman-Almashanu et al., 2002), and both sources of degraded fragments (either genomic DNA or high molecular weight DNA, which also derives from apoptosis) would still be expected to carry these characteristics. This would explain why during the BS-Seq protocol, the "bleeding" of the degraded DNA into the 40-100 bp fraction skewed the average of CpG methylation% towards higher levels (closer to 80%). Resultingly, the %CpG methylation of uscfDNA was measured in this study for the first time, and it appears to be approximately 60%.

Visually, the 5mCAdpBS-Seq protocol reduced the amount of degraded DNA from entering the final library in the 70-120bp region (Figure 4.3C). However, based on the global reduction in CpG methylation%, the degraded DNA likely ceased to occupy all regions lower than 180bp, including the ultrashort region. This resulted in fragment curves that more closely resembled the BRcfDNA-Seq libraries, where no degradation is expected to occur (Figure 4.3C). Other aspects, such as CpG Density and %G-Quad signatures, closely reflected the pattern from BRcfDNA-Seq libraries(Figure 4.3J and K). In contrast, under the BS-Seq protocol, CpG Density and G-Quad signatures were far less apparent. Therefore, the use of single-stranded 5mC-protected adapters prior to sequencing provides a more accurate portrayal of the native characteristics of not only uscfDNA but any cfDNA with a footprint shorter than 180bp.

We implemented both a technical and biological control to determine the conversion efficiency of the BS-Seq and 5mCAdpBS-Seq protocols. Using unmethylated non-human lambda spike-in, the inherent noise was shown to be <1% and <1.5% for CpG and non-CpG methylation (Figure 4.5). Interestingly, there was a slight increase in cytosine methylation levels in the 5mCAdpBS-Seq protocol for the digested lambda reads (Figure 4.5B&C). Similar to the nuclear DNA, a higher proportion of reads was found at >70bp for BS-Seq, suggesting that DNA in these size ranges was being included. All experiments should use CpG methylation of lambda as a quality control of bisulfite conversion efficiency.

Since the mitochondrial genome has been described to contain low or absent CpG% methylation (B. Liu et al., 2016; Mechta et al., 2017) it can act as a biological internal negative control for the 5mCAdpBS-Seq Protocol. We observed low levels <2% of both CpG and non-

CpG methylation in mitochondria cfDNA in size bins with most reads (30-75bp), suggesting our workflow did not artificially over-represent methylation levels. There was a pattern of increasing methylation variability in fragments in bins >150bp, potentially due to the lower number of reads in this footprint (Figure 4.3G and H). The mitochondrial DNA also reflected the accuracy of the 5mCAdpBS-Seq protocol as the fragment length histogram for mitochondria generated by 5mCAdpBS-Seq fragment profiles resembled BRcfDNA-Seq. BS-Seq had fragments contributing to a larger area under the curve for regions from 75 to 175bp, which could be the result of larger DNA decay artifacts.

In general, for fragments aligning to the human genome, we observed that as cfDNA fragments decreased in size, so did the CpG methylation% of the respective size bin (Figure 4.3D). With the exception of neurons and stem cells, non-CpG methylation is considered indistinguishable from non-conversion rates for most cell types which reflects the low non-CpG methylation observed in this study (Titcombe et al., 2022) (Figure 4.3E). Compared to the mncfDNA bin, both the BS-Seq and 5mCAdpBS-Seq protocol indicate that uscfDNA appears to have a lower CpG methylation%. In contrast, mncfDNA fragments which are derived from apoptosis (Heitzer et al., 2020) presented with ~70-80% CpG methylation matching the expected hypermethylated CpG profile in a typical genome (Strichman-Almashanu et al., 2002). In comparison, the lower CpG methylation % of uscfDNA illustrates that its origins are simply from mncfDNA undergoing further fragmentation.

There were both common and unique CpG sites between uscfDNA and mncfDNA, with less than half (41.4 ± 5%) of uscfDNA being unique. Since our definition for mncfDNA fragments encompassed reads from a larger bin (120-250bp vs. 40-70bp) and the average mncfDNA fragment is 3-times as long, this could explain why the total CpG sites in mncfDNA were more abundant compared to those of uscfDNA (Figure 4.6B).

The presence of methylated CpG residues within regulatory sequences can repress the expression of the corresponding gene (Dor & Cedar, 2018). Methyl-DNA binding proteins recruit factors that favor a compact chromatin conformation, thus reducing accessibility to transcription factors (Domcke et al., 2015). Between mncfDNA and uscfDNA, the CpG methylation traces had a similar pattern with a ~15% lower CpG methylation profile of uscfDNA fragments. From another perspective, the observation that the genomic element composition of CpG-containing fragments differs between uscfDNA and mncfDNA further indicates differences in their biogenesis. The uscfDNA bin had an enriched occupancy of fragments within simple repeat, promoters, exon, 5UTR, and CpG-Island elements regions, whereas the mncfDNA bin was increased in SINE and intergenic elements. The enrichment in promoters of uscfDNA was previously demonstrated in BRcfDNA-Seq and similar studies (J. Cheng et al., 2022; Hisano et al., 2021). There is a possibility, however, that bisulfite conversion may misalign repetitive sequences since only three distinct nucleotides are used during the alignment (Lerat et al., 2019).

Since uscfDNA was globally less CpG methylated compared to mncfDNA, we hypothesized that uscfDNA might be more associated with epigenetic marks(Zhang et al., 2015) related to active genes activity (Figure 4.7A), which could be more accessible due to the altered protein and nucleosome availability interactions (Domcke et al., 2015). Compared to the control shuffled regions, the uscfDNA fragments had the highest fold change in H3K4me3 and hypomethylated regions (Figure 4.7C). These genome regions may exhibit a hypomethylated-related organization that allows greater accessibility for nuclease activity to generate the appearance of hypomethylated uscfDNA in the circulation (Domcke et al., 2015; Teif et al., 2014). Another study has reported that the pattern of cfDNA fragmentation of H3K4me3 resembles the fragmentation pattern of regions of housekeeping genes in contrast to H3K9me3, which matches repressed genes (J. Guo et al., 2020). That report did not include uscfDNA analysis which might have demonstrated an even more distinct fragment pattern between active and non-active regions of the genome. It may be plausible that in circulation, a portion of uscfDNA is bound to H3K4me3 protein which has dissociated from the nucleosome, providing a protective effect. To confirm these findings, ChIP assays could be performed on plasma to determine if uscfDNA are bound to nucleosomal proteins or if they circulate freely.

For the genomic elements, we observed that methylation of CpG residues became more hypomethylated as they got closer to the center of CpG islands and promoter elements (Saxonov et al., 2006). This pattern was reflected in both uscfDNA and mncfDNA methylation profiles (Figure 4.8). Additionally, The methylation patterns upstream and downstream from the center of introns exons within cells have been reported in various studies (Gilsbach et al., 2014; H. Guo et al., 2014; Y. Li et al., 2010). The observation that the uscfDNA CpG methylation patterns still mirror those established patterns reported in the genome is suggestive that a subset of uscfDNA fragments could have originated from a genome before circulating in the blood as small fragments as opposed to being manufactured as uscfDNA fragments originally.

In contrast to the regulatory elements, simple repeats and LINE elements showed the greatest difference between cfDNA populations. Simple repeats, also known as short tandem repeats or microsatellites, are repeating units of 1-6 bases spread throughout the human genome and are prone to replication errors(Yu et al., 2021). For simple repeats, the uscfDNA demonstrated a dip in methylation as it approached the center of the element, while the mncfDNA only slightly dipped. The observed hypomethylation toward the start point of the simple repeat may be suggestive of differences in accessibility or regulatory events between uscfDNA and mncfDNA. In both simple repeats and LINE elements, the mncfDNA bin demonstrates an up-down-up methylation pattern towards and away from the element center point of these elements, which is not reflected by the uscfDNA bin. The similarities and differences between the CpG methylation patterns of different genomic elements for uscfDNA and mncfDNA likely reflect their unique regulatory mechanisms deserving further exploration.

Examining the common CpG regions for DMRs, there were regions where uscfDNA were more CpG methylated than mncfDNA. However, the majority of significantly different DMRs were from regions of decreased methylation in uscfDNA. This reflects the global observations that uscfDNA is hypomethylated compared to mncfDNA.

As a proof of concept, we attempted to uncover the tissue-of-origin of the plasma cell-free DNA using a pre-existing cell-free DNA deconvolution algorithm CelFie (Figure 4.9B). The deconvolution predicted that the mncfDNA derived from an assortment of blood cells that agreed with expected cell types in the blood(Razavi et al., 2019) and other prior cell-free DNA studies which used methylation DMRs to deconvolute literature(S. Guo et al., 2017; Moss et al., 2018). The uscfDNA also reported a profile with blood cells with an enrichment in eosinophils. Eosinophils have been reported to exhibit efficient DNA repair machinery for both double-strand and single-strand breaks(Salati et al., 2007). One possibility is that because uscfDNA is enriched in simple repeats, which are predisposed to double-strand break damage(Gadgil et al., 2020), the efficient repair process in blood cells (such as eosinophils) might lead to the generation of circulating uscfDNA by-products. Eosinophils are also reported to release DNA-based extracellular traps into circulation, which is another potential source of uscfDNA(Aoki et al., 2021; M. Mukherjee et al., 2018).

CpG-related cfDNA characteristics could potentially differentiate between non-cancer and NSCLC samples. When CpG methylation ratios for each size fragment were considered, the NSCLC samples appeared to be more hypermethylated in size bins <140bp (Figure 4.10). This observation contrasted with genome-wide hypomethylation normally observed in cancer cells compared to healthy cells(Jones & Baylin, 2002). However, the regions covered by cfDNA, and particularly uscfDNA, do not faithfully represent the genome in its entirety, as uscfDNA appears to be enriched in regulatory regions (Figure 4.6C) (J. Cheng et al., 2022; L. Y. Cheng et al., 2022; Hisano et al., 2021; Hudecova et al., 2021). Hypomethylation of transcription regions seems to occur less frequently in lung cancer(Hoffmann & Schulz, 2005; Pfeifer & Rauch, 2009; Rauch et al., 2008). Additionally, cfDNA is composed of DNA predominantly from blood cells more so than cancer tissue exclusively, which can explain the discrepancy.

In terms of genomic DNA elements for the uscfDNA bin, NSCLC samples presented significant changes in the percentage contribution of fragments of eight elements. For the mncfDNA bin, only the proportion of four elements differed in NSCLC samples. The change in the profile of elements is suggestive that cfDNA fragments entering the circulating environment may undergo alterations during a NSCLC situation. The increase in SINE, LINE, introns, and intergenic regions could be related to higher turnover from increased accessibility of these regions to nucleases.

In cancer, it has been reported that certain promoters and CpG Islands may become hypermethylated(Harden et al., 2003). In our sample set, both NSCLC uscfDNA and mncfDNA demonstrated substantial hypermethylation in the promoter, 5'UTR, CpG Islands, and exon elements compared to non-cancer subjects (Figure 4.11C and D). This hypermethylation was more evident in the mncfDNA population. Compared to the non-cancer samples, both the uscfDNA and mncfDNA NSCLC methylation profiles presented with increased signal variability that did not cluster together as tightly as the non-caner samples. The simple repeats for both uscfDNA and mncfDNA also presented a different methylation trace. The NSCLC simple repeat profile appeared "flattened" and absent of the steep decrease in hypomethylation to the center point of the simple repeat.

In contrast to the other elements, which were either hypermethylated or variable, we observed that the LINE elements of NSCLC subjects trended toward a hypomethylated state. In the genome, LINE elements have been described to undergo hypomethylation in cancer (Rauch et al., 2008). These high variability traces may be indicative of micro instability in the epigenetic regulation of these elements. The greater separation in mncfDNA may be due to the greater contribution of tumor-derived fragments, which have been shown to be enriched at 90-150bp (Mouliere et al., 2018). It is unclear if the changes in methylation patterns originate from an increasing load of tumor-cfDNA or from adjustments in activity from the immune system.

The limited number of DMRs for uscfDNA was the result of the overlap between the two uscfDNA fractions. Potentially deeper coverage could increase the number of candidates. Regardless, we were able to show that both uscfDNA and mncfDNA bins could be useful sources of DMR candidates between the two clinical cohorts (Figure 4.11E). The Increased expression of PK3P is associated with various types of cancer, including colon, lung, and bladder cancer (Furukawa et al., 2005; Ruan et al., 2021). CPLX1 is one of several factors that has been shown to be able to influence the activity of cyclin B1 (CCNB1), which is highly expressed in lung adenocarcinoma and associated with poor prognosis(Y. Li et al., 2022). CPLX1 has been documented to promote malignancy in gastric cancer (H. Tanaka et al., 2022). The expression of COL26A1 has been observed to be downregulated in subjects who respond well to PD-L1 inhibitors in transformed small-cell lung carcinoma. For mncfDNA candidates, mutations in ZNF595 have been indicated as a potential germline mutation in familial lung cancer(Kanwal et al., 2018) and region for prevalent somatic mutations in gastric cancer(Cui et al., 2015). The non-pseudo gene version of MLLT10 has been documented to be a promoter of tumor cell proliferation, migration, and invasion in NSCLC cell lines (Tian et al., 2020) and MLLT10P1 is commonly mutated in breast cancer subjects (Pongor et al., 2015). NERUDO2 has been shown to be hypermethylated in adenocarcinoma in situ tissues contrasting the hypomethylation we saw in our study (Selamat et al., 2011). Despite the potential biological rationale discussed, these DMRs are not currently validated. However, this approach shows the merit of DMR discovery, which could give rise to useful targets for future cancer detection.

The deconvolution prediction could also be a potential biomarker strategy for NSCLC detection (Figure 4.11H and I). Surprisingly for both uscfDNA and mncfDNA, we did not observe the signal from lung cell tissues despite the samples coming from NSCLC. Despite the cases being late-stage, the majority of cfDNA is still from blood cell origin (Razavi et al., 2019). For uscfDNA, the starkest change was a decrease in eosinophils % and a trend in increased neutrophils. Increased eosinophils have been associated with improved prognosis in lung cancer (Costello et al., 2005; Davis & Rothenberg, 2014). In the mncfDNA, the megakaryocytes were increased, which has also been described to be associated with cancer (Dejima et al., 2018; Huang et al., 2015; Soares, 1992) in the literature. Additionally, this analysis is only a proof-of-principle and should be taken with caution since the inferences are not validated.

Using whole-genome sequencing, other investigators have reported that uscfDNA predicted to contain G-Quad secondary structures are decreased in cancer subjects (Hudecova et al., 2021). In our study, this pattern was also observed in NSCLC samples that have undergone bisulfite conversion (Figure 5E). Interestingly, in NSCLC subjects, fragments that contained potential G-Quad structures showed increased CpG methylation levels compared to non-cancer subjects (Figure 4.12A and B). Within the genome, G-Quad has been described to regulate methylation behavior at CpG Islands (Mao et al., 2018; A. K. Mukherjee et al., 2019). It is possible that although there is a decrease in G-Quad structures present in the plasma reflects changes in altered CpG methylation and subsequent changes in transcription factors or chromosomal inaccessibility.

Epigenetic marks and their associative enzymes influence the activity of gene expression by affecting chromatin compaction, nucleosome dynamics, and transcription (Zhao & Shilatifard, 2019). There are many types of epigenetic marks, including chromatin conformation, histone modifications (e.g., acetylation, phosphorylation, and methylation), and DNA methylation (Zukowski et al., 2020). Mutations in chromatin-bound proteins frequently occur in cancer (H. Shen & Laird, 2013). We observed that %intersection with epigenetic marks was also altered in NSCLC subjects with the greatest decreases in %intersection of H3K27ac and H3K4me3 for both uscfDNA and mncfDNA (Figure 4.12C). As these two marks are associated with genes with high expression, their decrease in the NSCLC samples seen in our study may be suggestive of dysregulation in cancer and a potential viable global indicator.

In conclusion, the 5mCAdpBS-Seq single-stranded DNA library preparation is advantageous for uscfDNA methylation profile investigation due to the preservation of the native fragment length and methylation level in each size bin. Using this protocol, the methylation characteristics of uscfDNA appear distinctly different than mncfDNA, further illustrating that it should be considered a separate cfDNA molecule. As a methylated-based cancer biomarker, potentially useful features of uscfDNA are global CpG% methylation changes, genome element profiles, CpG-methylation traces for specific elements, DMRs, tissue-of-origin deconvolution, G-Quad signature changes, and epigenetic mark association. Although we have focused on cfDNA from plasma, the 5mCAdpBS-Seq protocol is useful for any contexts where very short DNA templates are present. This can include analysis of other biofluids with fragmented DNA (saliva and urine (Brooks et al., 2023; Chen et al., 2022)), cell-culture conditioned media environments (Bronkhorst et al., 2019), or theoretically in any in-vitro intracellular work where the accurate methylation analysis of short single-stranded DNA is required. Therefore, if investigators are interested in examining the methylation profile of a DNA sample with heterogeneous sizes, the 5mCAdpBS-Seq protocol should be considered.

4.5 References

Allis, C. D., & Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nature Reviews Genetics, 17(8), Article 8. https://doi.org/10.1038/nrg.2016.59

Aoki, A., Hirahara, K., Kiuchi, M., & Nakayama, T. (2021). Eosinophils: Cells known for over 140 years with broad and new functions. Allergology International, 70(1), 3–8. https://doi.org/10.1016/j.alit.2020.09.002

Babenko, V. N., Chadaeva, I. V., & Orlov, Y. L. (2017). Genomic landscape of CpG rich elements in human. BMC Evolutionary Biology, 17(Suppl 1), 19. https://doi.org/10.1186/s12862-016-0864-0

Balgkouranidou, I., Chimonidou, M., Milaki, G., Tsaroucha, E., Kakolyris, S., Georgoulias, V., & Lianidou, E. (2016). SOX17 promoter methylation in plasma circulating tumor DNA of patients with non-small cell lung cancer. Clinical Chemistry and Laboratory Medicine, 54(8), 1385–1393. https://doi.org/10.1515/cclm-2015-0776

Balgkouranidou, I., Chimonidou, M., Milaki, G., Tsarouxa, E. G., Kakolyris, S., Welch, D. R., Georgoulias, V., & Lianidou, E. S. (2014). Breast cancer metastasis suppressor-1 promoter methylation in cell-free DNA provides prognostic information in non-small cell lung cancer. British Journal of Cancer, 110(8), 2054–2062. https://doi.org/10.1038/bjc.2014.104

Baylin, S. B., Esteller, M., Rountree, M. R., Bachman, K. E., Schuebel, K., & Herman, J. G. (2001). Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. Human Molecular Genetics, 10(7), 687–692. https://doi.org/10.1093/hmg/10.7.687

Brabender, J., Usadel, H., Metzger, R., Schneider, P. M., Park, J., Salonga, D., Tsao-Wei, D. D., Groshen, S., Lord, R. V., Takebe, N., Schneider, S., Hölscher, A. H., Danenberg, K. D., & Danenberg, P. V. (2003). Quantitative O(6)-methylguanine DNA methyltransferase methylation analysis in curatively resected non-small cell lung cancer: Associations with clinical outcome. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 9(1), 223–227.

Bronkhorst, A. J., Ungerer, V., & Holdenrieder, S. (2019). Comparison of methods for the quantification of cell-free DNA isolated from cell culture supernatant. Tumor Biology, 41(8), 1010428319866369. https://doi.org/10.1177/1010428319866369

Brooks, P. J., Malkin, E. Z., Michino, S. D., & Bratman, S. V. (2023). Isolation of salivary cell-free DNA for cancer detection. PLOS ONE, 18(5), e0285214. https://doi.org/10.1371/journal.pone.0285214

Caggiano, C., Celona, B., Garton, F., Mefford, J., Black, B. L., Henderson, R., Lomen-Hoerth, C., Dahl, A., & Zaitlen, N. (2021). Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. Nature Communications, 12(1), Article 1. https://doi.org/10.1038/s41467-021-22901-x

Chan, K. C. A., Jiang, P., Chan, C. W. M., Sun, K., Wong, J., Hui, E. P., Chan, S. L., Chan, W. C., Hui, D. S. C., Ng, S. S. M., Chan, H. L. Y., Wong, C. S. C., Ma, B. B. Y., Chan, A. T. C., Lai, P. B. S., Sun, H., Chiu, R. W. K., & Lo, Y. M. D. (2013). Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proceedings of the National Academy of Sciences of the United States of America, 110(47), 18761–18768. https://doi.org/10.1073/pnas.1313995110

Chen, M., Chan, R. W. Y., Cheung, P. P. H., Ni, M., Wong, D. K. L., Zhou, Z., Ma, M.-J. L., Huang, L., Xu, X., Lee, W.-S., Wang, G., Lui, K. O., Lam, W. K. J., Teoh, J. Y. C., Ng, C.-F., Jiang, P., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2022). Fragmentomics of urinary cell-free DNA in nuclease knockout mouse models. PLoS Genetics, 18(7), e1010262. https://doi.org/10.1371/journal.pgen.1010262

Cheng, J., Morselli, M., Huang, W.-L., Heo, Y. J., Pinheiro-Ferreira, T., Li, F., Wei, F., Chia, D., Kim, Y., He, H.-J., Cole, K. D., Su, W.-C., Pellegrini, M., & Wong, D. T. W. (2022). Plasma contains ultrashort single-stranded DNA in addition to nucleosomal cell-free DNA. IScience, 25(7), 104554. https://doi.org/10.1016/j.isci.2022.104554

Cheng, L. Y., Dai, P., Wu, L. R., Patel, A. A., & Zhang, D. Y. (2022). Direct capture and sequencing reveal ultra-short single-stranded DNA in biofluids. IScience, 25(10), 105046. https://doi.org/10.1016/j.isci.2022.105046

Cosmic Database. (2018). COSMIC - Catalogue of Somatic Mutations in Cancer. https://cancer.sanger.ac.uk/cosmic

Costello, R., O'Callaghan, T., & Sébahoun, G. (2005). [Eosinophils and antitumour response]. La Revue De Medecine Interne, 26(6), 479–484. https://doi.org/10.1016/j.revmed.2005.02.013

Cui, J., Yin, Y., Ma, Q., Wang, G., Olman, V., Zhang, Y., Chou, W.-C., Hong, C. S., Zhang, C., Cao, S., Mao, X., Li, Y., Qin, S., Zhao, S., Jiang, J., Hastings, P., Li, F., & Xu, Y. (2015). Comprehensive characterization of the genomic alterations in human gastric cancer. International Journal of Cancer, 137(1), 86–95. https://doi.org/10.1002/ijc.29352

Da, K., Gl, S., & Jl, R. (2010). DNA methylation and epigenetic control of cellular differentiation. Cell Cycle (Georgetown, Tex.), 9(19). https://doi.org/10.4161/cc.9.19.13385

Daskalos, A., Nikolaidis, G., Xinarianos, G., Savvari, P., Cassidy, A., Zakopoulou, R., Kotsinas, A., Gorgoulis, V., Field, J. K., & Liloglou, T. (2009). Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. International Journal of Cancer, 124(1), 81–87. https://doi.org/10.1002/ijc.23849

Davis, B. P., & Rothenberg, M. E. (2014). Eosinophils and Cancer. Cancer Immunology Research, 2(1), 1–8. https://doi.org/10.1158/2326-6066.CIR-13-0196

de Goede, O. M., Razzaghian, H. R., Price, E. M., Jones, M. J., Kobor, M. S., Robinson, W. P., & Lavoie, P. M. (2015). Nucleated red blood cells impact DNA methylation and expression analyses of cord blood hematopoietic cells. Clinical Epigenetics, 7(1), 95. https://doi.org/10.1186/s13148-015-0129-6

Dejima, H., Nakanishi, H., Kuroda, H., Yoshimura, M., Sakakura, N., Ueda, N., Ohta, Y., Tanaka, R., Mori, S., Yoshida, T., Hida, T., Sawabata, N., Yatabe, Y., & Sakao, Y. (2018). Detection of abundant megakaryocytes in pulmonary artery blood in lung cancer patients using a microfluidic platform. Lung Cancer (Amsterdam, Netherlands), 125, 128–135. https://doi.org/10.1016/j.lungcan.2018.09.011

DeNizio, J. E., Liu, M. Y., Leddin, E. M., Cisneros, G. A., & Kohli, R. M. (2019). Selectivity and Promiscuity in TET-Mediated Oxidation of 5-Methylcytosine in DNA and RNA. Biochemistry, 58(5), 411–421. https://doi.org/10.1021/acs.biochem.8b00912

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L., & Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature, 528(7583), 575–579. https://doi.org/10.1038/nature16462

Dor, Y., & Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. Lancet (London, England), 392(10149), 777–786. https://doi.org/10.1016/S0140-6736(18)31268-6

Espada, J., Ballestar, E., Santoro, R., Fraga, M. F., Villar-Garea, A., Németh, A., Lopez-Serra, L., Ropero, S., Aranda, A., Orozco, H., Moreno, V., Juarranz, A., Stockert, J. C., Längst, G., Grummt, I., Bickmore, W., & Esteller, M. (2007). Epigenetic disruption of ribosomal RNA genes and nucleolar architecture in DNA methyltransferase 1 (Dnmt1) deficient cells. Nucleic Acids Research, 35(7), 2191–2198. https://doi.org/10.1093/nar/gkm118

Esteller, M. (2008). Epigenetics in cancer. The New England Journal of Medicine, 358(11), 1148–1159. https://doi.org/10.1056/NEJMra072067

Filipczak, P. T., Leng, S., Tellez, C. S., Do, K. C., Grimes, M. J., Thomas, C. L., Walton-Filipczak, S. R., Picchi, M. A., & Belinsky, S. A. (2019). P53-Suppressed Oncogene TET1 Prevents Cellular Aging in Lung Cancer. Cancer Research, 79(8), 1758–1768. https://doi.org/10.1158/0008-5472.CAN-18-1234

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., & Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-

methylcytosine residues in individual DNA strands. Proceedings of the National Academy of Sciences of the United States of America, 89(5), 1827–1831.

Furukawa, C., Daigo, Y., Ishikawa, N., Kato, T., Ito, T., Tsuchiya, E., Sone, S., & Nakamura, Y. (2005). Plakophilin 3 oncogene as prognostic marker and therapeutic target for lung cancer. Cancer Research, 65(16), 7102–7110. https://doi.org/10.1158/0008-5472.CAN-04-1877

Gadgil, R. Y., Romer, E. J., Goodman, C. C., Rider, S. D., Damewood, F. J., Barthelemy, J. R., Shin-Ya, K., Hanenberg, H., & Leffak, M. (2020). Replication stress at microsatellites causes DNA double-strand breaks and break-induced replication. The Journal of Biological Chemistry, 295(45), 15378–15397. https://doi.org/10.1074/jbc.RA120.013495

Gai, W., Ji, L., Lam, W. K. J., Sun, K., Jiang, P., Chan, A. W. H., Wong, J., Lai, P. B. S., Ng, S. S. M., Ma, B. B. Y., Wong, G. L. H., Wong, V. W. S., Chan, H. L. Y., Chiu, R. W. K., Lo, Y. M. D., & Chan, K. C. A. (2018). Liver- and Colon-Specific DNA Methylation Markers in Plasma for Investigation of Colorectal Cancers with or without Liver Metastases. Clinical Chemistry, 64(8), 1239–1249. https://doi.org/10.1373/clinchem.2018.290304

Gala-Lopez, B. L., Neiman, D., Kin, T., O'Gorman, D., Pepper, A. R., Malcolm, A. J., Pianzin, S., Senior, P. A., Campbell, P., Glaser, B., Dor, Y., Shemer, R., & Shapiro, A. M. J. (2018). Beta Cell Death by Cell-free DNA and Outcome After Clinical Islet Transplantation. Transplantation, 102(6), 978–985. https://doi.org/10.1097/TP.0000000000002083

Gilsbach, R., Preissl, S., Grüning, B. A., Schnick, T., Burger, L., Benes, V., Würch, A., Bönisch, U., Günther, S., Backofen, R., Fleischmann, B. K., Schübeler, D., & Hein, L. (2014). Dynamic DNA methylation orchestrates cardiomyocyte development, maturation and disease. Nature Communications, 5, 5288. https://doi.org/10.1038/ncomms6288

Gomes, A., Reis-Silva, M., Alarcão, A., Couceiro, P., Sousa, V., & Carvalho, L. (2014). Promoter hypermethylation of DNA repair genes MLH1 and MSH2 in adenocarcinomas and squamous cell carcinomas of the lung. Revista Portuguesa De Pneumologia, 20(1), 20–30. https://doi.org/10.1016/j.rppneu.2013.07.003

Grote, H. J., Schmiemann, V., Kiel, S., Böcking, A., Kappes, R., Gabbert, H. E., & Sarbia, M. (2004). Aberrant methylation of the adenomatous polyposis coli promoter 1A in bronchial aspirates from patients with suspected lung cancer. International Journal of Cancer, 110(5), 751–755. https://doi.org/10.1002/ijc.20196

Grunau, C., Clark, S. J., & Rosenthal, A. (2001). Bisulfite genomic sequencing: Systematic investigation of critical experimental parameters. Nucleic Acids Research, 29(13), E65-65. https://doi.org/10.1093/nar/29.13.e65

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., Jin, X., Shi, X., Liu, P., Wang, X., Wang, W., Wei, Y., Li, X., Guo, F., Wu, X., … Qiao, J. (2014). The DNA methylation landscape of human early embryos. Nature, 511(7511), 606–610. https://doi.org/10.1038/nature13544

Guo, J., Ma, K., Bao, H., Ma, X., Xu, Y., Wu, X., Shao, Y. W., Jiang, M., & Huang, J. (2020). Quantitative characterization of tumor cell-free DNA shortening. BMC Genomics, 21(1), 473. https://doi.org/10.1186/s12864-020-06848-9

Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K., & Zhang, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nature Genetics, 49(4), 635–642. https://doi.org/10.1038/ng.3805

Harden, S. V., Tokumaru, Y., Westra, W. H., Goodman, S., Ahrendt, S. A., Yang, S. C., & Sidransky, D. (2003). Gene promoter hypermethylation in tumors and lymph nodes of stage I lung cancer patients. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 9(4), 1370–1375.

Heitzer, E., Auinger, L., & Speicher, M. R. (2020). Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. Trends in Molecular Medicine, 26(5), 519–528. https://doi.org/10.1016/j.molmed.2020.01.012

Heller, G., Fong, K. M., Girard, L., Seidl, S., End-Pfützenreuter, A., Lang, G., Gazdar, A. F., Minna, J. D., Zielinski, C. C., & Zöchbauer-Müller, S. (2006). Expression and methylation pattern of TSLC1 cascade genes in lung carcinomas. Oncogene, 25(6), 959–968. https://doi.org/10.1038/sj.onc.1209115

Hisano, O., Ito, T., & Miura, F. (2021). Short single-stranded DNAs with putative non-canonical structures comprise a new class of plasma cell-free DNA. BMC Biology, 19(1), 225. https://doi.org/10.1186/s12915-021-01160-8

Hoffmann, M. J., & Schulz, W. A. (2005). Causes and consequences of DNA hypomethylation in human cancer. Biochemistry and Cell Biology = Biochimie Et Biologie Cellulaire, 83(3), 296–321. https://doi.org/10.1139/o05-036

Hopkins-Donaldson, S., Ziegler, A., Kurtz, S., Bigosch, C., Kandioler, D., Ludwig, C., Zangemeister-Wittke, U., & Stahel, R. (2003). Silencing of death receptor and caspase-8 expression in small cell lung carcinoma cell lines and tumors by DNA methylation. Cell Death and Differentiation, 10(3), 356–364. https://doi.org/10.1038/sj.cdd.4401157

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., & Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics, 13(1), 86. https://doi.org/10.1186/1471-2105-13-86

Huang, S. H., Xu, W., Waldron, J., Siu, L., Shen, X., Tong, L., Ringash, J., Bayley, A., Kim, J., Hope, A., Cho, J., Giuliani, M., Hansen, A., Irish, J., Gilbert, R., Gullane, P., Perez-Ordonez, B., Weinreb, I., Liu, F.-F., & O'Sullivan, B. (2015). Refining American Joint Committee on Cancer/Union for International Cancer Control TNM stage and prognostic groups for human papillomavirus-related oropharyngeal carcinomas. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 33(8), 836–845. https://doi.org/10.1200/JCO.2014.58.6412

Hudecova, I., Smith, C. G., Hänsel-Hertsch, R., Chilamakuri, C. S., Morris, J. A., Vijayaraghavan, A., Heider, K., Chandrananda, D., Cooper, W. N., Gale, D., Garcia-Corbacho, J., Pacey, S., Baird, R. D., Rosenfeld, N., & Mouliere, F. (2021). Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. Genome Research. https://doi.org/10.1101/gr.275691.121

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J., Sabunciyan, S., & Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature Genetics, 41(2), 178–186. https://doi.org/10.1038/ng.298

Jang, S. J., Soria, J. C., Wang, L., Hassan, K. A., Morice, R. C., Walsh, G. L., Hong, W. K., & Mao, L. (2001). Activation of melanoma antigen tumor antigens occurs early in lung carcinogenesis. Cancer Research, 61(21), 7959–7963.

Johnstone, S. E., Reyes, A., Qi, Y., Adriaens, C., Hegazi, E., Pelka, K., Chen, J. H., Zou, L. S., Drier, Y., Hecht, V., Shoresh, N., Selig, M. K., Lareau, C. A., Iyer, S., Nguyen, S. C., Joyce, E. F., Hacohen, N., Irizarry, R. A., Zhang, B., … Bernstein, B. E. (2020). Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer. Cell, 182(6), 1474-1489.e23. https://doi.org/10.1016/j.cell.2020.07.030

Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. Nature Reviews. Genetics, 13(7), 484–492. https://doi.org/10.1038/nrg3230

Jones, P. A., & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. Nature Reviews. Genetics, 3(6), 415–428. https://doi.org/10.1038/nrg816

Jurkowska, R. Z., Anspach, N., Urbanke, C., Jia, D., Reinhardt, R., Nellen, W., Cheng, X., & Jeltsch, A. (2008). Formation of nucleoprotein filaments by mammalian DNA methyltransferase Dnmt3a in complex with regulator Dnmt3L. Nucleic Acids Research, 36(21), 6656–6663. https://doi.org/10.1093/nar/gkn747

Kang, S., Li, Q., Chen, Q., Zhou, Y., Park, S., Lee, G., Grimes, B., Krysan, K., Yu, M., Wang, W., Alber, F., Sun, F., Dubinett, S. M., Li, W., & Zhou, X. J. (2017). CancerLocator: Non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biology, 18(1), 53. https://doi.org/10.1186/s13059-017-1191-5

Kanwal, M., Ding, X.-J., Ma, Z.-H., Li, L.-W., Wang, P., Chen, Y., Huang, Y.-C., & Cao, Y. (2018). Characterization of germline mutations in familial lung cancer from the Chinese population. Gene, 641, 94–104. https://doi.org/10.1016/j.gene.2017.10.020

Kim, D. H., Nelson, H. H., Wiencke, J. K., Christiani, D. C., Wain, J. C., Mark, E. J., & Kelsey, K. T. (2001). Promoter methylation of DAP-kinase: Association with advanced stage in non-small cell lung cancer. Oncogene, 20(14), 1765–1770. https://doi.org/10.1038/sj.onc.1204302

Kim, D. H., Nelson, H. H., Wiencke, J. K., Zheng, S., Christiani, D. C., Wain, J. C., Mark, E. J., & Kelsey, K. T. (2001). P16(INK4a) and histology-specific methylation of CpG islands by exposure to tobacco smoke in non-small cell lung cancer. Cancer Research, 61(8), 3419–3424.

Kim, D. S., Kim, M. J., Lee, J. Y., Kim, Y. Z., Kim, E. J., & Park, J. Y. (2007). Aberrant methylation of E-cadherin and H-cadherin genes in nonsmall cell lung cancer and its relation to clinicopathologic features. Cancer, 110(12), 2785–2792. https://doi.org/10.1002/cncr.23113

Kim, H., Kwon, Y. M., Kim, J. S., Han, J., Shim, Y. M., Park, J., & Kim, D.-H. (2006). Elevated mRNA levels of DNA methyltransferase-1 as an independent prognostic factor in primary nonsmall cell lung cancer. Cancer, 107(5), 1042–1049. https://doi.org/10.1002/cncr.22087

Kneip, C., Schmidt, B., Seegebarth, A., Weickmann, S., Fleischhacker, M., Liebenberg, V., Field, J. K., & Dietrich, D. (2011). SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer in plasma. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer, 6(10), 1632–1638. https://doi.org/10.1097/JTO.0b013e318220ef9a

Kohli, R. M., & Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. Nature, 502(7472), 472–479. https://doi.org/10.1038/nature12750

Lam, W. K. J., Gai, W., Sun, K., Wong, R. S. M., Chan, R. W. Y., Jiang, P., Chan, N. P. H., Hui, W. W. I., Chan, A. W. H., Szeto, C.-C., Ng, S. C., Law, M.-F., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2017). DNA of Erythroid Origin Is Present in Human Plasma and Informs the Types of Anemia. Clinical Chemistry, 63(10), 1614–1623. https://doi.org/10.1373/clinchem.2017.272401

Leddin, E. M., & Cisneros, G. A. (2019). Comparison of DNA and RNA substrate effects on TET2 structure. Advances in Protein Chemistry and Structural Biology, 117, 91–112. https://doi.org/10.1016/bs.apcsb.2019.05.002

Lehmann-Werman, R., Magenheim, J., Moss, J., Neiman, D., Abraham, O., Piyanzin, S., Zemmour, H., Fox, I., Dor, T., Grompe, M., Landesberg, G., Loza, B.-L., Shaked, A., Olthoff, K., Glaser, B., Shemer, R., & Dor, Y. (2018). Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. JCI Insight, 3(12), e120687, 120687. https://doi.org/10.1172/jci.insight.120687

Lehmann-Werman, R., Neiman, D., Zemmour, H., Moss, J., Magenheim, J., Vaknin-Dembinsky, A., Rubertsson, S., Nellgård, B., Blennow, K., Zetterberg, H., Spalding, K., Haller, M. J., Wasserfall, C. H., Schatz, D. A., Greenbaum, C. J., Dorrell, C., Grompe, M., Zick, A., Hubert, A., … Dor, Y. (2016). Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proceedings of the National Academy of Sciences of the United States of America, 113(13), E1826-1834. https://doi.org/10.1073/pnas.1519286113

Lerat, E., Casacuberta, J., Chaparro, C., & Vieira, C. (2019). On the Importance to Acknowledge Transposable Elements in Epigenomic Analyses. Genes, 10(4), Article 4. https://doi.org/10.3390/genes10040258

Li, W., Li, Q., Kang, S., Same, M., Zhou, Y., Sun, C., Liu, C.-C., Matsuoka, L., Sher, L., Wong, W. H., Alber, F., & Zhou, X. J. (2018). CancerDetector: Ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Research, 46(15), e89. https://doi.org/10.1093/nar/gky423

Li, Y., Leng, Y., Dong, Y., Song, Y., Wu, Q., Jiang, N., Dong, H., Chen, F., Luo, Q., & Cheng, C. (2022). Cyclin B1 expression as an independent prognostic factor for lung adenocarcinoma and its potential pathways. Oncology Letters, 24(6), 441. https://doi.org/10.3892/ol.2022.13561

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., Zhang, H., Chen, Q., Luo, R., Chen, M., He, Y., Jin, X., Zhang, Q., Yu, C., Zhou, G., … Zhang, X. (2010). The DNA methylome of human peripheral blood mononuclear cells. PLoS Biology, 8(11), e1000533. https://doi.org/10.1371/journal.pbio.1000533

Lin, R.-K., Hsu, H.-S., Chang, J.-W., Chen, C.-Y., Chen, J.-T., & Wang, Y.-C. (2007). Alteration of DNA methyltransferases contributes to 5′CpG methylation and poor prognosis in lung cancer. Lung Cancer (Amsterdam, Netherlands), 55(2), 205–213. https://doi.org/10.1016/j.lungcan.2006.10.022

Lin, R.-K., Wu, C.-Y., Chang, J.-W., Juan, L.-J., Hsu, H.-S., Chen, C.-Y., Lu, Y.-Y., Tang, Y.-A., Yang, Y.-C., Yang, P.-C., & Wang, Y.-C. (2010). Dysregulation of p53/Sp1 control leads to DNA methyltransferase-1 overexpression in lung cancer. Cancer Research, 70(14), 5807–5817. https://doi.org/10.1158/0008-5472.CAN-09-4161

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., & Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462(7271), 315–322. https://doi.org/10.1038/nature08514

Liu, B., Du, Q., Chen, L., Fu, G., Li, S., Fu, L., Zhang, X., Ma, C., & Bin, C. (2016). CpG methylation patterns of human mitochondrial DNA. Scientific Reports, 6(1), Article 1. https://doi.org/10.1038/srep23421

Liu, H., Liu, W., Wu, Y., Zhou, Y., Xue, R., Luo, C., Wang, L., Zhao, W., Jiang, J.-D., & Liu, J. (2005). Loss of epigenetic control of synuclein-gamma gene as a molecular indicator of metastasis in a wide range of human cancers. Cancer Research, 65(17), 7635–7643. https://doi.org/10.1158/0008-5472.CAN-05-1089

Lui, Y. Y. N., Chik, K.-W., Chiu, R. W. K., Ho, C.-Y., Lam, C. W. K., & Lo, Y. M. D. (2002). Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. Clinical Chemistry, 48(3), 421–427.

Mao, S.-Q., Ghanbarian, A. T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hänsel-Hertsch, R., Tannahill, D., & Balasubramanian, S. (2018). DNA G-Quadruplex structures mold the DNA methylome. Nature Structural & Molecular Biology, 25(10), 951–957. https://doi.org/10.1038/s41594-018-0131-8

Mastoraki, S., Balgkouranidou, I., Tsaroucha, E., Klinakis, A., Georgoulias, V., & Lianidou, E. (2021). KMT2C promoter methylation in plasma-circulating tumor DNA is a prognostic biomarker in non-small cell lung cancer. Molecular Oncology, 15(9), 2412–2422. https://doi.org/10.1002/1878-0261.12848

Mechta, M., Ingerslev, L. R., Fabre, O., Picard, M., & Barrès, R. (2017). Evidence Suggesting Absence of Mitochondrial DNA Methylation. Frontiers in Genetics, 8, 166. https://doi.org/10.3389/fgene.2017.00166

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Research, 33(18), 5868–5877. https://doi.org/10.1093/nar/gki901

Merlo, A., Herman, J. G., Mao, L., Lee, D. J., Gabrielson, E., Burger, P. C., Baylin, S. B., & Sidransky, D. (1995). 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nature Medicine, 1(7), 686–692. https://doi.org/10.1038/nm0795-686

Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., Fu, K.-Y., Kiss, E., Spalding, K. L., Landesberg, G., Zick, A., Grinshpun, A., Shapiro, A. M. J., Grompe, M., Wittenberg, A. D., … Dor, Y. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nature Communications, 9(1), Article 1. https://doi.org/10.1038/s41467-018-07466-6

Mouliere, F., Chandrananda, D., Piskorz, A. M., Moore, E. K., Morris, J., Ahlborn, L. B., Mair, R., Goranova, T., Marass, F., Heider, K., Wan, J. C. M., Supernat, A., Hudecova, I., Gounaris, I., Ros, S., Jimenez-Linan, M., Garcia-Corbacho, J., Patel, K., Østrup, O., … Rosenfeld, N. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. Science Translational Medicine, 10(466). https://doi.org/10.1126/scitranslmed.aat4921

Mukherjee, A. K., Sharma, S., & Chowdhury, S. (2019). Non-duplex G-Quadruplex Structures Emerge as Mediators of Epigenetic Modifications. Trends in Genetics : TIG, 35(2), 129–144. https://doi.org/10.1016/j.tig.2018.11.001

Mukherjee, M., Lacy, P., & Ueki, S. (2018). Eosinophil Extracellular Traps and Inflammatory Pathologies—Untangling the Web! Frontiers in Immunology, 9, 2763. https://doi.org/10.3389/fimmu.2018.02763

Munson, K., Clark, J., Lamparska-Kupsik, K., & Smith, S. S. (2007). Recovery of bisulfite-converted genomic sequences in the methylation-sensitive QPCR. Nucleic Acids Research, 35(9), 2893–2903. https://doi.org/10.1093/nar/gkm055

Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell, 99(3), 247–257. https://doi.org/10.1016/S0092-8674(00)81656-6

Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., Hölttä, M., Rosén, C., Olsson, C., Strobel, G., Wu, E., Dakin, K., Petzold, M., Blennow, K., & Zetterberg, H. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer's disease: A systematic review and meta-analysis. The Lancet. Neurology, 15(7), 673–684. https://doi.org/10.1016/S1474-4422(16)00070-3

Ooki, A., Maleki, Z., Tsay, J.-C. J., Goparaju, C., Brait, M., Turaga, N., Nam, H.-S., Rom, W. N., Pass, H. I., Sidransky, D., Guerrero-Preston, R., & Hoque, M. O. (2017). A Panel of Novel Detection and Prognostic Methylated DNA Markers in Primary Non-Small Cell Lung Cancer and Serum DNA. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 23(22), 7141–7152. https://doi.org/10.1158/1078-0432.CCR-17-1222

Pfeifer, G. P., & Rauch, T. A. (2009). DNA methylation patterns in lung carcinomas. Seminars in Cancer Biology, 19(3), 181–187. https://doi.org/10.1016/j.semcancer.2009.02.008

Pongor, L., Kormos, M., Hatzis, C., Pusztai, L., Szabó, A., & Győrffy, B. (2015). A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6,697 breast cancer patients. Genome Medicine, 7, 104. https://doi.org/10.1186/s13073-015-0228-1

Ponomaryova, A. A., Rykova, E. Y., Cherdyntseva, N. V., Skvortsova, T. E., Dobrodeev, A. Y., Zav'yalov, A. A., Bryzgalov, L. O., Tuzikov, S. A., Vlassov, V. V., & Laktionov, P. P. (2013). Potentialities of aberrantly methylated circulating DNA for diagnostics and post-treatment follow-up of lung cancer patients. Lung Cancer (Amsterdam, Netherlands), 81(3), 397–403. https://doi.org/10.1016/j.lungcan.2013.05.016

Powrózek, T., Krawczyk, P., Nicoś, M., Kuźnar-Kamińska, B., Batura-Gabryel, H., & Milanowski, J. (2016). Methylation of the DCLK1 promoter region in circulating free DNA and its prognostic value in lung cancer patients. Clinical & Translational Oncology: Official Publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico, 18(4), 398–404. https://doi.org/10.1007/s12094-015-1382-z

Raizis, A. M., Schmitt, F., & Jost, J. P. (1995). A bisulfite method of 5-methylcytosine mapping that minimizes template degradation. Analytical Biochemistry, 226(1), 161–166. https://doi.org/10.1006/abio.1995.1204

Ramasamy, D., Deva Magendhra Rao, A. K., Rajkumar, T., & Mani, S. (2021). Non-CpG methylation-a key epigenetic modification in cancer. Briefings in Functional Genomics, 20(5), 304–311. https://doi.org/10.1093/bfgp/elab035

Rasmussen, K. D., & Helin, K. (2016). Role of TET enzymes in DNA methylation, development, and cancer. Genes & Development, 30(7), 733–750. https://doi.org/10.1101/gad.276568.115

Rauch, T. A., Zhong, X., Wu, X., Wang, M., Kernstine, K. H., Wang, Z., Riggs, A. D., & Pfeifer, G. P. (2008). High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. Proceedings of the National Academy of Sciences of the United States of America, 105(1), 252–257. https://doi.org/10.1073/pnas.0710735105

Razavi, P., Li, B. T., Brown, D. N., Jung, B., Hubbell, E., Shen, R., Abida, W., Juluru, K., De Bruijn, I., Hou, C., Venn, O., Lim, R., Anand, A., Maddala, T., Gnerre, S., Vijaya Satya, R., Liu, Q., Shen, L., Eattock, N., … Reis-Filho, J. S. (2019). High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. Nature Medicine, 25(12), Article 12. https://doi.org/10.1038/s41591-019-0652-7

Reik, W., & Walter, J. (2001). Genomic imprinting: Parental influence on the genome. Nature Reviews Genetics, 2(1), Article 1. https://doi.org/10.1038/35047554

Ruan, S., Shi, J., Wang, M., & Zhu, Z. (2021). Analysis of Multiple Human Tumor Cases Reveals the Carcinogenic Effects of PKP3. Journal of Healthcare Engineering, 2021, 9391104. https://doi.org/10.1155/2021/9391104

Salati, S., Bianchi, E., Zini, R., Tenedini, E., Quaglino, D., Manfredini, R., & Ferrari, S. (2007). Eosinophils, but not neutrophils, exhibit an efficient DNA repair machinery and high nucleolar activity. Haematologica, 92(10), Article 10. https://doi.org/10.3324/haematol.11472

Sanchez, C., Snyder, M. W., Tanos, R., Shendure, J., & Thierry, A. R. (2018). New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. NPJ Genomic Medicine, 3, 31. https://doi.org/10.1038/s41525-018-0069-0

Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America, 103(5), 1412–1417. https://doi.org/10.1073/pnas.0510310103

Schmidt, B., Liebenberg, V., Dietrich, D., Schlegel, T., Kneip, C., Seegebarth, A., Flemming, N., Seemann, S., Distler, J., Lewin, J., Tetzner, R., Weickmann, S., Wille, U., Liloglou, T., Raji, O., Walshaw, M., Fleischhacker, M., Witt, C., & Field, J. K. (2010). SHOX2 DNA Methylation is a Biomarker for the diagnosis of lung cancer based on bronchial aspirates. BMC Cancer, 10(1), 600. https://doi.org/10.1186/1471-2407-10-600

Selamat, S. A., Galler, J. S., Joshi, A. D., Fyfe, M. N., Campan, M., Siegmund, K. D., Kerr, K. M., & Laird-Offringa, I. A. (2011). DNA Methylation Changes in Atypical Adenomatous Hyperplasia, Adenocarcinoma In Situ, and Lung Adenocarcinoma. PLOS ONE, 6(6), e21443. https://doi.org/10.1371/journal.pone.0021443

Shao, T., Song, P., Hua, H., Zhang, H., Sun, X., Kong, Q., Wang, J., Luo, T., & Jiang, Y. (2018). Gamma synuclein is a novel Twist1 target that promotes TGF-β-induced cancer cell migration and invasion. Cell Death & Disease, 9(6), Article 6. https://doi.org/10.1038/s41419-018-0657-z

Shen, H., & Laird, P. W. (2013). Interplay between the cancer genome and epigenome. Cell, 153(1), 38–55. https://doi.org/10.1016/j.cell.2013.03.008

Shen, S. Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M. H. A., Chadwick, D., Zuzarte, P. C., Borgida, A., Wang, T. T., Li, T., Kis, O., Zhao, Z., Spreafico, A., Medina, T. da S., Wang, Y., Roulois, D., Ettayebi, I., Chen, Z., Chow, S., … De Carvalho, D. D. (2018). Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature, 563(7732), 579–583. https://doi.org/10.1038/s41586-018-0703-0

Shimoda, N., Izawa, T., Yoshizawa, A., Yokoi, H., Kikuchi, Y., & Hashimoto, N. (2014). Decrease in cytosine methylation at CpG island shores and increase in DNA fragmentation during zebrafish aging. AGE, 36(1), 103–115. https://doi.org/10.1007/s11357-013-9548-5

Shivapurkar, N., Toyooka, S., Eby, M. T., Huang, C. X., Sathyanarayana, U. G., Cunningham, H. T., Reddy, J. L., Brambilla, E., Takahashi, T., Minna, J. D., Chaudhary, P. M., & Gazdar, A. F. (2002). Differential inactivation of caspase-8 in lung cancers. Cancer Biology & Therapy, 1(1), 65–69. https://doi.org/10.4161/cbt.1.1.45

Smith, Z. D., & Meissner, A. (2013). DNA methylation: Roles in mammalian development. Nature Reviews. Genetics, 14(3), 204–220. https://doi.org/10.1038/nrg3354

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. (2016). Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell, 164(0), 57–68. https://doi.org/10.1016/j.cell.2015.11.050

Soares, F. A. (1992). Increased numbers of pulmonary megakaryocytes in patients with arterial pulmonary tumour embolism and with lung metastases seen at necropsy. Journal of Clinical Pathology, 45(2), 140–142. https://doi.org/10.1136/jcp.45.2.140

Strichman-Almashanu, L. Z., Lee, R. S., Onyango, P. O., Perlman, E., Flam, F., Frieman, M. B., & Feinberg, A. P. (2002). A Genome-Wide Screen for Normally Methylated Human CpG Islands That Can Identify Novel Imprinted Genes. Genome Research, 12(4), 543–554. https://doi.org/10.1101/gr.224102

Sun, K., Jiang, P., Chan, K. C. A., Wong, J., Cheng, Y. K. Y., Liang, R. H. S., Chan, W., Ma, E. S. K., Chan, S. L., Cheng, S. H., Chan, R. W. Y., Tong, Y. K., Ng, S. S. M., Wong, R. S. M., Hui, D. S. C., Leung, T. N., Leung, T. Y., Lai, P. B. S., Chiu, R. W. K., & Lo, Y. M. D. (2015). Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proceedings of the National Academy of Sciences, 112(40), E5503–E5512. https://doi.org/10.1073/pnas.1508736112

Tanaka, H., Kanda, M., Shimizu, D., Tanaka, C., Inokawa, Y., Hattori, N., Hayashi, M., Nakayama, G., & Kodera, Y. (2022). Transcriptomic profiling on localized gastric cancer identified CPLX1 as a gene promoting malignant phenotype of gastric cancer and a predictor of recurrence after surgery and subsequent chemotherapy. Journal of Gastroenterology, 57(9), 640–653. https://doi.org/10.1007/s00535-022-01884-6

Tanaka, K., & Okamoto, A. (2007). Degradation of DNA by bisulfite treatment. Bioorganic & Medicinal Chemistry Letters, 17(7), 1912–1915. https://doi.org/10.1016/j.bmcl.2007.01.040

Tang, M., Xu, W., Wang, Q., Xiao, W., & Xu, R. (2009). Potential of DNMT and its Epigenetic Regulation for Lung Cancer Therapy. Current Genomics, 10(5), 336–352. https://doi.org/10.2174/138920209788920994

Teif, V. B., Beshnova, D. A., Vainshtein, Y., Marth, C., Mallm, J.-P., Höfer, T., & Rippe, K. (2014). Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. Genome Research, 24(8), 1285–1295. https://doi.org/10.1101/gr.164418.113

Teschendorff, A. E., Breeze, C. E., Zheng, S. C., & Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics, 18(1), 105. https://doi.org/10.1186/s12859-017-1511-5

Tian, Q.-Q., Xia, J., Zhang, X., Gao, B.-Q., & Wang, W. (2020). MiR-331-3p Inhibits Tumor Cell Proliferation, Metastasis, Invasion by Targeting MLLT10 in Non-Small Cell Lung Cancer. Cancer Management and Research, 12, 5749–5758. https://doi.org/10.2147/CMAR.S249686

Titcombe, P., Murray, R., Hewitt, M., Antoun, E., Cooper, C., Inskip, H. M., Holbrook, J. D., Godfrey, K. M., Lillycrop, K., Hanson, M., & Barton, S. J. (2022). Human non-CpG methylation patterns display both tissue-specific and inter-individual differences suggestive of underlying function. Epigenetics, 17(6), 653–664. https://doi.org/10.1080/15592294.2021.1950990

Titus, A. J., Gallimore, R. M., Salas, L. A., & Christensen, B. C. (2017). Cell-type deconvolution from DNA methylation: A review of recent applications. Human Molecular Genetics, 26(R2), R216–R224. https://doi.org/10.1093/hmg/ddx275

Troll, C. J., Kapp, J., Rao, V., Harkins, K. M., Cole, C., Naughton, C., Morgan, J. M., Shapiro, B., & Green, R. E. (2019). A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. BMC Genomics, 20(1), 1023. https://doi.org/10.1186/s12864-019-6355-0

Tuorto, F., Herbst, F., Alerasool, N., Bender, S., Popp, O., Federico, G., Reitter, S., Liebers, R., Stoecklin, G., Gröne, H.-J., Dittmar, G., Glimm, H., & Lyko, F. (2015). The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis. The EMBO Journal, 34(18), 2350–2362. https://doi.org/10.15252/embj.201591382

Ulz, P., Thallinger, G. G., Auer, M., Graf, R., Kashofer, K., Jahn, S. W., Abete, L., Pristauz, G., Petru, E., Geigl, J. B., Heitzer, E., & Speicher, M. R. (2016). Inferring expressed genes by whole-genome sequencing of plasma DNA. Nature Genetics, 48(10), 1273–1278. https://doi.org/10.1038/ng.3648

Vachtenheim, J., Horáková, I., & Novotná, H. (1994). Hypomethylation of CCGG sites in the 3′ region of H-ras protooncogene is frequent and is associated with H-ras allele loss in non-small cell lung cancer. Cancer Research, 54(5), 1145–1148.

Vaisvila, R., Ponnaluri, V. K. C., Sun, Z., Langhorst, B. W., Saleh, L., Guan, S., Dai, N., Campbell, M. A., Sexton, B. S., Marks, K., Samaranayake, M., Samuelson, J. C., Church, H. E., Tamanaha, E., Corrêa, I. R., Pradhan, S., Dimalanta, E. T., Evans, T. C., Williams, L., & Davis, T. B. (2021). Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Research, 31(7), 1280–1289. https://doi.org/10.1101/gr.266551.120

Vardimon, L., Kressmann, A., Cedar, H., Maechler, M., & Doerfler, W. (1982). Expression of a cloned adenovirus gene is inhibited by in vitro methylation. Proceedings of the National Academy of Sciences of the United States of America, 79(4), 1073–1077.

Vrba, L., Oshiro, M. M., Kim, S. S., Garland, L. L., Placencia, C., Mahadevan, D., Nelson, M. A., & Futscher, B. W. (2020). DNA methylation biomarkers discovered in silico detect cancer in liquid biopsies from non-small cell lung cancer patients. Epigenetics, 15(4), 419–430. https://doi.org/10.1080/15592294.2019.1695333

Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L., & Schübeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nature Genetics, 37(8), 853–862. https://doi.org/10.1038/ng1598

Werner, B., Yuwono, N. L., Henry, C., Gunther, K., Rapkins, R. W., Ford, C. E., & Warton, K. (2019). Circulating cell-free DNA from plasma undergoes less fragmentation during bisulfite treatment than genomic DNA due to low molecular weight. PLoS ONE, 14(10), e0224338. https://doi.org/10.1371/journal.pone.0224338

Wu, X.-Y., Chen, H.-C., Li, W.-W., Yan, J.-D., & Lv, R.-Y. (2020). DNMT1 promotes cell proliferation via methylating hMLH1 and hMSH2 promoters in EGFR-mutated non-small cell lung cancer. Journal of Biochemistry, 168(2), 151–157. https://doi.org/10.1093/jb/mvaa034

Xu, W., Lu, J., Zhao, Q., Wu, J., Sun, J., Han, B., Zhao, X., & Kang, Y. (2019). Genome-Wide Plasma Cell-Free DNA Methylation Profiling Identifies Potential Biomarkers for Lung Cancer. Disease Markers, 2019, 4108474. https://doi.org/10.1155/2019/4108474

Yang, H., Liu, Y., Bai, F., Zhang, J.-Y., Ma, S.-H., Liu, J., Xu, Z.-D., Zhu, H.-G., Ling, Z.-Q., Ye, D., Guan, K.-L., & Xiong, Y. (2013). Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. Oncogene, 32(5), 663–669. https://doi.org/10.1038/onc.2012.67

Yu, F., Makrigiorgos, A., Leong, K. W., & Makrigiorgos, G. M. (2021). Sensitive detection of microsatellite instability in tissues and liquid biopsies: Recent developments and updates. Computational and Structural Biotechnology Journal, 19, 4931–4940. https://doi.org/10.1016/j.csbj.2021.08.037

Zemmour, H., Planer, D., Magenheim, J., Moss, J., Neiman, D., Gilon, D., Korach, A., Glaser, B., Shemer, R., Landesberg, G., & Dor, Y. (2018). Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA. Nature Communications, 9(1), Article 1. https://doi.org/10.1038/s41467-018-03961-y

Zhang, T., Cooper, S., & Brockdorff, N. (2015). The interplay of histone modifications – writers that read. EMBO Reports, 16(11), 1467–1481. https://doi.org/10.15252/embr.201540945

Zhao, Z., & Shilatifard, A. (2019). Epigenetic modifications of histones in cancer. Genome Biology, 20(1), 245. https://doi.org/10.1186/s13059-019-1870-5

Zheng, J., Li, Z., Zhang, X., Zhang, H., Zhu, S., Sun, J., & Wang, Y. (2022). Comparison of dsDNA and ssDNA-based NGS library construction methods for targeted genome and methylation profiling of cfDNA (p. 2022.01.12.475986). bioRxiv. https://doi.org/10.1101/2022.01.12.475986

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., & Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature, 500(7463), Article 7463. https://doi.org/10.1038/nature12433

Zukowski, A., Rao, S., & Ramachandran, S. (2020). Phenotypes from cell-free DNA. Open Biology, 10(9), 200119. https://doi.org/10.1098/rsob.200119

5

FUTURE DIRECTIONS FOR ULTRASHORT SINGLE-STRANDED CELL-FREE DNA

5.1 Introduction

Throughout this thesis, we have demonstrated the presence of an additional species of cell-free DNA in plasma. In this final chapter, we discuss the promising future directions that can be taken for uscfDNA. These future directions can be categorized into two major directions. Firstly, one would be to examine the biogenesis of uscfDNA. Secondly, it would be interesting to explore further what potential clinical contributions uscfDNA can provide. Throughout this chapter, we will review the rationale of these potential directions and which strategies we can implement to examine them.

5.2 Potential Biological Origins and Mechanisms

5.2.1 uscfDNA Origins from Certain Cell Types

Although we attempted to deconvolution the tissue of origin in Chapter 4, one direction would be to further establish the profile of cell types that contribute to uscfDNA in the blood. As a control, we showed that the potential tissue of origin mncfDNA was mainly blood cell derived, which is consistent with previous studies (Caggiano et al., 2021; Moss et al., 2018). Our uscfDNA methylation data was suggestive that uscfDNA may have origins from blood cells, although the profile differed from mncfDNA. Our methylation data tissue of origin reveals that uscfDNA could potentially derive from eosinophils, neutrophils, or monocytes. However, this analysis was built on a methylation deconvolution platform based on mncfDNA. This approach can be refined with an algorithm and database designed for uscfDNA deconvolution with validated controls.

Another hypothesis is that uscfDNA could originate from extracellular traps. Neutrophils and eosinophils release extracellular traps made of modified chromatin and bactericidal proteins (de Bont et al., 2019; Ueki et al., 2013). This subset of circulating nucleic acids expectorated into the blood may possess different genomic traits than the internal genome of neutrophils or eosinophils. One previous study utilized a cultured model to sequence the released ejected netosis DNA, referring to it as the "netome" (Scieszka et al., 2022). A potential study could be to analyze the genomic characteristics of netosis samples to see if it contains traits that closer resemble that of uscfDNA, mncfDNA, or neither. A similar approach could be performed with a sequenced sample from eosinophilic extracellular trap DNA, although this sequenced resource is not readily available. Therefore, constructing a similar model in cell culture for eosinophils would be valuable to generate this resource for comparison.

5.2.2 uscfDNA Potential Relation to R-loops

In addition to attempting to allocate the origin of uscfDNA from certain cells, it may be helpful to determine if uscfDNA comes from certain genomic processes. One such candidate process may be uscfDNA's involvement with R-loops. R-loops are unique triple-stranded nucleic acid complexes that consist of a DNA:RNA hybrid and a displaced single-stranded DNA (Hegazy et al., 2020). R-loops occur in both bacteria and mammals (Aguilera & García-Muse, 2012) and can form during a variety of biological circumstances, including transcription regulation (Grunseich et al., 2018), DNA damage repair (Bhatia et al., 2014; Lang et al., 2017), and regulation of chromatin landscape (García-Pichardo et al., 2017). Recent evidence has demonstrated that aberrations in R-loop formation are involved in human diseases such as neurological disorders, autoimmune diseases, and cancer.

The DNA:RNA hybrid displaces single-stranded DNA and is more structurally stable than corresponding dsDNA complementary structures. It has been shown that degradation of the RNA strand by RNase was necessary for the resolution of the R-loop structure (Wahba et al., 2011). Various factors promote R-loop formation and stabilization. R-loop formation is more efficient in G-rich strands, and DNA secondary structures such as G-Quadruplexes can form on the displaced strand of DNA (Sundquist & Klug, 1989). Nicks in the DNA downstream of the promoter can also favor R-loop formation by preventing potent reannealing of unwound duplex DNA (Roy et al., 2010).

Due to the single-stranded nature, enrichment in promoter sequences, and presence of potential sequences G-Quadruplexes, it could be hypothesized that uscfDNA could be related to R-loop from DNA: RNA hybrids. Although it is unclear if uscfDNA is derived from the excision of the non-template ssDNA strand or the template DNA:RNA hybrid strand during the resolution steps of the R-loop. In Chapter 3, we observed that the G-Quadruplex sequences are equally associated with the primary and theoretical complement strand, suggesting that uscfDNA could be derived from either strand of the R-loop complex. Studies have indicated that the displaced ssDNA in the R-loop is prone to single base damage and ssDNA nucleases, which generate nicks in the DNA (Freudenreich, 2018). This exposed ssDNA undergoes deamination to uracil (Stavnezer et al., 2008) and oxidative damage (Entezam et al., 2010). This exposure of the non-template strand is sometimes part of a complex mechanism to promote class-switch recombination for antibody gene diversification (Yu et al., 2003). Hence in non-cancer, the presence of uscfDNA in circulation in blood could be indicative of proper maintenance of R-loops and immune response.

However, this same tendency for DNA damage on the non-template strand can also generate deleterious outcomes if not controlled properly. Accumulating oxidative or deamination damage to the non-template ssDNA strand could lead to DNA breaks, R-loop displacements, DNA nicks, or gaps contributing to genomic instability. Therefore, a relationship between increasing genomic stability could be monitored by changes in the profile of uscfDNA.

5.2.3 uscfDNA as Part of DNA Repair

Additionally, there are several kinds of DNA repair pathways that cells can undergo when DNA is damaged, which may generate an excised DNA molecules or strands. Whether this excised DNA is released into the cytoplasm and eventually the adjacent cell-free environment is currently unclear. These include base excision repair, nucleotide excision repair, direct repair, mismatch repair, base-excision repair, nucleotide-excision repair, homologous recombination, and non-homologous end-joining (Postel-Vinay et al., 2012). Nucleotide excision repair has been shown to remove oligonucleotides from 24-32nt in length (Wakasugi & Sancar, 1998). Studies in mismatch repair using Escherichia coli indicate that mismatch repair ssDNA excision can extend as far as 45bp (Liu et al., 2019). Base excision repair removes a single modified base but has also been shown only to repair patches of 2-6 and 6-12 nucleotides in cell lines (Sattler et al., 2003). Although these studies have been performed in an assortment of in-vitro models, it is plausible that these excised single-stranded DNA segments could be released in circulation in a more complex organism. Interestingly, a study using Chinese hamster ovary (CHO) and HeLa cell models showed that the excised ssDNA from the nucleotide excision repair mechanism is first bound to transcription factor IIH and either released through an ATP-dependent mechanism and eventually associated with single-stranded binding protein replication protein A or targeted by nucleases (Kemp et al., 2012). This study provides a mechanism of release and fate of excised nucleotides and could be related to the existence of uscfDNA in plasma. Additionally, the lengths of excised single-stranded DNA could also be dependent on the organism, and the 40-70nt of the uscfDNA reflects a process that occurs only in humans and not in other organisms. Examining the end-motifs of excised ssDNA between uscfDNA and known DNA repair mechanisms could help resolve this hypothesis if the same end-motif profiles of these DNA repair mechanisms are well-defined.

5.2.4 uscfDNA May be Derived from Repeat Elements

Repeat DNA is defined as DNA sequences present in multiple copies in the genome. Over 50% of the human genome comprises DNA repeats (Lander et al., 2001). Repeat DNA elements can be subdivided into two groups; the first are tandem repeats, and the other are interspersed repeats. Tandem repeats account for only 6% of the genome and are repetitions of the same sequence in a head-to-tail orientation present over heterochromatin and centromeric regions (Trost et al., 2020). Tandem repeats are further categorized as microsatellites, minisatellites, centromeric/pericentric satellites, and telomeric/subtelomeric repeats (Gezer et al., 2022). Microsatellites are repeats of 1-9 nucleotide motifs and are prone to mutations and instability associated with colorectal cancer (Boland & Goel, 2010). In contrast, minisatellites are GC-rich repeats consisting of 10-100bp motifs, also associated with high mutation rates ranging from 0.5% to >20% (Bois, 2003). Satellite sequences are also in present centromeric/pericentric regions accounting for 3% of the human genome, which is involved in chromosome organization, segregation, kinetochore formation, and regulation of the heterochromatin (Pezer et al., 2012). α-satellites and human satellite 2 (HSATII) (Hall et al., 2017) are repetitive 171bp sequences and 26bp repeats, respectively, in centromeric/pericentric regions, which aberrated are present in tumor cells leading to expansion of genomic copy numbers (Bersani et al., 2015).

In contrast to tandem repeats, interspersed repeats constitute 45% of the human genome and are thought to result from retro transposable elements inserting themselves into new genomic regions throughout evolution (Criscione et al., 2014). Retro transposable elements can be classified into long terminal repeats, which are identical sequences of DNA several hundred bp long, and shorter non-long terminal repeats made up of long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINESs). SINES can be further categorized into the LINE-1 and ALU families (Levin & Moran, 2011).

Combined, both types of repeats make up more than half of the human genome. It would be logical that cfDNA fragments are also derived from these abundant elements. In Chapter 4, CpG sites in uscfDNA fragments were associated with SINES, LINES, and simple repeats (microsatellites). Interestingly, uscfDNA appears elevated in microsatellites compared to mncfDNA. A more granularized analysis of the repetitive sequence profile of uscfDNA provides another dimension of characterization. Since DNA repeats are prone to mutations and genomic instability, DNA repair mechanisms may be in play to prevent the aberrations from destabilizing the genomic integrity. The assortment of DNA repair mechanisms explained earlier may be connected to the release of DNA strands into the circulation.   Many tools are available to identify and profile repetitive elements in next-generation sequencing DNA data (Novák et al., 2010; Tarailo-Graovac & Chen, 2009). This analysis may also be useful as another potential feature for differentiating clinical cohorts using uscfDNA.

5.2.5 Relationship of uscfDNA in Extracellular Vesicles

It is still ambiguous whether cell-free DNA is circulating in a naked form, bound to protein, or packaged in extracellular vesicles (exosomes/microvesicles/apoptotic bodies). Several studies have attempted to clarify this relationship, but there are contrasting findings. One study suggests that the majority of mncfDNA are found in exosome fractions(Fernando et al., 2017). Alternatively, other studies indicate that mncfDNA is present in both exosomes, and a major fraction is in the supernatant (Sun et al., 2021). Interestingly, they claimed that the supernatant component demonstrated a higher signal of somatic mutations-containing cfDNA than the other fractions, although the exosome fraction still contained a portion ctDNA signal. Another report indicated similar findings concluding that mncfDNA is either bound to proteins or associated with extracellular vesicles (Moldovan et al., 2022). Similarly, they described that tumor-derived cfDNA is not enriched in exosomes compared to circulation.

Repeating similar fractionation studies in plasma except with a focus on the residency of uscfDNA could be an informative study. Plasma and other samples can be fractionated into different extracellular vesicles by a variety of methods (Konoshenko et al., 2018). These different fractions can undergo BRcfDNA-Seq and be analyzed for size-distribution analysis of the cfDNA present. Further, if uscfDNA is present in different fractions, we can establish genomic characteristic profiles by analyzing the subsequent NGS data. For example, perhaps a certain population of uscfDNA is present only in extracellular vesicles such as those associated with the DNA repair pathway or from nucleosome occupancy.

5.3 Scientific Experimental Models

5.3.1 Cellular Models

Thus far, the limitation of cfDNA-based research is the requirement to work on human biospecimens, which are both limited in volume and variable-rich. If a vivo model system, such as cell-model, can provide equivalent information to cfDNA in a biofluid, it would allow for fundamental mechanism experiments. Several studies have studied the media of cell models for mncfDNA characteristics with the understanding that the majority of mncfDNA comes from cell death or apoptosis. Cell culture models have been shown to release mncfDNA into the supernatant media (Bronkhorst et al., 2016, 2019; Ungerer et al., 2022). These studies show that it is viable for cell-free DNA to be studied using a conditioned media model. In addition to mncfDNA, the conditioned media also contains many sizes of DNA, ranging bands from 3000bp to the 100bp showing evidence of mncfDNA and di-nucleosomal cell-free DNA. Most of these studies did not perform NGS sequencing, so the genomic details of the cfDNA were not examined. Resultingly, there was limited information about the existence of uscfDNA in these models. The BRcfDNA-Seq technique (low molecular weight cfDNA extraction and single-stranded library kit to visualize) would be required to answer the question about the presence of uscfDNA. Subsequent experiments testing double-stranded library kits versus single-stranded library kits and digestion assays with ssDNA-specific and dsDNA-specific enzymes would aid in demonstrating resemblance to uscfDNA observed in plasma. It is highly possible that supernatant media is not as complex as plasma since it only reflects one cell type growing in an artificial environment. In contrast, plasma has more complex interactions between different cell types and organ sites that can influence the appearance of cfDNA. Since we hypothesize that uscfDNA may come from immune blood type cells, culturing immune cell may be a more useful model. Regardless, if any cellular model indicates uscfDNA is reproducible, further cellular model manipulations, such as nuclease enzyme activity knockdown/knockout within the cells, may aid in revealing more about the biogenesis of uscfDNA and mncfDNA.

5.3.2 Animals Models

Due to the complexity of simulating plasma with cell models, other research groups have attempted to look at cfDNA in animal models. Early experiments used animals to demonstrate the kinetics of clearance of injected nucleic acids in mice (Chused et al., 1972; Emlen & Mannik, 1978; Gosse et al., 1965). Recently, cell-free DNA experiments studying nuclease knockout models for dnase1L3 and dNase1 were published to observe how the mncfDNA patterns were affected(Han & Lo, 2021). Aside from looking at the size distribution, they also examined the end-motif changes of the rodent cfDNA, providing a more granular inspection of the genomic characteristics. Hence, seeing that mncfDNA is present in animals, a parallel search for uscfDNA would open up similar nuclease studies. However, similar to Chapter 2, we would need to test if uscfDNA in animal models would be equivalent enough to proceed. In the supplementary data in an adjacent uscfDNA publication where they used a direct pull-down method in biofluids(Cheng et al., 2022), they also tested the plasma of bovine, pig, and rabbit. Interestingly, the apparent fragment curves for bovine plasma showed a peak at 62.5nt, while pig plasma showed a peak at 25nt. Since their method used a direct DNA pull-down with beads, it differed from the BRcfDNA-Seq protocol. If these animal studies were replicated with this system and certain animals showed different size distributions of cell-free DNA populations, it may indicate some details about mammal evolution.

5.4 Existence of uscfDNA in Other Biofluids

In this thesis, we have exclusively explored plasma which appears to showcase a unique uscfDNA population at ~50nt in addition to the 167bp mncfDNA. One observation is that conventionally plasma comes from the venous blood for convenience and safety. Other scientists have explored differences between capillary and venous plasma and described that the size distribution is similar (Breitbach et al., 2014; Ehrich et al., 2023). They showed a 50% lower cfDNA fraction in the venous fraction, which may be alluded to by the abundance of observed lymphocytes in the arterioles (Yang et al., 2001). The appearance of cfDNA in arterial blood could differ since there may be distinct processes at different physiological regions of the body, such as the properties of cfDNA before and after the lungs, liver, and kidneys. The liver and kidneys have been described as regions of degradation and clearance for nucleic acids (Botezatu et al., 2000; Gauthier et al., 1996). Therefore, these studies would reveal the differences in the pattern between the arterial, capillary, and venous blood filtration or metabolism that occurs through circulation through these organs.

Analysis of other biofluids has also demonstrated different cfDNA fragmentation patterns. Many groups have described urine as a biomarker for bladder cancer. The fragment size appears to have a mncfDNA region and an abundant region below 70-100bp (Chen et al., 2022; Mouliere et al., 2021).

Another interesting biofluid would be to explore the cell-free DNA pattern of supernatant saliva. In a preliminary study performed in our lab, processing the supernatant using BRcfDNA-Seq demonstrated that the fragmentation pattern of saliva contains mncfDNA and a distinct spiked profile of lower molecular weight DNA from 50-100bp. The spiked peaks demonstrate 10.4 bp periodicities indicating DNase1 susceptibility around the nucleosomes (Klug & Lutter, 1981). Interestingly, unlike plasma, there was no singular peak at ~50nt. In contrast, there was greater cfDNA density from 50-100bp, which was not seen in plasma. Therefore, like urine, there may be differences in DNA metabolism that occur in the saliva. This may be derived from blood circulating in a closed and well-monitored system. In contrast, saliva interacts with an external environment where food, bacteria, and oral cells interact in a dynamic environment. Additionally, in blood, if immune cells detect abundant cfDNA, they can associate it with a danger-associated molecular pattern (DAMP) and move to remove it (Stortz et al., 2019). Therefore, unlike saliva, blood may have more reactive cells and behavior to maintain a homeostatic environment leading to the observed cfDNA pattern.

Another biofluid of interest to analyze with BRcfDNA-Seq would be cerebral spinal fluid (CSF). CSF has been proposed as an alternative biofluid medium for brain cancer ctDNA analysis (De Mattos-Arruda et al., 2015; Mouliere et al., 2018). There are limited studies on CSF fragment size distribution, but it appears that, unlike plasma, CSF has a peak at 133bp, and a large proportion of cfDNA is smaller than 150bp. This could also indicate differences in nuclease metabolism with the cell-free DNA worth exploring.

5.5 Clinical Applications

5.5.1 Diagnostic and Classification Potential

In Chapters 3 and 4, we have demonstrated proof of concept that uscfDNA has distinct characteristics such as functional peaks, G-Quad prevalence, fragmentomic patterns, and methylation patterns that can be potential biomarkers for late-stage NSCLC differentiation. The pipeline and analysis would need to be performed on an increased sample size to validate these observations. Additionally, although interesting, there is minimal clinical need to screen for the presence of late-stage advanced cancers through liquid biopsy. There is a greater need for finding new biomarkers for early stages. Thus, applying BRcfDNA-Seq and uscfDNA for screening for early stages of NSCLC is an urgent direction to be explored.

Another interesting direction would be to examine how uscfDNA contributes to cancer detection at other anatomic sites. For somatic mutation-containing ctDNA, different cancer types are associated with different amounts of ctDNA(Bettegowda et al., 2014). For example, another study examining 10,000 patients shows that the ctDNA concentration is highest in small-cell lung carcinoma and lowest in thyroid and renal cancers (Zhang et al., 2021). Therefore, non-somatic mutation characteristics built into uscfDNA may eventually be as helpful as mncfDNA fragmentomics in various cancer types (Thierry, 2023). This analysis approach has already demonstrated efficacy in various cancer types (Cristiano et al., 2019) with several follow-up studies showing promise in lung cancer(Wang et al., 2023), osteosarcoma (Udomruk et al., 2023), and HCC(Foda et al., 2022).

5.5.2 Therapy Monitoring

There is also potential that uscfDNA also contributes to monitoring the efficacy of therapy. Precision medicine in lung cancer (Politi & Herbst, 2015) has become an established treatment route where the biomarker (genetic or protein-based) profile of the tissue tumor determines treatment decisions. Several studies have tested the use of somatic mutation containing mononucleosomal-sized ctDNA as an alternative to repeat tumor tissue biopsy genotyping (Assaf et al., 2023; Kim et al., 2021; Murtaza et al., 2013; Oxnard et al., 2014). It is possible that aspects of uscfDNA can provide alternative metrics to aid in this process. However, since the BRcfDNA-Seq is not designed to analyze specific mutations from a clinically relevant panel, providing information regarding an increase or decrease of ctDNA variant allele frequency changes with tumor size is challenging. Interestingly fragmentomics (which does not consider somatic mutation in cfDNA) has demonstrated the ability to monitor treatment (Cristiano et al., 2019). Analysis of fragmentomic patterns of genes of interest has also shown a correlation with patients treated with PD-(L)1 immune-checkpoint inhibitors clinical response (Esfahani et al., 2022). Therefore, there is potential for these characteristics of uscfDNA (G-Quadruplex prevalence or functional elements) described to be used as biomarkers for therapy monitoring.

5.6 Required Technical Advancements

Several technical advancements would be helpful for further uscfDNA studies. A priority would be a method to quantify the amount of uscfDNA in a sample accurately. Currently, we determine the uscfDNA to mncfDNA relationship by measuring the number or reads that are categorized as uscfDNA (40-70bp) and mncfDNA (120-250bp). However, this relationship is only relative and based on bioinformatic analysis. A direct measure would also be very useful.

There are challenges to direct uscfDNA detection due to its short and single-stranded nature. DNA extraction methods are mainly based on intercalating dye with the helix of the nucleic acid strands or a quantifiable PCR reaction for housekeeping genes. To add to this complexity, plasma is composed of a complex mixture of uscfDNA, mncfDNA, di-, tri-nucleosomal cell-free DNA, large molecular weight genomic DNA, and even extracellular RNA species. Current cfDNA measurements must be specific and avoid over or underestimating the designated target for quantification. Many "ssDNA" fluorescent dyes on the market can still bind to dsDNA if present, making them non-specific in mixed samples (Nakayama et al., 2016). Thus, dye development may be required to develop useful uscfDNA assays.

This capability would allow a researcher to determine if uscfDNA concentration changes during physiological conditions. In previous literature, researchers have looked at several questions regarding physiology. For alcohol and menstruation, there have been no conclusive on whether mncfDNA concentration changes. For acute exercise, studies have reproducibly demonstrated sharp increases in cfDNA. Pesticides have also been shown to be associated with increased cfDNA. Ionization radiation should have decreased in cfDNA. However, inconsistent cfDNA changes were observed for smoking, BMI, hypertension, circadian rhythm, gender, age, and chronic exercise (Yuwono et al., 2021). There is a possibility that the uscfDNA present could be monitored in these other non-cancer applications to provide another layer of information.

5.7 Bioinformatic Questions

Our proof of uscfDNA being single-stranded is through deductive digestions and contrasting ssDNA and dsDNA libraries library preparations as per Chapter 2. Another strategy will be to bioinformatically determine if the uscfDNA has complementary sequences. This would require looking for plausible fragments in uscfDNA that could have been originally paired together. The BRcfDNA-Seq pipeline uses a single-stranded library preparation. During that process, DNA in the sample is heat-denatured and separated prior to adapter ligation. The mncfDNA could be used as a control since it is well understood that they are double-stranded. The concept of looking at complementary sequences in samples has previously been attempted in the analysis of ancient DNA. Several bioinformatic attempts have been made by other groups to find bioinformatic strategies to find mates (Bokelmann et al., 2020). For mncfDNA, which is dsDNA, the probability of recovering the other strand depends on the efficiency of the library (Bokelmann et al., 2020; Gansauge & Meyer, 2013). The development of MatchSeq was intended to computationally reconstruct double-stranded DNA from individual DNA strands within the sequencing file. This strategy, however, had several limitations. Firstly, they require the single-stranded library to be sequenced deeply to ensure that each unique DNA strand will appear digitally. Secondly, libraries should be prepared from small quantities of DNA so that the complexity of the DNA is limited to ensure a higher chance of finding the authentic dsDNA mate. These limitations make it challenging to apply to cfDNA immediately, but it could be a potentially interesting direction.

5.8 Conclusion

In closing, the future of cell-free DNA and uscfDNA has many possible directions. In this chapter, we have overviewed some directions that uscfDNA-oriented work can tread towards. However, fundamental work needs to be established at the biological, molecular, and bioinformatic levels to aid in the further progress of uscfDNA development. These advances would aid in many ways, both in contributing to the biological understanding of cfDNA physiology and for uscfDNA to potentially be a significant player in the liquid biopsy tool kit for clinicians and patients in the future.

5.9 References

Aguilera, A., & García-Muse, T. (2012). R loops: From transcription byproducts to threats to genome stability. Molecular Cell, 46(2), 115–124. https://doi.org/10.1016/j.molcel.2012.04.009

Assaf, Z. J. F., Zou, W., Fine, A. D., Socinski, M. A., Young, A., Lipson, D., Freidin, J. F., Kennedy, M., Polisecki, E., Nishio, M., Fabrizio, D., Oxnard, G. R., Cummings, C., Rode, A., Reck, M., Patil, N. S., Lee, M., Shames, D. S., & Schulze, K. (2023). A longitudinal circulating tumor DNA-based model associated with survival in metastatic non-small-cell lung cancer. Nature Medicine, 29(4), 859–868. https://doi.org/10.1038/s41591-023-02226-6

Bersani, F., Lee, E., Kharchenko, P. V., Xu, A. W., Liu, M., Xega, K., MacKenzie, O. C., Brannigan, B. W., Wittner, B. S., Jung, H., Ramaswamy, S., Park, P. J., Maheswaran, S., Ting, D. T., & Haber, D. A. (2015). Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. Proceedings of the National Academy of Sciences of the United States of America, 112(49), 15148–15153. https://doi.org/10.1073/pnas.1518008112

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., Antonarakis, E. S., Azad, N. S., Bardelli, A., Brem, H., Cameron, J. L., Lee, C. C., Fecher, L. A., Gallia, G. L., Gibbs, P., … Diaz, L. A. (2014). Detection of circulating tumor DNA in early- and late-stage human malignancies. Science Translational Medicine, 6(224), 224ra24. https://doi.org/10.1126/scitranslmed.3007094

Bhatia, V., Barroso, S. I., García-Rubio, M. L., Tumini, E., Herrera-Moyano, E., & Aguilera, A. (2014). BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. Nature, 511(7509), 362–365. https://doi.org/10.1038/nature13374

Bois, P. R. J. (2003). Hypermutable minisatellites, a human affair? Genomics, 81(4), 349–355. https://doi.org/10.1016/s0888-7543(03)00021-1

Bokelmann, L., Glocke, I., & Meyer, M. (2020). Reconstructing double-stranded DNA fragments on a single-molecule level reveals patterns of degradation in ancient samples. Genome Research, 30(10), 1449–1457. https://doi.org/10.1101/gr.263863.120

Boland, C. R., & Goel, A. (2010). Microsatellite Instability in Colorectal Cancer. Gastroenterology, 138(6), 2073-2087.e3. https://doi.org/10.1053/j.gastro.2009.12.064

Botezatu, I., Serdyuk, O., Potapova, G., Shelepov, V., Alechina, R., Molyaka, Y., Ananév, V., Bazin, I., Garin, A., Narimanov, M., Knysh, V., Melkonyan, H., Umansky, S., & Lichtenstein, A. (2000). Genetic analysis of DNA excreted in urine: A new approach for detecting specific genomic DNA sequences from cells dying in an organism. Clinical Chemistry, 46(8 Pt 1), 1078–1084.

Breitbach, S., Sterzing, B., Magallanes, C., Tug, S., & Simon, P. (2014). Direct measurement of cell-free DNA from serially collected capillary plasma during incremental exercise. Journal of Applied Physiology (Bethesda, Md.: 1985), 117(2), 119–130. https://doi.org/10.1152/japplphysiol.00002.2014

Bronkhorst, A. J., Ungerer, V., & Holdenrieder, S. (2019). Comparison of methods for the quantification of cell-free DNA isolated from cell culture supernatant. Tumor Biology, 41(8), 1010428319866369. https://doi.org/10.1177/1010428319866369

Bronkhorst, A. J., Wentzel, J. F., Aucamp, J., van Dyk, E., du Plessis, L., & Pretorius, P. J. (2016). Characterization of the cell-free DNA released by cultured cancer cells. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1863(1), 157–165. https://doi.org/10.1016/j.bbamcr.2015.10.022

Caggiano, C., Celona, B., Garton, F., Mefford, J., Black, B. L., Henderson, R., Lomen-Hoerth, C., Dahl, A., & Zaitlen, N. (2021). Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. Nature Communications, 12(1), Article 1. https://doi.org/10.1038/s41467-021-22901-x

Chen, M., Chan, R. W. Y., Cheung, P. P. H., Ni, M., Wong, D. K. L., Zhou, Z., Ma, M.-J. L., Huang, L., Xu, X., Lee, W.-S., Wang, G., Lui, K. O., Lam, W. K. J., Teoh, J. Y. C., Ng, C.-F., Jiang, P., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2022). Fragmentomics of urinary cell-free DNA in nuclease knockout mouse models. PLoS Genetics, 18(7), e1010262. https://doi.org/10.1371/journal.pgen.1010262

Cheng, L. Y., Dai, P., Wu, L. R., Patel, A. A., & Zhang, D. Y. (2022). Direct capture and sequencing reveal ultra-short single-stranded DNA in biofluids. IScience, 25(10), 105046. https://doi.org/10.1016/j.isci.2022.105046

Chused, T. M., Steinberg, A. D., & Talal, N. (1972). The clearance and localization of nucleic acids by New Zealand and normal mice. Clinical and Experimental Immunology, 12(4), 465–476.

Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M., & Neretti, N. (2014). Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics, 15(1), 583. https://doi.org/10.1186/1471-2164-15-583

Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., Jensen, S. Ø., Medina, J. E., Hruban, C., White, J. R., Palsgrove, D. N., Niknafs, N., Anagnostou, V., Forde, P., Naidoo, J., Marrone, K., Brahmer, J., Woodward, B. D., Husain, H., … Velculescu, V. E. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature, 1. https://doi.org/10.1038/s41586-019-1272-6 de Bont, C. M., Boelens, W. C., & Pruijn, G. J. M. (2019). NETosis, complement, and coagulation: A triangular relationship. Cellular & Molecular Immunology, 16(1), Article 1. https://doi.org/10.1038/s41423-018-0024-0

De Mattos-Arruda, L., Mayor, R., Ng, C. K. Y., Weigelt, B., Martínez-Ricarte, F., Torrejon, D., Oliveira, M., Arias, A., Raventos, C., Tang, J., Guerini-Rocco, E., Martínez-Sáez, E., Lois, S., Marín, O., de la Cruz, X., Piscuoglio, S., Towers, R., Vivancos, A., Peg, V., … Seoane, J. (2015). Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. Nature Communications, 6, 8839. https://doi.org/10.1038/ncomms9839

Ehrich, M., Sagaser, K. G., Porreco, R. P., Bellesheim, D., Patil, A. S., Shulman, L. P., & Van Den Boom, D. (2023). Capillary blood collection: Exploring a new method to promote noninvasive prenatal screening access. American Journal of Obstetrics and Gynecology, S0002-9378(23)00153-9. https://doi.org/10.1016/j.ajog.2023.03.008

Emlen, W., & Mannik, M. (1978). Kinetics and mechanisms for removal of circulating single-stranded DNA in mice. The Journal of Experimental Medicine, 147(3), 684–699. https://doi.org/10.1084/jem.147.3.684

Entezam, A., Lokanga, A. R., Le, W., Hoffman, G., & Usdin, K. (2010). Potassium bromate, a potent DNA oxidizing agent, exacerbates germline repeat expansion in a Fragile X premutation mouse model. Human Mutation, 31(5), 611–616. https://doi.org/10.1002/humu.21237

Esfahani, M. S., Hamilton, E. G., Mehrmohamadi, M., Nabet, B. Y., Alig, S. K., King, D. A., Steen, C. B., Macaulay, C. W., Schultz, A., Nesselbush, M. C., Soo, J., Schroers-Martin, J. G., Chen, B., Binkley, M. S., Stehr, H., Chabon, J. J., Sworder, B. J., Hui, A. B.-Y., Frank, M. J., … Alizadeh, A. A. (2022). Inferring gene expression from cell-free DNA fragmentation profiles. Nature Biotechnology, 40(4), 585–597. https://doi.org/10.1038/s41587-022-01222-4

Fernando, M. R., Jiang, C., Krzyzanowski, G. D., & Ryan, W. L. (2017). New evidence that a large proportion of human blood plasma cell-free DNA is localized in exosomes. PLOS ONE, 12(8), e0183915. https://doi.org/10.1371/journal.pone.0183915

Foda, Z. H., Annapragada, A. V., Boyapati, K., Bruhm, D. C., Vulpescu, N. A., Medina, J. E., Mathios, D., Cristiano, S., Niknafs, N., Luu, H. T., Goggins, M. G., Anders, R. A., Sun, J., Mehta, S. H., Thomas, D. L., Kirk, G. D., Adleff, V., Phallen, J., Scharpf, R. B., … Velculescu, V. E. (2022). Detecting liver cancer using cell-free DNA fragmentomes. Cancer Discovery, CD-22-0659. https://doi.org/10.1158/2159-8290.CD-22-0659

Freudenreich, C. H. (2018). R-loops: Targets for Nuclease Cleavage and Repeat Instability. Current Genetics, 64(4), 789–794. https://doi.org/10.1007/s00294-018-0806-z

Gansauge, M.-T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nature Protocols, 8(4), Article 4. https://doi.org/10.1038/nprot.2013.038

García-Pichardo, D., Cañas, J. C., García-Rubio, M. L., Gómez-González, B., Rondón, A. G., & Aguilera, A. (2017). Histone Mutants Separate R Loop Formation from Genome Instability Induction. Molecular Cell, 66(5), 597-609.e5. https://doi.org/10.1016/j.molcel.2017.05.014

Gauthier, V. J., Tyler, L. N., & Mannik, M. (1996). Blood clearance kinetics and liver uptake of mononucleosomes in mice. Journal of Immunology (Baltimore, Md.: 1950), 156(3), 1151–1156.

Gezer, U., Bronkhorst, A. J., & Holdenrieder, S. (2022). The Utility of Repetitive Cell-Free DNA in Cancer Liquid Biopsies. Diagnostics (Basel, Switzerland), 12(6), 1363. https://doi.org/10.3390/diagnostics12061363

Gosse, C., Pecq, J. B. L., Defrance, P., & Paoletti, C. (1965). Initial Degradation of Deoxyribonucleic Acid after Injection in Mammals. Cancer Research, 25(6 Part 1), 877–883.

Grunseich, C., Wang, I. X., Watts, J. A., Burdick, J. T., Guber, R. D., Zhu, Z., Bruzel, A., Lanman, T., Chen, K., Schindler, A. B., Edwards, N., Ray-Chaudhury, A., Yao, J., Lehky, T., Piszczek, G., Crain, B., Fischbeck, K. H., & Cheung, V. G. (2018). Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. Molecular Cell, 69(3), 426-437.e7. https://doi.org/10.1016/j.molcel.2017.12.030

Hall, L. L., Byron, M., Carone, D. M., Whitfield, T. W., Pouliot, G. P., Fischer, A., Jones, P., & Lawrence, J. B. (2017). Demethylated HSATII DNA and HSATII RNA Foci Sequester PRC1 and MeCP2 into Cancer-Specific Nuclear Bodies. Cell Reports, 18(12), 2943–2956. https://doi.org/10.1016/j.celrep.2017.02.072

Han, D. S. C., & Lo, Y. M. D. (2021). The Nexus of cfDNA and Nuclease Biology. Trends in Genetics: TIG, 37(8), 758–770. https://doi.org/10.1016/j.tig.2021.04.005

Hegazy, Y. A., Fernando, C. M., & Tran, E. J. (2020). The balancing act of R-loop biology: The good, the bad, and the ugly. The Journal of Biological Chemistry, 295(4), 905–913. https://doi.org/10.1074/jbc.REV119.011353

Kemp, M. G., Reardon, J. T., Lindsey-Boltz, L. A., & Sancar, A. (2012). Mechanism of release and fate of excised oligonucleotides during nucleotide excision repair. The Journal of Biological Chemistry, 287(27), 22889–22899. https://doi.org/10.1074/jbc.M112.374447

Kim, C., Xi, L., Cultraro, C. M., Wei, F., Jones, G., Cheng, J., Shafiei, A., Pham, T. H.-T., Roper, N., Akoth, E., Ghafoor, A., Misra, V., Monkash, N., Strom, C., Tu, M., Liao, W., Chia, D., Morris, C., Steinberg, S. M., … Guha, U. (2021). Longitudinal Circulating Tumor DNA Analysis in Blood and Saliva for Prediction of Response to Osimertinib and Disease Progression in EGFR-Mutant Lung Adenocarcinoma. Cancers, 13(13), 3342. https://doi.org/10.3390/cancers13133342

Klug, A., & Lutter, L. C. (1981). The helical periodicity of DNA on the nucleosome. Nucleic Acids Research, 9(17), 4267–4284. https://doi.org/10.1093/nar/9.17.4267

Konoshenko, M. Y., Lekchnov, E. A., Vlassov, A. V., & Laktionov, P. P. (2018). Isolation of Extracellular Vesicles: General Methodologies and Latest Trends. BioMed Research International, 2018, 8545347. https://doi.org/10.1155/2018/8545347

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860–921. https://doi.org/10.1038/35057062

Lang, K. S., Hall, A. N., Merrikh, C. N., Ragheb, M., Tabakh, H., Pollock, A. J., Woodward, J. J., Dreifus, J. E., & Merrikh, H. (2017). Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis. Cell, 170(4), 787-799.e18. https://doi.org/10.1016/j.cell.2017.07.044

Levin, H. L., & Moran, J. V. (2011). Dynamic interactions between transposable elements and their hosts. Nature Reviews. Genetics, 12(9), 615–627. https://doi.org/10.1038/nrg3030

Liu, J., Lee, R., Britton, B. M., London, J. A., Yang, K., Hanne, J., Lee, J.-B., & Fishel, R. (2019). MutL sliding clamps coordinate exonuclease-independent Escherichia coli mismatch repair. Nature Communications, 10(1), Article 1. https://doi.org/10.1038/s41467-019-13191-5

Moldovan, N., Verkuijlen, S., Pol, Y. van der, Bosch, L., Weering, J. R. T. van, Bahce, I., Pegtel, D. M., & Mouliere, F. (2022). Circulating extracellular vesicles in lung cancer patients are not enriched in tumor-derived DNA fragments as revealed by whole genome sequencing (p. 2022.07.22.501161). bioRxiv. https://doi.org/10.1101/2022.07.22.501161

Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., Fu, K.-Y., Kiss, E., Spalding, K. L., Landesberg, G., Zick, A., Grinshpun, A., Shapiro, A. M. J., Grompe, M., Wittenberg, A. D., … Dor, Y. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nature Communications, 9(1), Article 1. https://doi.org/10.1038/s41467-018-07466-6

Mouliere, F., Mair, R., Chandrananda, D., Marass, F., Smith, C. G., Su, J., Morris, J., Watts, C., Brindle, K. M., & Rosenfeld, N. (2018). Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. EMBO Molecular Medicine, 10(12), e9323. https://doi.org/10.15252/emmm.201809323

Mouliere, F., Smith, C. G., Heider, K., Su, J., van der Pol, Y., Thompson, M., Morris, J., Wan, J. C. M., Chandrananda, D., Hadfield, J., Grzelak, M., Hudecova, I., Couturier, D.-L., Cooper, W., Zhao, H., Gale, D., Eldridge, M., Watts, C., Brindle, K., … Mair, R. (2021). Fragmentation patterns and personalized sequencing of cell-free DNA in urine and plasma of glioma patients. EMBO Molecular Medicine, 13(8), e12881. https://doi.org/10.15252/emmm.202012881

Murtaza, M., Dawson, S.-J., Tsui, D. W. Y., Gale, D., Forshew, T., Piskorz, A. M., Parkinson, C., Chin, S.-F., Kingsbury, Z., Wong, A. S. C., Marass, F., Humphray, S., Hadfield, J., Bentley, D., Chin, T. M., Brenton, J. D., Caldas, C., & Rosenfeld, N. (2013). Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature, 497(7447), Article 7447. https://doi.org/10.1038/nature12065

Nakayama, Y., Yamaguchi, H., Einaga, N., & Esumi, M. (2016). Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions. PLoS ONE, 11(3), e0150528. https://doi.org/10.1371/journal.pone.0150528

Novák, P., Neumann, P., & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics, 11(1), 378. https://doi.org/10.1186/1471-2105-11-378

Oxnard, G. R., Paweletz, C. P., Kuang, Y., Mach, S. L., O'Connell, A., Messineo, M. M., Luke, J. J., Butaney, M., Kirschmeier, P., Jackman, D. M., & Jänne, P. A. (2014). Noninvasive detection of response and resistance in EGFR-mutant lung cancer using quantitative next-generation genotyping of cell-free plasma DNA. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 20(6), 1698–1705. https://doi.org/10.1158/1078-0432.CCR-13-2482

Pezer, Z., Brajković, J., Feliciello, I., & Ugarkovć, D. (2012). Satellite DNA-mediated effects on genome regulation. Genome Dynamics, 7, 153–169. https://doi.org/10.1159/000337116

Politi, K., & Herbst, R. S. (2015). Lung Cancer in the Era of Precision Medicine. Clinical Cancer Research : An Official Journal of the American Association for Cancer Research, 21(10), 2213–2220. https://doi.org/10.1158/1078-0432.CCR-14-2748

Postel-Vinay, S., Vanhecke, E., Olaussen, K. A., Lord, C. J., Ashworth, A., & Soria, J.-C. (2012). The potential of exploiting DNA-repair defects for optimizing lung cancer treatment. Nature Reviews Clinical Oncology, 9(3), Article 3. https://doi.org/10.1038/nrclinonc.2012.3

Roy, D., Zhang, Z., Lu, Z., Hsieh, C.-L., & Lieber, M. R. (2010). Competition between the RNA transcript and the nontemplate DNA strand during R-loop formation in vitro: A nick can serve as a strong R-loop initiation site. Molecular and Cellular Biology, 30(1), 146–159. https://doi.org/10.1128/MCB.00897-09

Sattler, U., Frit, P., Salles, B., & Calsou, P. (2003). Long-patch DNA repair synthesis during base excision repair in mammalian cells. EMBO Reports, 4(4), 363–367. https://doi.org/10.1038/sj.embor.embor796

Scieszka, D., Lin, Y.-H., Li, W., Choudhury, S., Yu, Y., & Freire, M. (2022). NETome: A model to Decode the Human Genome and Proteome of Neutrophil Extracellular Traps. Scientific Data, 9(1), 702. https://doi.org/10.1038/s41597-022-01798-1

Stavnezer, J., Guikema, J. E. J., & Schrader, C. E. (2008). Mechanism and regulation of class switch recombination. Annual Review of Immunology, 26, 261–292. https://doi.org/10.1146/annurev.immunol.26.021607.090248

Stortz, J. A., Hawkins, R. B., Holden, D. C., Raymond, S. L., Wang, Z., Brakenridge, S. C., Cuschieri, J., Moore, F. A., Maier, R. V., Moldawer, L. L., & Efron, P. A. (2019). Cell-free nuclear, but not mitochondrial, DNA concentrations correlate with the early host inflammatory response after severe trauma. Scientific Reports, 9(1), Article 1. https://doi.org/10.1038/s41598-019-50044-z

Sun, L., Du, M., Kohli, M., Huang, C.-C., Chen, X., Xu, M., Shen, H., Wang, S., & Wang, L. (2021). An Improved Detection of Circulating Tumor DNA in Extracellular Vesicles-Depleted Plasma. Frontiers in Oncology, 11, 691798. https://doi.org/10.3389/fonc.2021.691798

Sundquist, W. I., & Klug, A. (1989). Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. Nature, 342(6251), Article 6251. https://doi.org/10.1038/342825a0

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics, Chapter 4, 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25

Thierry, A. R. (2023). Circulating DNA fragmentomics and cancer screening. Cell Genomics, 3(1), 100242. https://doi.org/10.1016/j.xgen.2022.100242

Trost, B., Engchuan, W., Nguyen, C. M., Thiruvahindrapuram, B., Dolzhenko, E., Backstrom, I., Mirceta, M., Mojarad, B. A., Yin, Y., Dov, A., Chandrakumar, I., Prasolava, T., Shum, N., Hamdan, O., Pellecchia, G., Howe, J. L., Whitney, J., Klee, E. W., Baheti, S., … Yuen, R. K. C. (2020). Genome-wide detection of tandem DNA repeats expanded in autism. Nature, 586(7827), 80–86. https://doi.org/10.1038/s41586-020-2579-z

Udomruk, S., Phanphaisarn, A., Kanthawang, T., Sangphukieo, A., Sutthitthasakul, S., Tongjai, S., Teeyakasem, P., Thongkumkoon, P., Orrapin, S., Moonmuang, S., Klangjorhor, J., Pasena, A., Suksakit, P., Dissook, S., Puranachot, P., Settakorn, J., Pusadee, T., Pruksakorn, D., & Chaiyawat, P. (2023). Characterization of Cell-Free DNA Size Distribution in Osteosarcoma Patients. Clinical Cancer Research, 29(11), 2085–2094. https://doi.org/10.1158/1078-0432.CCR-22-2912

Ueki, S., Melo, R. C. N., Ghiran, I., Spencer, L. A., Dvorak, A. M., & Weller, P. F. (2013). Eosinophil extracellular DNA trap cell death mediates lytic release of free secretion-competent eosinophil granules in humans. Blood, 121(11), 2074–2083. https://doi.org/10.1182/blood-2012-05-432088

Ungerer, V., Bronkhorst, A. J., Uhlig, C., & Holdenrieder, S. (2022). Cell-Free DNA Fragmentation Patterns in a Cancer Cell Line. Diagnostics, 12(8), Article 8. https://doi.org/10.3390/diagnostics12081896

Wahba, L., Amon, J. D., Koshland, D., & Vuica-Ross, M. (2011). RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. Molecular Cell, 44(6), 978–988. https://doi.org/10.1016/j.molcel.2011.10.017

Wakasugi, M., & Sancar, A. (1998). Assembly, subunit composition, and footprint of human DNA repair excision nuclease. Proceedings of the National Academy of Sciences, 95(12), 6669–6674. https://doi.org/10.1073/pnas.95.12.6669

Wang, S., Meng, F., Li, M., Bao, H., Chen, X., Zhu, M., Liu, R., Xu, X., Yang, S., Wu, X., Shao, Y., Xu, L., & Yin, R. (2023). Multidimensional Cell-Free DNA Fragmentomic Assay for Detection of Early-Stage Lung Cancer. American Journal of Respiratory and Critical Care Medicine, 207(9), 1203–1213. https://doi.org/10.1164/rccm.202109-2019OC

Yang, Z. W., Yang, S. H., Chen, L., Qu, J., Zhu, J., & Tang, Z. (2001). Comparison of blood counts in venous, fingertip and arterial blood and their measurement variation. Clinical and Laboratory Haematology, 23(3), 155–159. https://doi.org/10.1046/j.1365-2257.2001.00388.x

Yu, K., Chedin, F., Hsieh, C.-L., Wilson, T. E., & Lieber, M. R. (2003). R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nature Immunology, 4(5), Article 5. https://doi.org/10.1038/ni919

Yuwono, N. L., Warton, K., & Ford, C. E. (2021). The influence of biological and lifestyle factors on circulating cell-free DNA in blood plasma. ELife, 10, e69679. https://doi.org/10.7554/eLife.69679

Zhang, Y., Yao, Y., Xu, Y., Li, L., Gong, Y., Zhang, K., Zhang, M., Guan, Y., Chang, L., Xia, X., Li, L., Jia, S., & Zeng, Q. (2021). Pan-cancer circulating tumor DNA detection in over 10,000 Chinese patients. Nature Communications, 12(1), 11. https://doi.org/10.1038/s41467-020-20162-8

6

METHODS

6.1 Clinical Samples

6.1.1 Non-cancer Plasma Clinical Samples

Plasma from healthy donors was commercially purchased from Innovative Research (IPLASK2E10ML) in K2EDTA tubes. According to vendor instructions, whole blood was spun at 5000xG for 15 minutes, and plasma was removed using a plasma extractor. The age and gender of the donors can be found in the charts (Table 6.1, Table 6.2, and Table 6.4). Purchased samples were anonymous and did not contain any additional personal details aside from age, sex, and race, and thus UCLA IRB approval was not applicable.

6.1.2 Source of NSCLC Plasma Samples

Plasma from late-stage NSCLC patients was obtained from UCLA in the NIH-funded project (4UH3CA206126-03: Advancing EFIRM-Liquid Biopsy (eLB) to a CLIA-Certified Laboratory Developed Test (eLB-LDT) for Detection of Actionable EGFR Mutations in NSCLC Patients, IRB#17-000997). Biopsy specimens were examined histologically, and the presence of EGFR mutations was determined using the Therascreen EGFR RGQ PCR Kit (EGFR IVD Kit)(Syed, 2016). The staging criteria used were those from the American Joint Committee on Cancer (AJCC) TNM system (Huang et al., 2015).

Table 6.1 Non-cancer patient demographics for Chapter 2

Purpose	Gender	Age
Digestions Donor 1	Male	47
Digestions Donor 2	Female	57
Digestions Donor 3	Male	35
Healthy 10 Replicate Donor 1	Male	45
Healthy 10 Replicate Donor 2	Male	18
Healthy 10 Replicate Donor 3	Male	23
Healthy 10 Replicate Donor 4	Male	26
Healthy 10 Replicate Donor 5	Male	38
Healthy 10 Replicate Donor 6	Male	33
Healthy 10 Replicate Donor 7	Male	22
Healthy 10 Replicate Donor 8	Male	37
Healthy 10 Replicate Donor 9	Male	27
Healthy 10 Replicate Donor 10	Male	41
Healthy Donor for QiaM on QiaC Flowthrough 1	Male	19
Healthy Donor for QiaM on QiaC Flowthrough 2	Male	25

Table 6.2. Non-cancer patient demographics for Chapter 3

Code	Sex	Age
NC01	F	60
NC02	M	41
NC03	M	47
NC04	F	26
NC05	F	32
NC06	F	35
NC07	F	46
NC08	F	48
NC09	F	39
NC10	M	38
NC11	M	18
NC12	M	52
NC13	M	37
NC14	M	41
NC15	F	30
NC16	F	30
NC17	F	31
NC18	F	46

Table 6.3 NSCLC patient demographics for Chapter 3

Code	Sex	Age	Stage	TNM Staging	Histological Type
LC01	M	56	IVA	cT4N3M1a(AJCC 8th)	Adenocarcinoma
LC02	F	69	IV	cT4N3M1B(AJCC 7th)	Adenocarcinoma
LC03	M	86	IVB	cT4N3M1c(AJCC 8th)	Adenocarcinoma
LS04	F	60	IVA	cT2aN3M1a(AJCC 8th)	Adenocarcinoma
LS05	F	64	IVA	cT1N3M1a(AJCC 8th)	Adenocarcinoma
LS06	F	46	IV	T1AN0M1A(AJCC 7th)	Adenocarcinoma
LS07	F	67	IV	T4N3M1B(AJCC7th)	Adenocarcinoma
LS08	F	61	IV	T4N3M1B(AJCC7th)	Adenocarcinoma
LS09	F	64	IV	T2aN3M1b(AJCC7th)	Adenocarcinoma
LS10	M	63	IV	T4N3M1b(AJCC7th)	Adenocarcinoma
LS11	F	69	IV	cT4N2M1a(AJCC7th)	Adenocarcinoma
LS12	M	58	IV	T4N3M1b(AJCC7th)	Adenocarcinoma
LS13	F	76	IVB	cT4N3M1c(AJCC 8th)	Adenocarcinoma
LS14	M	68	IVA	cT2N0M1a(AJCC 8th)	Adenocarcinoma

Table 6.4. Non-cancer patient demographics for Chapter 4

Number	Lot Number	Age	Sex
1	666	38	M
2	668	52	M
3	681	18	M
4	698	26	F
5	700	35	F

Table 6.5 NSCLC Patient Demographics for Chapter 4

Number	Lot Number	Age	Stage	Sex
1	120E	47	3A	F
2	147E	75	4	F
3	161E	79	3B	F
4	231E	62	3A	M

6.2 Nucleic Acid Extraction

6.2.1 QiaC and QiaM Extraction

Using the QIAmp Circulating Nucleic Acid Kit (Qiagen, 55114), we followed two of the manufacturer protocol: Purification of Circulating Nucleic Acids from 1mL of Plasma (QiaC) and Purification of Circulating microRNA from 1ml of Plasma (QiaM). Proteinase-K digestion was carried out as instructed. Carrier RNA was not used. The ATL Lysis buffer (Qiagen, 19076) was used as indicated in the microRNA protocol. The final elution volume was 20µl.

6.2.2 SPRI Extraction

100µL of Proteinase K (20mg/mL, Zymogen, D3001-2-1215) and 56µL 20% SDS (Invitrogen, AM9820) was added to 1mL of human plasma and incubated for 30 minutes at 60°C. After cooling to ambient room temperature, 540µL SPRI-select beads (Beckman Coulter, B22318) and 3000µL of 100% isopropanol (Fisher, BP26181) were added to the plasma and incubated for 10 minutes on the benchtop. The plasma was then centrifuged at 4000xG for five minutes. The supernatant was removed and discarded. The pellet was resuspended using 1mL of 1x TE

Buffer (Invitrogen, AM9848) and divided into 500µl aliquots into two phase lock tubes (Quantabio, 10847-802). An equal volume (500µL) of phenol:chloroform:isoamyl alcohol with equilibrium buffer was added (Sigma, P2069-100mL), and contents were vortexed for 15 seconds. The tubes were then centrifuged at 19000xG for five minutes. This was repeated twice (vortexed and centrifuged). The upper clear supernatant was pipetted and transferred to a 15mL conical tube SPRI-select beads, and 3000µL of 100% isopropanol were added to the plasma and incubated for 10 minutes on the benchtop. The tube was placed on a magnetic rack for five minutes to allow the beads to migrate.   The supernatant was discarded, and the beads were washed twice with 5ml of 85% ethanol. Once the second ethanol wash was removed, the beads were left to air dry for 10 minutes. The beads were then resuspended in 30µL of elution buffer (Qiagen, 19086) and incubated for 2 minutes. Afterwards, the beads were transferred to a 1.5mL tube and magnet rack to separate the beads from the resuspended DNA. Once the solution was clear (~2 minutes), the 30µL of elution was transferred to another 1.5mL tube and combined with 1µL of 20mg/ml glycogen (Thermo, R0561), 44µL of 1xTE Buffer, 25µL of 3M sodium acetate (Quality Biological INC, 50-751-7660), 250µL of 100% ethanol and placed at -80^{o}C overnight. The tube was then centrifuged at 19000xG for 15 minutes. The supernatant was removed and replaced with 200µL of 80% ethanol. This was done two more times. The supernatant was removed, and the pellet was resuspended in a 30µL of elution buffer and combined with 90µL of SPRI-select beads, 90µL of 100% isopropanol, and incubated for 10 minutes. The tube was placed on a magnetic rack for five minutes to allow the beads to migrate. The supernatant was discarded, and the beads were washed twice with 200µL of 80%

ethanol. Once the second ethanol wash was removed, the beads were left to air dry for 10 minutes. The beads were then resuspended in 40µL of Qiagen elution buffer.

6.2.3 QiaM Methylation Extraction (Chapter 4)

2 mL of plasma was extracted with QIAmp Circulating Nucleic Acid Kit (Qiagen, 55114) following the manufacturer protocol: Purification of Circulating microRNA from 2ml of Plasma (QiaM). Proteinase-K digestion was carried out as instructed. Carrier RNA was not used. The ATL Lysis buffer (Qiagen, 19076) was used as indicated in the microRNA protocol. The final elution for both protocols was 20µl.

6.3 Library Preparations Protocols

6.3.1 ssDNA Library Kit Preparation (BRcfDNA-Seq Library Preparation)

Single-stranded DNA library preparation was performed using the SRSLY^{TM} PicoPlus DNA NGS Library Preparation Base Kit with the SRSLY 12 UMI-UDI Primer Set, UMI Add-on Reagents, and purified with Clarefy Purification Beads (Claret Bioscience, CBS-K250B-24, CBS-UM-24, CBS-UR-24, CBS-BD-24). Since there is currently no optimized method to measure uscfDNA, 18µL of extracted cfDNA was used as input and heat-denatured as instructed. The low molecular weight retention protocol was followed for all bead cleanup steps to retain a high proportion of small fragments. The index reaction PCR was run for 11 cycles. In experiments including digested lambda DNA, a total of 50pg was added with the input cfDNA.

6.3.2 dsDNA Library Kit preparation

For double-stranded DNA libraries, the NEB Ultra II (New England Bio, E7645S) was used with a 9µL aliquot of extracted cfDNA according to the manufacturer's instructions with some modifications: the adapter ligation was performed using 2.5 µl of NEBNext® Multiplex Oligos for Illumina (Unique Dual Index UMI Adaptors RNA Set 1 - NEB, cat# E7416S); the post-adapter ligation purification was performed using 50 µl of purification beads and 50 µl of purification beads' buffer, while the second (or post-PCR) purification was performed using 60µl of purification beads (to retain smaller fragments). The PCR was performed using the MyTaq HS mix (Bioline, BIO-25045) for 10 PCR cycles.

6.3.3 BS-Seq Library Preparation

First, for the BS-Seq protocol, 20µl of extracted DNA underwent bisulfite conversion using Zymo Research DNA Methylation Lightning kit (Zymo Research, cat# D5030) with an elution volume of 20uL. Subsequently, single-stranded libraries were constructed as described in "BRcfDNA-Seq Library Preparation ." During the final index PCR, the Index PCR Master Mix was substituted with the Kapa HIFI HotStart Uracil+ ReadyMix. The Bisulfite PCR protocol is as follows: 98oC for 3 minutes, [98oC for 30 seconds, 60oC for 30 seconds, 72oC for 1:00] for 11 cycles, 72oC for 1 minute, then hold at 12oC. All bead cleanup steps followed the low molecular weight retention purification protocol.

6.3.4 5mCAdpBS-Seq Library Preparation

The first step of the single-stranded library preparation (pre-methylated single-stranded adapter ligation) was performed on extracted cfDNA prior to bisulfite conversion. Custom 5mC protected SRSLY adapters were provided by Claret Bioscience and used in place of the regular adapters in the adapter ligation step. After the bead cleanup in the first step, the product was resuspended to 20µL. Then 20µl of adapter-ligated DNA underwent bisulfite conversion using Zymo Research DNA Methylation Lightning kit (Zymogen, cat# D5030) with an elution volume of 15µL into the UMI-UDI step of the single-strand library preparation protocol. The remaining steps (Addition of UMI by Primer extension and Index PCR) were performed as described in "BRcfDNA-Seq Library Preparation ." During the final index PCR, the Index PCR Master Mix was substituted with the Kapa HIFI HotStart Uracil+ ReadyMix. The Bisulfite PCR protocol is as follows: 98oC for 3 minutes, [98oC for 30 seconds, 60oC for 30 seconds, 72oC for 1:00] for 11 cycles, 72oC for 1 minute, then hold at 12oC. All bead cleanup steps followed the low molecular weight retention purification protocol.

6.4 Sequencing

Final library concentrations were measured using the Qubit Fluorometer (Thermo, Q33327), and quality was assessed using the Tapestation 4200 using D1000 High-Sensitivity Tapes (Agilent, G2991BA and 5067-5584). Final libraries were run on Nova-Seq SP 300 (150x2) to reach 40 million reads per sample.

Table 6.6. Synthetic oligomers and primers

Name	Size	ss/ds	Lambda phage region	Notes
Lambda dsDNA Control	459 bp	ds	27’944:28’402	PCR product, no UMI
5′-CAAACTGCGCAACTCGTGAAAGGTAGGCGGATCCCCTTCGAAGGAAAGACCTGATGCTTTTCGTGCGCGCATAAAATACCTTGATACTGTGCCGGATGAAAGCGGTTCGCGACGAGTAGATGCAATTATGGTTTCTCCGCCAAGAATCTCTTTGCATTTATCAAGTGTTTCCTTCATTGATATTCCGAGAGCATCAATATGCAATGCTGTTGGGATGGCAATTTTTACGCCTGTTTTGCTTTGCTCGACATAAAGATATCCATCTACGATATCAGACCACTTCATTTCGCATAAATCACCAACTCGTTGCCCGGTAACAACAGCCAGTTCCATTGCAAGTCTGAGCCAACATGGTGATGATTCTGCTGCTTGATAAATTTTCAGGTATTCGTCAGCCGTAAGTCTTGATCTCCTTACCTCTGATTTTGCTGCGCGAGTGGCAGCGACATGGTTTGTTGT-3'				
Lambda ssDNA Control	350 nt	ss	7’582:7’930	IDT synthesized

5′- CCTGGCCAGAATGCAATAACGGGAGGCGCTGTGGCTGATTTCGATAACCTGTTCGATGCTGCCATTGCCCGCGCCGATGAAACGATACGCGGGTACATGGGAACGTCAGCCACCATTACATCCGGTGAGCAGTCAGGTGCGGTGATACGTGGTGTTTTTGATGACCCTGAAAATATCAGCTATGCCGGACAGGGCGTGCGCGTTGAAGGCTCCAGCCCGTCCCTGTTTGTCCGGACTGATGAGGTGCGGCAGCTGCGGCGTGGAGACACGCTGACCATCGGTGAGGAAAATTTCTGGGTAGATCGGGTTTCGCCGGATGATGGCGGAAGTTGTCATCTCTGGCTTGGAC-3'		
I7 Extension Primer Sequence (i7 ext)	75 nt	5′- CAAGCAGAAGACGGCATACGAGATNNNNNNNNNXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'
Forward Index Primer Sequence (i5)	70 nt	5′- AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
Reverse Index Primer Sequence (Ui7)	21 nt	5′- CAAGCAGAAGACGGCATACGA-3′

6.5 Nuclease Digestions

6.5.1 Nuclease Digestions for Analysis of Strandedness

Prior to library preparation, the extracted cfDNA was digested with various strand-specific nucleases. For all reactions, 500pg of control oligos (350nt ssDNA and 460bp dsDNA lambda sequence, IDT) was spiked into 20µL of cfDNA. DNA was purified by combining 30µL of reaction buffer with 90µL of SPRI-select beads and 90µL of 100% isopropanol and incubated for 10 minutes. The tube was placed on a magnetic rack for five minutes to allow for the beads to migrate. The supernatant was discarded, and the beads were washed twice with 200µL of 80% ethanol. Once the second ethanol wash was removed, the beads were left to air dry for 10 minutes. The beads were then resuspended in 20µL of Qiagen elution buffer.

Non-strand specific DNA digestion: 20µL cfDNA was combined with 1µL DNase I (Invitrogen, 18-068-015), 3µL 10xDNase 1 Buffer, 6µL of ddH2O incubated for 15 minutes at 37°C and heat-inactivated for 15 minutes at 80°C with 1µL of 0.5M EDTA.

ssDNA-specific Digestion: 20µL cfDNA was combined with 1µL 1x S1 (Thermo, EN0321), 6µL 5x S1 Buffer, 3µL of ddH2O incubated for 30 minutes at room temperature and heat-inactivated for 15 minutes at 80°C with 2µL of 0.5M EDTA (.

ssDNA-specific Digestion: 20µL cfDNA was combined with 1µL 0.1x P1 (NEB, M0660S), 3µL NEBuffer r1.1, 6µL of ddH2O incubated for 30 minutes at 37°C and inactivated with 2µL of 0.5M EDTA.

ssDNA-specific Digestion: 20µL cfDNA was combined with 3µL Exonuclease 1 (NEB, M0293S), 3µL 10x Exo 1 Buffer, 4µL of ddH2O incubated for 30 minutes at 37°C and heat inactivated for 15 minutes at 80°C with 1µL of 0.5M EDTA.

dsDNA-specific Digestion: 20µL cfDNA was combined with 2µL dsDNase (ArcticZyme, 70600-201), 8µL of ddH2O incubated for 30 minutes at 37°C and heat inactivated for 15 minutes at 65°C with 1mM DTT.

Nick Repair Analysis: 20µL cfDNA was combined with 1µL PrePCR Repair (NEB, M0309S), 5µL ThermoPol Buffer (10x), 0.5µL of NAD+ (100x), 2µL of Takara 2.5mM dNTP, 21.5 ddH2O incubated for 30 minutes at 37°C and placed on ice.

RNA Digestion: 20µL of cfDNA was combined with 1µL of RNase Cocktail (Thermo, AM228). For 20 minutes at 30°C prior to input into the library preparation.

6.5.2 Lambda DNA Control Restriction Enzyme Reactions

1.5µl (1µg) of unmethylated lambda (Promega, D1521) was used for all reactions. After the restriction enzyme reaction, the DNA was purified by combining 20µL of reaction mixture and combined with 60µL of SPRI-select beads and 60µL of 100% isopropanol and incubated for 10 minutes. The tube was placed on a magnetic rack for five minutes to allow for the beads to migrate. The supernatant was discarded, and the beads were washed twice with 200µL of 80% ethanol. Once the second ethanol wash was removed, the beads were left to air dry for 10 minutes. The beads were then resuspended in 20µL of Qiagen elution buffer.

CviKL Restriction Enzyme: 1.5µL of Lambda DNA was combined with 2µl (10x) rCutSmart Buffer, 1ul CviKL enzyme (NEB, R0710S) and 15.5µl H_2O. The mixture was heated to 25°C for 60 minutes, and then the enzyme deactivated by heating it to 65°C for 20 minutes.

NlaIII Restriction Enzyme: 1.5µL of Lambda DNA was combined with 2µl (10x) rCutSmart Buffer, 1µL NlaIII (NEB, R0125S), and 15.5µl H_2O. The mixture was heated to 37°C for 15 minutes, and then the enzyme deactivated by heating it to 65°C for 20 minutes.

AluI Restriction Enzyme: 1.5µL of Lambda DNA was combined with 2ul (10x) rCutSmart Buffer, AluI (NEB, R0137S), and 15.5µl H_2O. The mixture was heated to 37°C for 60 minutes, and then the enzyme deactivated by heating it to 65°C for 20 minutes.

Digested lambda products were combined to produce a mixture with a final concentration of 50pg/µL.

6.6 Bioinformatic Pre-Processing

Initial experiments for merging paired reads into single-end reads were performed using BBMerge (Bushnell et al. 2017). Then single-end .fastq files were trimmed with (fastp (Chen et al., 2018), using adapter sequence AGATCGGAAGAGCACACGTCTGAACTCCAGTCA (r1) and AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT (r2) and filtered for a Phred score of >15. Non-BS-treated and BS-treated libraries aligned against the human reference genome [GenBank:GCA_000001305.2] using BWA-mem(Li and Durbin 2009) and BSBolt's default settings(Farrell et al. 2021), respectively. The LambdaPhage sequence was appended to the end of the human reference. Sequence reads were demultiplexed using SRSLYumi (SRSLYumi 0.4 version, Claret Bioscience), python package. The duplicated reads were removed using Picard Toolkit (http://broadinstitute.github.io/picard/) after sorting, filtering, and removing soft and hard clipped reads with Samtools (samtools 1.9 version). Quality control was performed with Qualimap Version 2.2.2c. To isolate mitochondria DNA reads, samtools were used to filter out those reads mapped to that genome. The bam file aligning to human, mitochondria, and lambdaPhage was first binned into reads 10bp of from 20 to 200bp using alignmentsieve (deepTools 3.5.0) (Ramírez et al. 2016).

6.7 BRcfDNA-Seq Bioinformatic Analysis

6.7.1 Genome-wide Ideogram

The .bam files were split into genomic bins of 1 million reads along the genome (e.g., Chr1:0-1,000,000) for two in-silico categories: uscfDNA (40-70bp) and mncfDNA (120-250bp).

Karyograms are self-normalized so that the legend reflects the intrasample dynamic range. Ideograms were constructed from .bam files that were 1 million bp using rideogram R package (Hao et al., 2020).

6.7.2 Functional Element Analysis

Functional peaks were detected using macs2 (2.2.7.1 version) (Zhang et al., 2008) and then analyzed with HOMERannotatePeaks (version 4.11.1) to determine which functional element category each peak is associated with. Only 3'UTR, TTS (Transcription termination site), Exon, Intron, Intergenic, Promoter, and 5'UTR categories were used based on the UCSC HG38 annotations database (Rosenbloom et al., 2015). Protein-coding and ncRNA gene types were used. For each category, the top 10 peaks were used to generate a list of the top 20 most common peaks between non-cancer and NSCLC. The chord diagram indicating the common peaks for both cohorts' promoter, introns, or exonic regions was assembled using Flourish (https://flourish.studio). Individual peaks were defined as the % contribution (peak score/ summed peak score of the select 20 per category). For example, if the peak score for Snx16 was 433, it was divided by the total peak score of the top 20 (2400) to arrive at a score of 0.18.

6.7.3 Fragment Curve Profiles

Non-normalized fragment curve profiles were calculated using Samtools (Li et al., 2009) by plotting a histogram of the % reads of each length in the 20-350bp bin.

6.7.4 Fragmentomics

The .bam files were split into genomic bins of 1 million bp along the genome (e.g., Chr1:0-1,000,000) for two in-silico categories: uscfDNA (40-70bp) and mncfDNA (120-250bp). For each genomic bin, we calculated the fragment scores by totaling the read count of those from

40-53bp (A) and 54-70bp (B) for uscfDNA and 120-167bp (A) and 168-250bp (B) for mncfDNA and by using the following equation$\frac{(\frac{A}{A+B})}{(\frac{B}{A+B})}$. The scores bin was plotted sequentially to form the genome fragment score curves.

6.7.5 End-motif Score

The first four base pairs from the 5' end were extracted and compiled using a custom python script. The End-Motif Diversity Score (Shannon Entropy) was calculated by analyzing the distribution of frequencies of motifs (total of 256 motifs) and compared between different sample populations. As per (Jiang et al., 2020), the normalized Shannon entropy mathematical equation was used, incorporating the contribution of all 256 motifs, with P_i being the frequency of a particular motif (e.g., CCCA).

$$Motif\ Score = \Sigma_{i=1}^{256} - P_i * \log{(P_i)}/\log{(256)}$$

6.7.6 G-Quadruplex (G-Quad) Percentage

The G-Quad percentage was calculated by first converting binned .bam files to .bed and then to .fasta using bamtobed (bedtools ver 2.18) and getfasta (bedtools ver 2.18) (Quinlan & Hall, 2010). G-Quad signatures were detected using fastaRegexFinder.py to analyze the sequences in the reads (https://github.com/dariober/bioinformatics-cafe/tree/master/fastaRegexFinder). This python pipeline examines if the sequences contain this pattern in this equation "([gG]{3,}\w{1,7}){3,}[gG]{3,}". This translates to the identification of 3 or more G nucleotides followed by 1 to 7 of any other bases and must be repeated three or more times and end with three or more Gs. The G-Quad counts were divided by the total read counts to identify the G-

Quad percentage and normalized by the average bp of the fragments of each bin (uscfDNA: 50bp | mncfDNA: 167bp).

6.8 Methylation Bioinformatic Analysis

6.8.1 CpG and non-CpG% Methylation

This was determined using BSbolt's methylation call on each bin size to determine the %CG methylation and %CHH methylation(Farrell et al., 2021). MapQ scores were calculated from each size bin and determined using the reports from Qualimap Version 2.2.2c(García-Alcalde et al., 2012).

6.8.2 CpG Regions

CpG regions were calculated using BSbolt methylation call on both NT-seq and BS-treated Libraries on the 10bp increment bin sizes (from 20-200) and looking at total CpG positions reported. G-Quad percentage was calculated by first converting binned .bam files to .bed using bamtobed (bedtools ver 2.18) and then from .bed to .fasta using getfasta (bedtoosl ver 2.18)(Quinlan and Hall 2010).

6.8.3 G-Quad Signatures Count

Similar to Chapter 4, G-Quad percentage was calculated by first converting binned .bam files to .bed using bamtobed (bedtools ver 2.18) then from .bed to .fasta using getfasta (bedtools ver 2.18)(Quinlan and Hall 2010). fastaRegexFinder.py was used to analyze the sequences in the reads (https://github.com/dariober/bioinformatics-cafe/tree/master/fastaRegexFinder). In general, this python pipeline examines if the sequences contain this pattern in this equation "([gG]{3,}\w{1,7}){3,}[gG]{3,}". This translates to identify 3 or more G nucleotides followed by 1 to 7 of any other bases and must be repeated 3 or more times and end with 3 or more G. The G-Quad counts was divided by the total read counts to identify the G-Quad percentage. A normalized ratio was calculated by dividing the read counts by the median value of the bin length (eg. 20-29 is 25). Only primary fragments that contained G-Quad sequences were counted (eg. complementary sequences that contained G-Quads were excluded). The coordinates of the G-Quad sequences were used to generate G-Quad Only bam files for CpG methylation% calling.

6.8.4 CpG Methylation % Quantification Trend Plots

We generated quantification plots for eleven genomic elements (SINEs, LINEs, Simple Repeats, Exons, Introns, Intergenic, Promoters, CpG Islands, 5' Untranslated Region, 3' Untranslated Region, and Translation Termination Site (TTS)) were performed in SeqMonk (version 1.48.1) https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/. The .CGmap.gz files generated by BSbolt were converted to cg.bismark.cov.gz files using zless function to rearrange the column format. The files were imported, and probes were defined using Feature Probe Generator for the categories. Probes were defined as over feature from -5000bp to +5000bp. Probes were then defined using Running Window Generator. Window | step size boundaries were set as 100 and 100. After, define quantification was selected with "features to quantitate" as existing probes, minimum count to include position as 1, minimum observations to include feature set as 1, and combined value to report as mean. Next, quantification trend plots were constructed CDS, CG Island, or TSS promoter. Remove exact duplicates were checked and "make probes" from -5000bp to +5000bp for both over features and centered on feature. Graphs were plotted using Graphpad Prism 9.

6.8.5 Differentially Methylated Regions

Samples were aggregated using metilene_input.pl from metilene package (Jühling et al., 2016) using a minimum coverage of 1. DMRs between samples were identified using metilene (Jühling et al., 2016) using the settings --mincpgs 3 –maxdist 100 –minMethDiff 0.1 –valley 0.7. Closest gene was analyzed using bedtools closest with default settings with a hg38 gene reference from UCSC RefSeq (refgene) from https://genome.ucsc.edu/cgi-bin/hgTables.

6.8.6 Tissue Deconvolution

Samples were analyzed using CelFiE (CEL Free DNA decomposition Expectation maximization) algorithm with default parameters as described(Caggiano et al. 2021).

6.8.7 Epigenetic Marks Overlap Region Ratios

Ratios were calculated using bedtools (version 2.30.0) intersect(Quinlan & Hall, 2010) with the -wo argument. Intersected bp counts were divided by total bp counts of the bed file. Control bed files were generated using bedtools shuffle with human reference described above and by sourcing size fragments count and distribution from the uscfDNA or mncfDNA bed file from each respective subject. Experiment reference files were retrieved from the BLUEPRINT Data Analysis Portal(Fernández et al., 2016). Specific subjects used were the following: eosinophil (S006XE53 and S006XEH2), macrophage (C005VG51 and C005VGH1), monocyte (C000S5A1b and C000S5H2), and neutrophil (C0010KA1bs and C0010KH2). The % intersected base pairs were normalized against control shuffled bed files to compare non-cancer and NSCLC samples.

6.9 Statistical Analysis

For fragmentomics, functional elements, and end-motif, we calculate significant regions of interest by performing paired (between uscfDNA and mncfDNA of non-cancer samples) and non-paired multiple t-tests with a false discovery rate of 1% using the two-stage step-up method of Benjamini, Krieger, and Yekutieli (Zehetmayer & Posch, 2012). For comparison less than 20, non-paired multiple paired t-tests were performed with Holm-Šídák correction with alpha at 0.05(Guo & Romano, 2007). A Student's t-test was performed with Welch's correction (after ANOVA if necessary) for single comparisons. Using significant targets from the domains, we performed multivariable analysis using the online principal component analysis tool (https://biit.cs.ut.ee/clustvis/) (Metsalu & Vilo, 2015). Error bars represent SEM. Stars indicate adjusted q-values are presented with * p <0.05, ** p <0.01, *** p< 0.001, and **** p <0.0001.

6.10 References

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England), 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Farrell, C., Thompson, M., Tosevska, A., Oyetunde, A., & Pellegrini, M. (2021). BiSulfite Bolt: A bisulfite sequencing analysis platform. GigaScience, 10(5), giab033. https://doi.org/10.1093/gigascience/giab033

Fernández, J. M., de la Torre, V., Richardson, D., Royo, R., Puiggròs, M., Moncunill, V., Fragkogianni, S., Clarke, L., Flicek, P., Rico, D., Torrents, D., de Santa Pau, E. C., & Valencia, A. (2016). The BLUEPRINT Data Analysis Portal. Cell Systems, 3(5), 491-495.e5. https://doi.org/10.1016/j.cels.2016.10.021

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. Bioinformatics (Oxford, England), 28(20), 2678–2679. https://doi.org/10.1093/bioinformatics/bts503

Guo, W., & Romano, J. (2007). A generalized Sidak-Holm procedure and control of generalized error rates under independence. Statistical Applications in Genetics and Molecular Biology, 6, Article3. https://doi.org/10.2202/1544-6115.1247

Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., & Chen, J. (2020). RIdeogram: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Computer Science, 6, e251. https://doi.org/10.7717/peerj-cs.251

Huang, S. H., Xu, W., Waldron, J., Siu, L., Shen, X., Tong, L., Ringash, J., Bayley, A., Kim, J., Hope, A., Cho, J., Giuliani, M., Hansen, A., Irish, J., Gilbert, R., Gullane, P., Perez-Ordonez, B., Weinreb, I., Liu, F.-F., & O'Sullivan, B. (2015). Refining American Joint Committee on Cancer/Union for International Cancer Control TNM stage and prognostic groups for human papillomavirus-related oropharyngeal carcinomas. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 33(8), 836–845. https://doi.org/10.1200/JCO.2014.58.6412

Jiang, P., Sun, K., Peng, W., Cheng, S. H., Ni, M., Yeung, P. C., Heung, M. M. S., Xie, T., Shang, H., Zhou, Z., Chan, R. W. Y., Wong, J., Wong, V. W. S., Poon, L. C., Leung, T. Y., Lam, W. K. J., Chan, J. Y. K., Chan, H. L. Y., Chan, K. C. A., … Lo, Y. M. D. (2020). Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discovery, 10(5), 664–673. https://doi.org/10.1158/2159-8290.CD-19-0622

Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016). metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Research, 26(2), 256–262. https://doi.org/10.1101/gr.196394.115

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Metsalu, T., & Vilo, J. (2015). ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Research, 43(W1), W566–W570. https://doi.org/10.1093/nar/gkv468

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., … Kent, W. J. (2015). The UCSC Genome Browser database: 2015 update. Nucleic Acids Research, 43(Database issue), D670-681. https://doi.org/10.1093/nar/gku1177

Syed, Y. Y. (2016). therascreen® EGFR RGQ PCR Kit: A Companion Diagnostic for Afatinib and Gefitinib in Non-Small Cell Lung Cancer. Molecular Diagnosis & Therapy, 20(2), 191–198. https://doi.org/10.1007/s40291-016-0189-0

Zehetmayer, S., & Posch, M. (2012). False discovery rate control in two-stage designs. BMC Bioinformatics, 13, 81. https://doi.org/10.1186/1471-2105-13-81

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology, 9(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137