Nucleosomics analysis of cell-free DNA for patient stratification

Mariya Shtumpf

A thesis submitted for the degree of MSD in Molecular Medicine

School of Life Sciences

University of Essex

December 2022

Supervisor: Dr Vladimir B. Teif

Acknowledgements

I thank my lab members for being like a family to me throughout my university days, especially my supervisor Dr Vladimir Teif for his guidance and support. I am very thankful to Divya and Clark for their help and advice as fellow master students and to Dr Hulkar Mamayusupova for her objective and thorough criticism which I always found extremely valuable. I also thank Luminita, Shivam and Isabella for their support throughout the course of this project.

I thank Prof Anish Thomas and his lab for providing the data from the small cell lung cancer study, and Prof Leonard Schalkwyk and Yucheng Wang for providing the DNA methylation in ageing dataset, and for all the informative and fascinating discussions we had.

I am very grateful to Dr Greg Brooke for all his patience, advice and helpful suggestions.

I am also very thankful to my examiners Dr Toni Marco and Dr Mikhail Spivakov for a fascinating discussion and their very helpful suggestions for this dissertation and words of support.

I thank Prof Nelson Fernández for his kind support and valuable discussions, especially about the immunology direction of my project.

I would also thank my late friend Angela for kindly looking after me throughout my academic journey from my school days to almost the end of my masters.

I thank my dear Toby and his family for their infinitely kind support and encouragement.

And finally yet importantly, I thank my parents and my sister Katie for all their great help and support.

Contents

List of abbreviations

bp: base pair

cfDNA: cell-free DNA

ChIP-seq: Chromatin Immunoprecipitation followed by sequencing

CLL: Chronic Lymphocytic Leukaemia

IGHV: Immunoglobulin Heavy Chain Region

M-CLL: IGHV-mutated CLL

U-CLL: IGHV-unmutated CLL

ctDNA: circulating tumour DNA

HCC: Hepatocellular Carcinoma

MNase-seq: Micrococcal Nuclease digestion followed by sequencing

NETs: Neutrophil Extracellular Traps

NGS: Next Generation Sequencing

NRL: Nucleosome Repeat Length

PCA: Principal Component Analysis

SCLC: Small Cell Lung Cancer

SRA: Sequence Read Archive

TF: Transcription Factor

TSS: Transcription Start Site

WGS: Whole Genome Sequencing

y.o.: years old

Abstract

Nucleosome positioning is a key process in gene regulation and may vary throughout cell differentiation stages and in response to healthy and disease-related processes. In this project I examined different methods of using nucleosome positioning reconstructed from cell-free DNA (cfDNA) as part of a method for patient diagnostics and stratification. These results contributed to the development of two novel methods: one involving stratification of patients based on condition-specific changes in nucleosome occupancy, and another one based on distances between nucleosomes. Using these two methods, I analysed nucleosome positioning reconstructed from cfDNA in healthy people and patients with breast cancer and small cell lung cancer (SCLC), as well as in paired normal/tumour breast tissues. In addition, I have analysed the effect of ageing on the nucleosome patterns reconstructed from cfDNA. Both methods successfully distinguish cancer from healthy, as well as different age groups. CfDNA diagnostic methods based on nucleosome positioning have a potential to bring significant improvements in the clinic.

1. Literature review

Medical tests based on cell-free DNA (cfDNA) are referred to as “liquid biopsy” since they allow to avoid classic tissue biopsy when it comes to sampling DNA from solid tumours (Volik *et al.*, 2016). The presence of cfDNA in blood plasma was first reported in 1948 (Mandel and Metais, 1948). Later it was established that cfDNA arises as a result of digestion of nuclear DNA (Williamson, 1970). The active application of cfDNA in medical diagnostics started only in the last decade with the advancements in next generation sequencing (NGS) (Ignatiadis *et al*., 2021). Its current applications range from cancer diagnostic assays (Frenel *et al*., 2015) and prenatal testing (Kitzman *et al*., 2012) to the study of ageing (Teo *et al*., 2019) and even the assessment of astronaut well-being during spaceflights (Bezdan *et al*., 2020). However, while the sequencing technologies and our knowledge of the genomics of disease is constantly improving, the field of medical diagnostics is increasingly more in need of methods that would fare better in terms of sensitivity and cost.

1.1. cfDNA nucleosomics in medical diagnostics. Genomic nucleosome positions are specific for a given cell type and change over time as the cell adjusts to new modes of functioning. It reflects the workings of gene regulatory machinery on chromatin (reviewed in Clarkson *et al*., 2019) and has important implications in diagnosis and management of cancer. Sequenced data is usually acquired through the analysis of DNA from tissue samples where the acquisition of tissue requires biopsy through surgery in the case of solid tumours, which is invasive and carries risk to patient health. CfDNA, which consists of double-stranded fragments of DNA digested between nucleosomes and released by dying cells (through apoptosis and necrosis) (Ungerer *et al*., 2021) or through active release (as neutrophil extracellular traps (NETs) through the inflammatory process known as NETosis), can provide nucleosome positioning information. This makes cfDNA a promising non-invasive diagnostic marker for assessing disease and physiological states.

Nucleosome positioning changes accumulate much faster than mutations and are more sensitive to environmental changes (e.g. treatment or disease progression) and therefore can be more promising than the existing diagnostic tests which mostly target mutations. The half-life of cfDNA in blood plasma is short enough to be measured in minutes, so it delivers an almost real-time representation of nucleosome positioning in tissues of origin. Thus, nucleosome positioning has promising diagnostic value as it would offer fewer limitations compared to other tests.

1.1.1. cfDNA nucleosomics in cancer. Protein biomarker- and nucleic acid-based assays are commonly used in diagnosis and monitoring of cancer patients (Huang *et al*., 2020). CfDNA-based methods are attracting increasingly more attention because of their non-invasive nature and, potentially, high tumour specificity by using quantitative detection or targeted sequencing.

Elevated levels of blood cfDNA have been associated with cancer as a consequence of increased necrotic or apoptotic activity since tumour cells tend to divide more frequently than healthy ones, resulting in higher levels of cfDNA (Raja *et al*., 2018) and producing what is known as circulating tumour DNA (ctDNA) (Shu *et al.*, 2017). Therefore, the quantification and profiling of ctDNA offer a great prognostic value, which results in them being seen as very promising novel minimally invasive biomarkers (Diehl *et al*., 2008).

1.1.2. cfDNA nucleosomics in ageing. As human life expectancy grows, the elderly population is rapidly expanding, with the proportion of people aged 80 or over being expected to increase sevenfold in the next 80 years (United Nations, 2017). Since increased lifespan does not guarantee increased health-span, geriatric health is a matter of priority. The biomarkers of ageing with clinical relevance are attracting increasingly more interest. The availability of such biomarkers and knowledge of processes underlying ageing can be of use to medicine in general, and to the development of the "holy grail" of liquid biopsy – a method to diagnose any disease or even to predict the development of a condition before the symptoms worsen and it becomes too advanced for cure.

Biomarkers circulating in blood plasma are becoming increasingly relevant to molecular gerontology, especially in defining biological ageing as opposed to chronological ageing (Capri *et al*., 2015). Biomarkers that estimate mortality in ageing are already known, for example C-reactive protein and insulin-like growth factor-1 (Castagne *et al*., 2018). Recent studies show that individuals of the same chronological age can differ in biological age, which can be predictive of mortality and poor health (Chen *et al.*, 2016). Amongst the various proposed biomarkers of biological age, such as N-glycans and DNA methylation (Horvath, 2013; Miura *et al*., 2016), cfDNA is especially favoured due to the minimally invasive nature of sample collection and the increasing affordability of DNA sequencing. In spite of this, the relationship between cfDNA and ageing is little known.

Since 1970s, there were many attempts to determine genome-scale changes of nucleosome positioning in ageing, e.g. quantified by the distances between nucleosomes or the distribution of sizes of DNA fragments protected by nucleosomes from nuclease digestion. However, the question of whether these variables undergo systematic or predictable changes with ageing is still open. This question becomes critical for new generations of patient diagnostics based on the analysis of cfDNA, where cfDNA fragments originate from nucleosome-protected genomic regions undergoing digestion by apoptotic nucleases and other enzymes. Classical studies of NRL changes with ageing, performed in pre-NGS era using chromatin MNase digestion, have been inconclusive. Indeed, early investigations in mice and rats suggested that NRL in cerebral cortex neurons decreases by ~10 base pairs (bp) during ageing, while NRL in cerebellum neurons and liver cells increases (Zongza *et al*., 1979). This was contrasted by works claiming large increase of NRL in neurons of ageing rats (Berkowitz *et al*., 1983), while other authors reported that NRL does not change in the mouse ageing model, at least not in the liver, brain and heart tissues (Gaubatz *et al*., 1979). It is worth noting that cortical neurons mentioned above represent an outlier from most other cell types in terms of their unusually small NRL of 165 bp (Clark *et al*., 2020), and therefore the brain may have very non-homogenous nucleosome repositioning trends. As long as cfDNA-based diagnostics is considered, most cfDNA present in blood plasma comes from apoptotic lymphocytes rather than other cell types. Pre-NGS era studies using lymphocytes from people of different age did not detect NRL changes within the age range between 24 and 78 years (Smith *et al*., 1989). Also, it was not possible to detect NRL differences in human skin fibroblasts from people with ages ranging from 16 to 60 years old (y.o.) (Ishimi *et al*., 1987).

NGS-era studies clearly demonstrated the existence of ageing-associated chromatin defects which precede DNA damage accumulation in human fibroblasts (Pegoraro *et al*., 2009). However, there is still no simple answer to the question of systematic changes of nucleosome positioning with age. In all available models of ageing, senescence has been associated with histone loss along with general chromatin redistribution, loss of balance in activation and repression by histone modifications and changes in transcription, along with changes in DNA methylation and global nuclear re-organisation (Sen *et al*., 2016). It has been shown that ageing-associated histone loss is associated with about 50% decrease of overall nucleosome occupancy in ageing yeast (Hu *et al*., 2014). However, studies performed in mice concluded that mammalian core histone levels are not significantly affected in ageing (Chen *et al*., 2020). Furthermore, it has been reported that neutrophils of older mice (20-months old compared with 4-months old) have even higher nucleosome occupancy and smaller NRL near bound transcription factors (TFs) as determined by ATAC-seq (Lu *et al*., 2021). Thus, deciphering the NRL change in mammalian ageing still remains a challenge. Here we will resolve this challenge using recently published experimental datasets of sequenced cfDNA from blood plasma in people of different ages: 25-, 70- and 100-year-old individuals from a study by Teo *et al*. (2019) and healthy people of ages 24-50 from a study by Peneder *et al*. (2021), a cancer study without the context of ageing.

The pioneering study by Teo *et al*. aimed to assess global chromatin changes in response to ageing by analysing cfDNA from the plasma of healthy volunteers and patients from three distinct age groups. Nucleosome positioning data inferred from sequenced plasma cfDNA showed a rearrangement of chromatin that had been previously associated with biological ageing in animal models. In addition, a relative loss of cfDNA signal was detected throughout the ageing process at transcription start sites (TSS), termination sites, dimeric AluY elements and 5′UTR of L1HS retrotransposons. This is the first study to use a non-invasive method to study global *in vivo* epigenome changes in ageing and here we explore this dataset further with implications in cancer management.

1.1.3. CfDNA nucleosomics of medical conditions involving inflammation. Inflammation is associated with a wide range of medical conditions and can be detected from characteristic cfDNA patterns (Franceschi *et al*., 2014). The ageing process often includes chronic systemic inflammation (sometimes referred to as "inflammaging" or "inflammageing" in UK spelling). Higher concentrations of cfDNA have been reported in the plasma of elderly people (Jylhava *et al*., 2011). This may be caused by the increase in ageing cells that also release senescence-associated proteins that induce inflammation (Franceschi *et al*., 2014). Cellular debris, metabolites, nucleic acids of intrinsic and extraneous origin (e.g. nuclear acids, mitochondrial DNA and microRNAs) can act as damage-associated molecular patterns that trigger an inflammatory response as they accumulate. Their accumulation is a part of healthy physiology and increases throughout the ageing process while their destruction through autophagy decreases with age, thus increasing general inflammation (Franceschi *et al*., 2017). Higher abundance of cfDNA in the blood of younger people is also seen in the context of systemic inflammation and poor health (Jylhava *et al*., 2013).

1.2. Computational methods of cfDNA analysis. The first cfDNA-based sequencing assays were based on the analysis of mutations (Frenel *et al*., 2015). An important limitation in these types of analysis is that their sensitivity is limited by sequencing depth, which can increase the cost of diagnostic assays. Also, they perform best in high concentrations of cfDNA, which cannot always be achieved in the clinic and seem to increase with progression of disease, therefore these methods are not feasible for earlier stages (Zviran *et al*., 2020). This can be overcome by a cfDNA analysis method based on multiple regions with more complex and transient modifications, entering the field of disease-related epigenetic changes in the tissues of origin, e.g. DNA methylation (Liu *et al*., 2020) or hydroxymethylation (Song *et al*., 2017). Methods targeting the cfDNA methylome are currently in active use in the clinic, however these types of assays are still limited by sequencing depth and at the moment are not affordable enough for frequent use. Also, both epigenetic marks and mutations may not be present at onset and occur over a long period of time with progression of disease and do not tend to change in response to therapy. Therefore, methods that target more subtle modifications at a wider range of genomic regions would be more favourable. Nucleosome positioning in particular appears to be an ideal candidate since it can be reconstructed from cfDNA fragmentation patterns and responds easily to treatment and progression of disease.

1.2.1. “Fragmentomics” and “nucleosomics”. The new fields that emerged with novel methods based on the analysis of cfDNA fragmentation patterns and the reconstruction of nucleosome positioning are known as “fragmentomics” and “nucleosomics” (Im *et al*., 2021). Fragmentomics is concerned with the range of cfDNA fragment sizes (Snyder *et al*., 2016; Underhill *et al*., 2016; Mouliere *et al*., 2018) and enzyme signature nucleotides at start sites where the fragments are digested (Chandrananda *et al*., 2015). The lengths of individual fragments are defined by their origin – the majority arise in the process of necrosis and apoptosis, or through the action of neutrophil extracellular traps (NETosis). Enzymes that drive apoptosis digest DNA into fragments that are shorter than the average length of mononucleosomal DNA (Han *et al*., 2020). These shorter fragments appear to be more prevalent in cancer patients compared to controls (van der Pol and Mouliere 2019). This is distinct from the inflammatory process that is NETosis, where unusually large cfDNA fragments are produced by neutrophils as they release chromatin-based structures (neutrophil extracellular traps or NETs) that capture and dispose of pathogens (Kustanovich *et al*., 2019). These ultra-long fragments can be expected to be associated with inflammation, e.g. in COVID-19 (Ng *et al*., 2021), lupus erythematosus (Snyder *et al*., 2016) and diabetes (Wong *et al*., 2015). In this way, various processes of cell death leave their footprints in cfDNA fragment size distribution. It can also vary according to origin of the sample, for example urine cfDNA appears to be shorter than plasma cfDNA (van der Pol and Mouliere, 2019). In the development of early methods in fragmentomics, it was hoped that something as simple as a ratio of the counts of short versus long fragments can be used to calculate the fraction of ctDNA to cfDNA (van der Pol and Mouliere, 2019). However, with the current understanding of the aforementioned variation of cfDNA fragment sizes according to cellular origin, it may not be feasible. Other fragmentomics methods are based on the analysis of nucleotide patterns at cut sites, which can be informative because different types of nucleases (which vary by cell death process) have different nucleotide sequence preference (Han *et al*., 2020). Thus, comparing nucleotide patterns at cfDNA fragment ends could potentially yield important information about a medical condition (van der Pol and Mouliere, 2019).

1.2.2. Nucleosomics for cfDNA-based diagnostics. Methods based on cfDNA nucleosomics are potentially very desirable as they do not require the knowledge of particular genomic features or existing hypotheses about a disease, and the need for further discovery and development is now on the side of bioinformatics. Some novel methods applied machine learning to identify tissues of origin or classify a sample as healthy or cancerous relying on the density of cfDNA fragments that overlap with promoters (Snyder *et al*., 2016; Wan *et al*., 2019). There is also a method that combines some "simple" genomic features to perform principal component analysis (PCA). This method seems to work well by classifying a sample as healthy or cancerous in a binary fashion with various types of cancer merged together. Various methods of sample stratification and visualisation have been studied in our laboratory previously (Piroeva *et al.*, 2022). PCA based on nucleosome occupancy proved to be most effective when distinguishing CLL from healthy samples, as well as two subtypes of CLL (with mutated or unmutated immunoglobulin heavy chain region (IGHV), named M-CLL and U-CLL respectively). Here we continue to apply this method in the context of ageing, as well as other cancer types.

1.2.3. Use of transcription factor binding sites in cfDNA nucleosomics. Ulz *et al*. (2019) created a pipeline (https://github.com/PeterUlz/TranscriptionFactorProfiling) for assessing TF activity by analysing cfDNA sequencing data and nucleosome occupancy. In this method, transcription factor binding sites are obtained from GTRD database (http://gtrd.biouml.org/) (Yevshin *et al*., 2019) and then coverage data around each transcription factor binding site are averaged and the activity for the TFs is calculated. This approach allows to map tumour-specific TF activity *in vivo* and offers the potential to make the non-coding genomic regions available for clinical use. As an alternative, the Teif lab has considered using both experimentally determined and computationally predicted TF binding sites that are specific for a given cancer (Takahashi *et al*., 2022).

1.2.4. Machine learning based on cfDNA patterns. With a constant influx of advances in computational methods in genomic analysis and a growing availability of sequencing data, machine learning is becoming more of interest in the field of medical therapeutics and diagnostics. Some of the earliest efforts to use machine learning on cfDNA data originated in the field of non-invasive prenatal diagnostics, one of the well-known cfDNA success stories in the clinic. Neocleous *et al*. (2015) applied artificial neural network (ANN) methods to assess the suitability of cfDNA in estimating the risk of developmental abnormalities. ANN-based methodology was reported to correctly identify all cases of trisomy 21 in this study, potentially providing an effective and non-invasive early-stage screening tool for foetal aneuploidies.

Che *et al*. (2022a) applied a machine learning workflow named GIP*Xplore* to large cfDNA shallow whole genome sequencing (WGS) datasets and identified cancer signatures. This pipeline reduces data dimensionality by using the first 50 principal components of genome-wide features obtained from linear transformation of many individual cfDNA profiles, which are used to discover the underlying biological signals (such as the differences in cfDNA fragmentation between different conditions, e.g. cancer vs healthy) based on the stratification of causal patterns that allow the creation of the supervised machine learning models. Also, the pipeline allows to identify disease status and type to examine the use of these transformed features as a diagnostic marker, possibly even for a range of cancers at once (e.g. pan-cancer). The same principle was successfully applied to autoimmune diseases and inflammatory bowel disease (Che *et al*., 2022b).

Similarly, Lee *et al*. (2022) created a high-specificity diagnostic pipeline by integrating machine learning. The study stratifies hepatocellular carcinoma (HCC) patients by levels of plasma cfDNA and the expression of α-fetoprotein, a biomarker of HCC, through *k*-means clustering, a machine learning method that helps organise data into clusters according to its features, and then uses PCA for dimensionality reduction to develop an integrated cfDNA scoring system.

Thus, many methods for the analysis of cfDNA are already available, including some which are based on nucleosome positioning, which is potentially a very sensitive and cost-effective method. Further development of bioinformatic pipelines and integration of new data analysis techniques for the study of cfDNA are of great importance and promise to disease management in the clinic. Here I provide the results of several pilot studies testing alternative directions of cfDNA analysis based on nucleosomics.

2. Methods

2.1. Sources of sequenced data. My computational analysis is based on datasets detailed in Table 1. These include original experimental data from breast cancer patients generated in the Teif lab (referred to as Jacob *et al*., 2022), experimental data from SCLC patients provided by our collaborator Prof Anish Thomas (National Institutes of Health, USA, referred to as Takahashi *et al*., 2022), as well as datasets obtained from third-party publications.

The dataset of Snyder *et al*. (2016) available in open access is one of the first whole genome deep-sequenced cfDNA datasets for several cancer types (including breast cancer) and four healthy controls. Three of the healthy samples in this study came from individual volunteers and one was pooled from an unknown number of healthy individuals. In this study I use the Snyder *et al*. data for the healthy controls and six breast cancer patients, all from individuals with stage 4 cancer but different cancer subtypes.

The dataset of Jacob *et al*. (2022) contains original experimental data from the Teif lab. I used data from four patients, numbered using the following anonymised patient IDs: 138, 311, 1670 and 1690. For patients 138, 1670 and 1690 both MNase-seq and MNase-assisted histone H3 ChIP-seq datasets were available with a total of two samples per patient where one comes from the breast cancer tumour and another from surrounding healthy tissue of the same patient. For patient 311 only MNase-seq datasets were available, consisting of paired normal/tumour breast tissues. In addition, for patients 138 and 1670 we had sequenced cfDNA with two replicates for patient 1670 and one replicate for 138.

The dataset of Takahashi *et al*. (2022) was kindly provided to us by our collaborators led by Prof Anish Thomas (National Institutes of Health). This dataset is a longitudinal study of cfDNA from 20 patients with SCLC, including three timepoints consisting of pre-treatment, disease progression and post-treatment stages. The metadata for this dataset included information about tumour fraction in cfDNA and two treatment outcomes consisting of individual patient resistance to the treatment performed in this study and sensitivity to cisplatin (which was applied in a separate treatment round).

The dataset of Teo *et al*. (2019) included sequenced cfDNA from 12 volunteers in three different age groups: three healthy 25-year-old individuals, three healthy 70-year-old individuals and six centenarians. Out of the six centenarians, three were unhealthy as defined by lack of autonomy compared to the 3 healthy centenarians. Each age group included at least one male and at least one female.

The dataset of Peneder *et al*. (2021) was kindly provided to us within a material transfer agreement, and included data for 22 healthy individuals aged 24-50, marked “Ctrl” and followed by the person’s anonymised ID in the form of a number from 1 to 22. This dataset was originally reported by Peneder *et al*. in the frame of a larger study as an axillary control dataset assessing an age-matched cancer cohort, but here we use only the data from the healthy people for the purpose of investigating the effects of age.

The dataset of Piroeva *et al*. (2022) which has been prepared in the Teif lab contains processed nucleosome positioning data in B-cells from peripheral blood of patients with chronic lymphocytic leukaemia (CLL) and healthy controls, based on the raw sequencing data reported in Mallm *et al*. (2019). The dataset contains data from three pools of healthy people, two replicates per each pool, and 28 samples from patients with CLL consisting of two groups: 12 patients with M-CLL (IGHV-mutated CLL) and 2 patients with U-CLL (IGHV-unmutated CLL), with two replicates each.

Table 1 summarises the types of data used in this work. All cfDNA data analysed in this study came from blood plasma; some samples were taken from individual patients or volunteers, and some were pooled from multiple individuals, as indicated in the table.

Table 1. Sources of sequencing data used in this project.

Condition	**Experiment type**	**No. of samples/patients**	**Citation**	**Accession number**
Breast cancer	cfDNA	6	Snyder *et al*., 2016	SRR2130004, SRR2130011, SRR2130032, SRR2130033, SRR2130043, SRR2130045
Healthy	cfDNA	4 (including 1 sample pooled from an unknown number of individuals and 3 from individual people)	Snyder *et al*., 2016	SRR2129993, SRR2130050, SRR2130051, SRR2130052
Paired tumour and healthy breast tissues	MNase-seq and MNase-assisted histone H3 ChIP-seq	4	Jacob *et al*., 2022 (sequencing data obtained in the Teif lab)	
Breast cancer	cfDNA	3	--//--	
Lung cancer	cfDNA	20, three progression timepoints and two treatment timepoints	Takahashi *et al*., 2022	

Ageing (25-100 y.o.)	cfDNA	12	Teo *et al*., 2019	SRR7170698-SRR7170709
Ageing (24-50 y.o.)	cfDNA	22	Peneder *et al*., 2021	EGAD00001007080
Healthy people	MNase-assisted histone H3 ChIP-seq in B cells	28 pooled into 3 batches, 2 replicates	Piroeva *et al*., 2022, Mallm *et al*., 2019	EGAS00001002518

2.2. Calculations setup. All sequenced genomic datasets were processed using the high-performance computation cluster at the University of Essex (ceres.essex.ac.uk) through bash programming in the Linux environment with PuTTY terminal (at https://www.putty.org/). The HOMER suite (version 4.10) (Heinz *et al*., 2010) was used to calculate the aggregate occupancy profiles and NucTools (version 5.0) (Vainshtein *et al*., 2017) for nucleosome occupancy profiles. Principal component analysis and cfDNA fragment size analysis were calculated using R (version 3.6). DAVID (v2022q4) (Sherman *et al*., 2022), an online tool, was used to search for the applicable gene ontology terms. All graphing was performed in OriginPro 2020 (originlab.com). The Perl scripts used in this study which were written earlier are available as part of the NucTools package (Vainshtein *et al*., 2017). Additional Perl and R scripts developed in the Teif lab are available on GitHub: https://github.com/TeifLab/cfDNAtools.

2.3. Acquisition of external sequencing data. Files with raw reads in .fastq format were downloaded from the Short Read Archive (SRA) in the case of Snyder *et al*. (2016) and Teo *et al*. (2019) by using the “wget” command in Linux interactively or through SRA Tools to download the files from the SRA archive and split them into two as the original libraries are paired-end in both studies. Peneder *et al*. (2021) data were downloaded within a material transfer agreement. The sequenced data from Jacob *et al*. (2022) were acquired originally by the Teif lab. The raw sequenced data described in Piroeva *et al*. (2022) are available in the EGA database (EGAS00001002518). The SCLC data from Takahashi *et al*., 2022 were kindly provided by Prof Anish Thomas. The coordinates of the top-5000 CpG sites associated with ageing were kindly provided by Yucheng Wang from the lab of Prof Leo Schalkwyk. All data was mapped to the human genome assembly hg19 which has been used throughput this project.

***2.4. Sequenced reads alignment and pre-processing*.** Sequenced reads were mapped using Bowtie (version 1.2.2) (Langmead *et al*., 2009) to the hg19 human reference genome with parameters set for paired-end reads, allowing up to 2 mismatches, only considering uniquely mappable reads and suppressing all alignments for a read if more than 1 reportable alignments exist for it. The remaining pre-processing was carried out with NucTools. The output of Bowtie, files in .map format, were converted to .bed format using “bowtie2bed.pl” script, which converted the .map file to a .bed format which is gzip-compressed, and the paired-end reads (two lines per paired read) were combined into one line. Using script “extend_PE_reads.pl“, the mapped paired-end reads were sorted by read name and one line per nucleosome into the following columns (left to right): chromosome (chr[chromosome number]), nucleosome start coordinate, nucleosome end coordinate, nucleosome length in bp. The mapped .bed files were split into individual chromosomes using “extract_chr_bed.pl”, which separated the .bed file with the mapped reads for the whole genome into smaller .bed files, each containing reads for one individual chromosome.

2.5. Analysis of cfDNA fragment size distribution. Fragment size histograms were calculated using an R script, “fragment_length_histogram.r” from cfDNAtools available at https://github.com/TeifLab/cfDNAtools, which takes .bed files with nucleosomes as input and produces histograms with fragment sizes in .txt format. The resulting files were plotted in Origin.

2.6. Calculating average occupancies across chromosomes. Nucleosome occupancy per individual sample was calculated using “bed2occupancy_average.pl” script from NucTools, which takes aligned reads in .bed format and converts them to .occ files for each chromosome with occupancy calculated within 100 bp windows. I used NucTools to convert the files from .occ to .bed and duplicated column 2 to fit format requirements for the next step.

2.7. Comparison of nucleosome occupancy in two different conditions. The comparison of nucleosome occupancy landscapes was performed using the NucTools workflow as detailed previously (Vainstein *et al.*, 2017; Piroeva *et al*., 2022). This provided the basis for defining genomic regions that lost/gained or had stable nucleosome occupancy genome-wide in cancer and ageing. The genomic regions which had increased nucleosome occupancy in cancer vs control (or in aged vs young people) were referred thereafter as “gained nucleosomes”. Similarly, the genomic regions which lost nucleosome occupancy were referred to as “lost nucleosomes”.

In the NucTools analysis, by default the genome was split into 100 bp regions. Regions with unchanged nucleosome occupancy for each condition were defined using “stable_nucs_replicates.pl” which normalised nucleosome occupancy profiles by coverage for each chromosome for each sample. The profiles were averaged within each 100-bp window for all samples in a condition (separately for healthy and breast cancer). The regions with average relative error of <0.5 were named "stable" nucleosome regions and used for subsequent analysis. The stable regions were used for binary comparison of occupancy between the conditions in question (cancer vs healthy, old vs young). The comparison was performed using “compare_two_conditions.pl” which determined the relative change in occupancy (O_{diff}) using the following formula: $O_{diff} = 2 * (<O_{N1}> - <O_{N2}>) / (<O_{N1}> + <O_{N2}>)$, where $< O_{N1} >$ and $< O_{N2} >$ are averaged nucleosome occupancy in conditions 1 and 2 respectively. Regions where the relative change in occupancy exceeded a given threshold were defined as the regions where nucleosome occupancy was lost or gained when comparing conditions. For the data from Snyder *et al*. a sliding window of 100 bp and $O_{diff} = 0.95$ were used to determine relative change in nucleosome occupancy for comparing breast cancer and healthy samples, which yielded 3085 regions that lost occupancy and 2666 regions that gained occupancy). For comparing pre vs post-treatment in SCLC I used a genomic window of 5000bp and a threshold of 0.5 for defining “stable” nucleosomes and 0.3 for lost/gained nucleosomes which produced 918 and 627 regions respectively, for comparing SCLC treatment response the same parameters produced 534 regions that lost nucleosome occupancy and 824 regions that gained nucleosome occupancy; for data from Teo *et al*. I used the threshold of 0.2 for defining “stable” nucleosomes and 0.99 for lost/gained nucleosomes which produced 22 and 271 regions respectively and for data from

Peneder *et al*. I used a threshold of 0.5 for defining “stable” regions and 0.8 for defining regions that lost/gained occupancy, which produced 5710 regions that gained occupancy (which gave 34 gene names when intersected with promoters) and 367 regions that lost occupancy (which gave 60 gene names when intersected with promoters). P-values were calculated with two-sample t-test in OriginPro 2020.

2.8. Intersecting datasets with lost/gained regions. I used the “bedtools intersect” command to find intersecting regions between the datasets and the files containing lost or gained occupancies. The genomic regions that lost or gained nucleosome occupancy in one condition compared to another were intersected with all the available datasets from that study. Some of these datasets were those that were not included in the model, i.e. another type of cancer or an intermediate age group that is not defined as young or old, to test whether this model can be used for prediction and stratification.

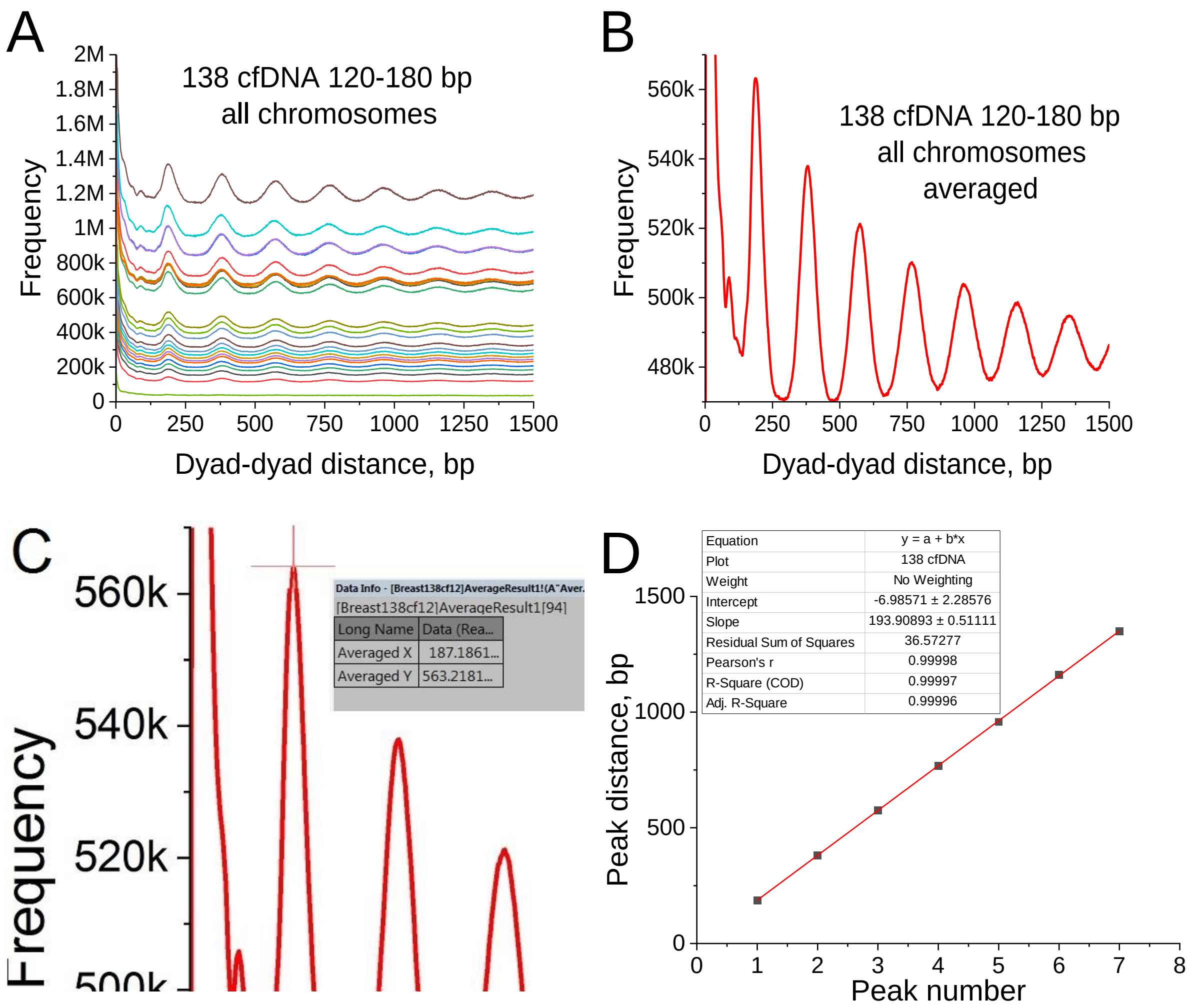
***2.9. Principal component analysis*.** PCA was applied to the normalised nucleosome occupancy scores inside regions defined as those that lost nucleosome occupancy and those that gained nucleosome occupancy using “prcomp” command in R. Then the principal components of the resulting files with were plotted in OriginPro 2020.

***2.10. Gene ontology analysis*.** The .bed files with intersected regions were intersected further with hg19 promoter coordinates to obtain a list of coordinates and names of genes that have promoters in those regions. Then the fourth column with the list of genes was extracted through the “cut” command in Linux, producing a file with a list of gene names. These were copy-pasted into the input field of the DAVID online gene ontology tool and run on default parameters with the exception of output where all options where included. The bar plots were generated in Origin with terms for biological processes, pathways and tissues combined.

2.11. DNA fragment size extraction. DNA fragments with size ranges 160-170 and 120-180 were extracted from .bed files with nucleosome coordinates using cfDNAtools. The calculations were performed separately for each fragment size range and for each sample. BedTools was used for intersections between fragment sizes and genomic regions of interest according to hg19.

2.12. Peak calling of broad regions of nucleosome occupancy change between tumour and normal breast tissues**.** MACS2 (for Python 2.7) (Zhang *et al*., 2008) command "callpeak" was used to call peaks in each sample of each fragment size range with tumour samples as input and healthy samples as controls with parameters "--broad –nomodel". This created a set of broad peaks of nucleosome occupancy chance (as opposed to lost/gained nucleosome regions determined with 100-bp window used at other analysis steps). To ensure that the regions are broad enough for NRL calculations, the broad peaks were further extended by 1000bp on both sides using the BedTools "slop" command. These regions were then used to calculate NRL as detailed below.

2.13. Nucleosome repeat length**.** The nucleosome repeat length (NRL) is the average inter-nucleosome distance, which can be defined genome-wide or for a subset of genomic regions in a given cell state, and then compared between a physiological condition and a healthy state. The protocol for calculating NRL from paired-end reads was developed earlier (Vainshtein *et al*., 2017) and was used here with the help of the NucTools software package. This method works more precisely with paired-end reads as opposed to single-end because it targets nucleosome centres (dyads) directly and uses a different filtering technique. The NucTools pipeline calculates the distribution of frequencies of distances between nucleosome dyads with single-bp resolution separately per each chromosome which are then averaged across all chromosomes. Then the peaks on the distribution of nucleosome-nucleosome distances are identified using the "Annotation" tool in Origin. The location of each peak was plotted as a function of the peak number and then linear regression was used to measure the slope of the linear fit, which gives the corresponding NRL value. I used inter-nucleosome distances of up to 2000 base pairs, applied a limit of 40 million reads for each chromosome and excluded locations with 50 reads or over mapping with their centre on the same base pair. NRL values were calculated both genome-wide and in the proximity of certain types of genomic regions. Regions around Alu and LINE-1 repeats were annotated according to the UCSC Genome Browser’s Repeat Masker (version 1.04.00) for the human genome assembly hg19 and extended by 1,000 bp on both sides using BedTools command “slop”. Figure 1 illustrates the process of calculating NRL. First, files with nucleosome distances are plotted together (Figure 1A), then averaged (Figure 1B), the maximum value of each peak is measured (Figure 1C) and then the peaks are plotted as a scatter graph with a linear fit where the slope equals the NRL value (Figure 1D). The violin plots and Pearson correlation value were calculated and plotted using Origin. Paired-sample t-test was calculated for tumour and normal MNase-seq samples and p-value is given directly above them on Figures 6 and 7.



Data Info - [Breast138cf12]AverageResult1!(A"Aver.
[Breast138cf12]AverageResult1[94]

Long Name	Data (Rea...
Averaged X	187.1861...
Averaged Y	563.2181...

Equation	y = a + b*x
Plot	138 cfDNA
Weight	No Weighting
Intercept	-6.98571 ± 2.28576
Slope	193.90893 ± 0.51111
Residual Sum of Squares	36.57277
Pearson's r	0.99998
R-Square (COD)	0.99997
Adj. R-Square	0.99996

Figure 1. The procedure of NRL calculation following steps A-D. A). Files with nucleosome distances are plotted together. B) Average of the profiles shown in (A) across all chromosomes. C) Defining the X coordinates (genomic distance in bp) for each peak. D) Linear regression of the peak positions determined in (C) against the corresponding peak numbers. The linear fit shown on the figure gives the NRL value (193.9 ± 0.5 bp).

***2.14. Calculation of nucleosome occupancy profiles*.** Aggregate nucleosome occupancy profiles around genomic features of interest were calculated with HOMER (Heinz *et al*., 2010). HOMER tag directories were created for each sample by taking the mapped .bed file as input. HOMER script “annotatePeaks.pl” matched each region with the gene of the closest TSS and identified the genomic annotation of the centre of that region and by applying parameters hg19 -size 2000 -hist 10, calculated aggregate nucleosome occupancy profiles for each sample. The average of the individual samples within a condition was determined using software Origin Pro 2021 (originlab.com) which was also used for plotting the resulting profiles and calculating Pearson's correlation.

3. Results

3.1. Overview of the types of analysis and the datasets. In this project I tested two different "nucleosomics" approaches for patient diagnostics. The first one considers changes of nucleosome occupancies in condition-specific regions and condition-specific changes of nucleosome positioning and DNA methylation, the second one relies on disease- or condition-specific changes of distances between nucleosomes. These approached were applied in the frame of several projects, addressing the following biomedical systems:

- Classification of healthy vs cancer cfDNA samples applied to the following datasets:
    - cfDNA from breast cancer patients and healthy controls reported by Snyder *et al*., 2016.
    - Original datasets obtained in the University of Essex based on MNase-seq and MNase-assisted histone H3 ChIP-seq in paired healthy/cancer breast tissues and corresponding cfDNA (Jacob *et al*., 2022).
    - cfDNA from lung cancer patients taken before and after treatment and during disease progression (Takahashi *et al*., 2022).
- Classification of cfDNA samples based on a person's age applied to:

- Dataset from Teo *et al*., 2019, cfDNA from 25-, 70- and 100-year-old people
- Dataset from Peneder *et al*., 2021, cfDNA from healthy people aged 24-50
- Nucleosome maps obtained by MNase-assisted histone H3 ChIP-seq in healthy B-cells (Mallm *et al*., 2019; Piroeva *et al*., 2022)

***3.2. Classification of cfDNA samples based on nucleosome loss/gain in cancer-sensitive regions*.** Several types of cfDNA analysis can potentially allow to distinguish between a healthy and a cancer sample. One method I tested is the comparison of nucleosome occupancy in cancer-sensitive regions and applying principal component analysis, in which dimensionality reduction helps visualise the differences in variation between healthy and cancer samples. Another method I tested is through calculating the nucleosome repeat length (NRL) whereby the distances between nucleosomes are compared in cancer vs healthy cfDNA. In both cases the comparison can be made between individual types of cancer vs healthy or several cancer types (pan-cancer) vs healthy. Below are the examples of my application of these methods in several datasets and various types of cancer.

3.2.1. Relative nucleosome occupancy in cancer-sensitive regions in breast cancer versus healthy controls using cfDNA from Snyder et al. Snyder *et al*. (2016) sequenced cfDNA from patients with several different cancer types and healthy volunteers. I decided to concentrate on breast cancer which included data from six patients with various subtypes of breast cancer. To compare nucleosome occupancy in two conditions, in this case breast cancer vs healthy, I defined “stable” nucleosomes that remained unchanged in each condition and then extracted the genomic regions that have changed nucleosome occupancy in one condition compared to another, which were named as the regions that “lost” or “gained” occupancy.

A sliding window of 100 bp and O_{diff} = 0.95 were used to determine relative change in nucleosome occupancy for comparing breast cancer and healthy samples. This produced 3085 regions that lost occupancy and 2666 regions that gained occupancy. Figure 2 shows principal component analysis (PCA) result for the six breast cancer patients compared to the four healthy individuals according to the method detailed above.

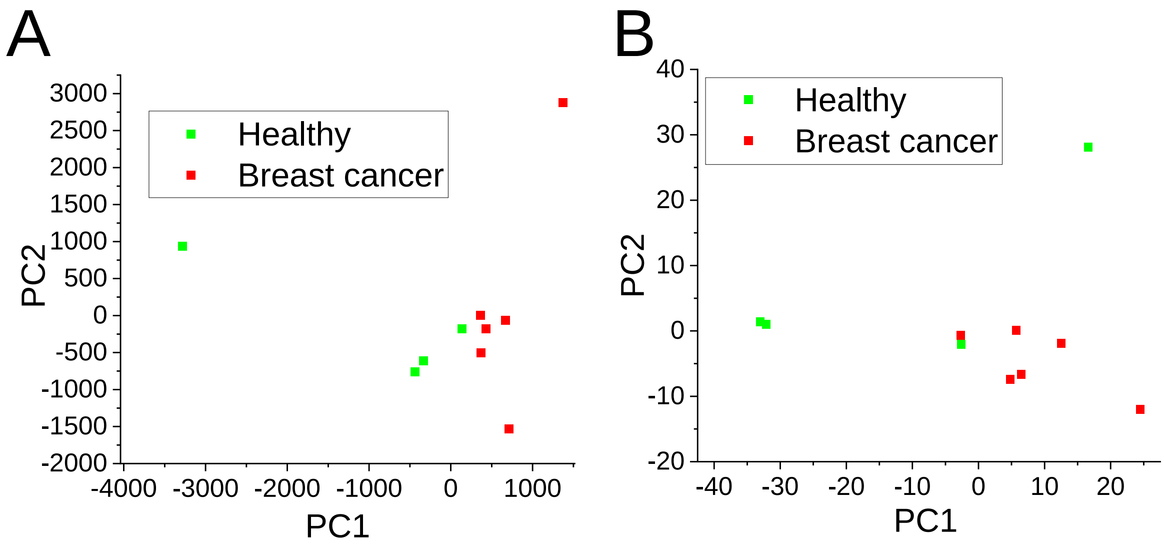


Figure 2. Principal component analysis (PCA) of nucleosome occupancy in cancer-sensitive regions based on whole-genome cfDNA sequencing for breast cancer patients and healthy controls from Snyder *et al*. A) PCA based on genomic regions where nucleosome occupancy decreased (“lost nucleosomes” in breast cancer compared to healthy controls. B) PCA based on genomic regions where nucleosome occupancy increased (“gained nucleosomes”) in breast cancer compared to healthy controls.

3.2.2. Classification of SCLC cfDNA samples based on nucleosome occupancy in cancer-sensitive regions. In this analysis, cfDNA from 20 patients with small cell lung cancer (SCLC) was taken at three time points. Tumour fraction information was also available for each sample.

3.2.3. Relative nucleosome occupancy in cancer-sensitive regions to classify cfDNA samples by cancer stage and progression in SCLC. Nucleosome occupancy at individual nucleosome positions inside condition-specific regions was calculated and averaged per condition, and the regions that lost or gained nucleosomes in SCLC compared to healthy were analysed using principal components. I applied this method to available SCLC data using threshold 0.5 for defining “stable” nucleosomes and 0.3 for lost/gained nucleosomes which produced 918 and 627 regions respectively.

Figure 3A shows distinct clustering that separates pre-treatment, progression and post-treatment samples with progression clustering clearly between the two, despite being excluded in the model that defines lost/gained nucleosomes. Figure 3C shows PCA for gained regions with similar results but with more overlap between the groups. I labelled data points with low tumour fraction (Figure 3B for lost regions and Figure 3D for gained regions) and showed that the exclusion of those samples would significantly improve clustering by removing some of the outliers.

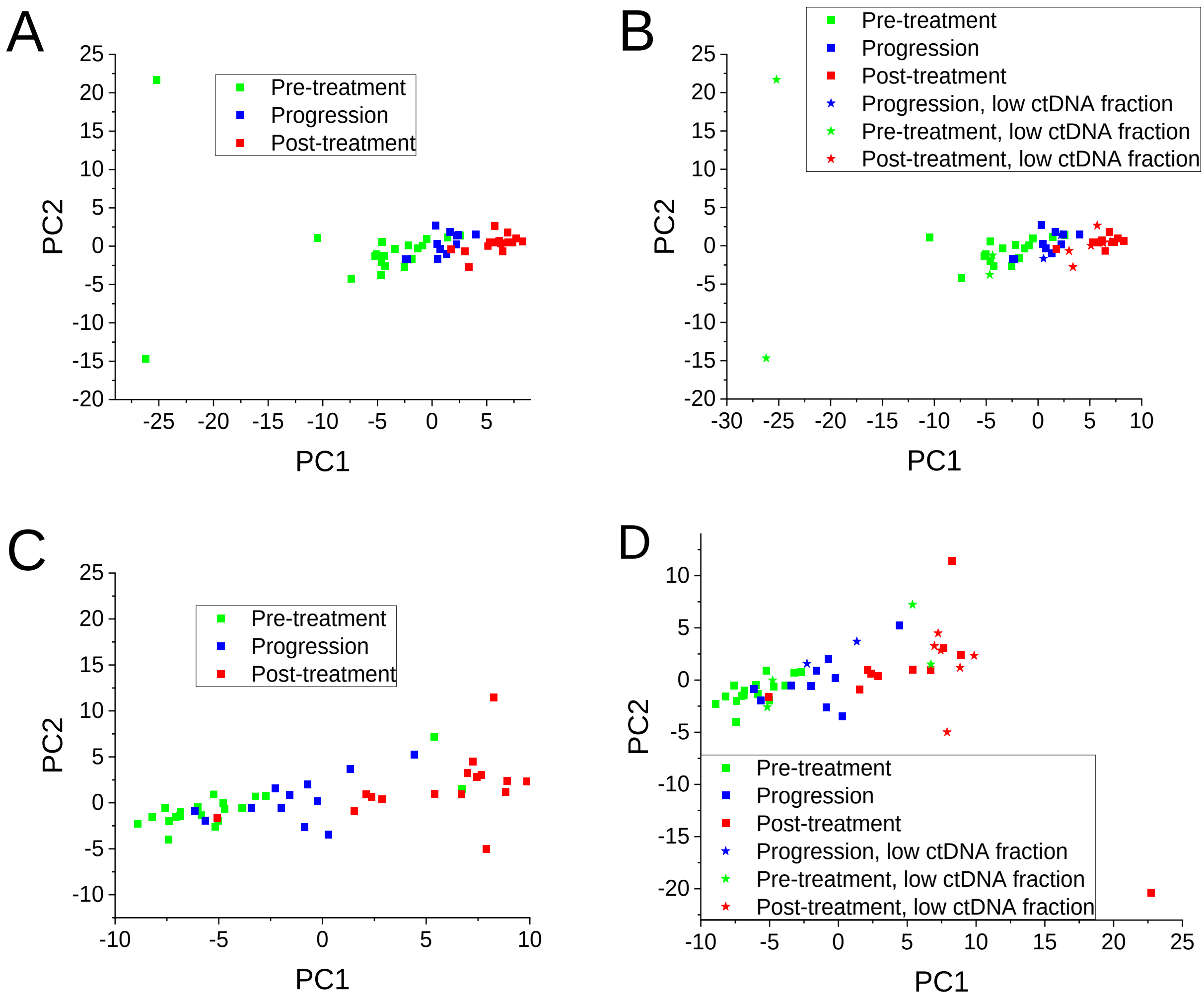


Figure 3. PCA based on nucleosome occupancy in cancer-sensitive regions for cfDNA from SCLC patients taking at three timepoints. A) PCA based on regions that lost nucleosomes. B) PCA based on regions that lost nucleosomes with samples that came from low ctDNA fraction indicated as star-shaped points. C) PCA based on regions that gained nucleosomes. D) PCA based on regions that gained nucleosomes with samples that came from low ctDNA fraction indicated as star-shaped points.

3.2.4. Classification of patient treatment response based on nucleosome occupancy in cancer-sensitive regions. Since the analysis described in the previous section showed promising results by distinguishing different timepoints in the SCLC study, I decided to test whether this method can also classify patients as treatment resistant or treatment responsive. I applied the same method with the same parameters as previously, which produced 534 regions that lost nucleosome occupancy and 824 regions that gained nucleosome occupancy. Figure 4A shows distinct clustering that separates treatment sensitive and treatment resistant patients in regions that lost nucleosomes and PCA in C) is based on regions that gained nucleosomes. Removing the samples with low ctDNA fraction (Figure 4B and D respectively) significantly improved the clustering.

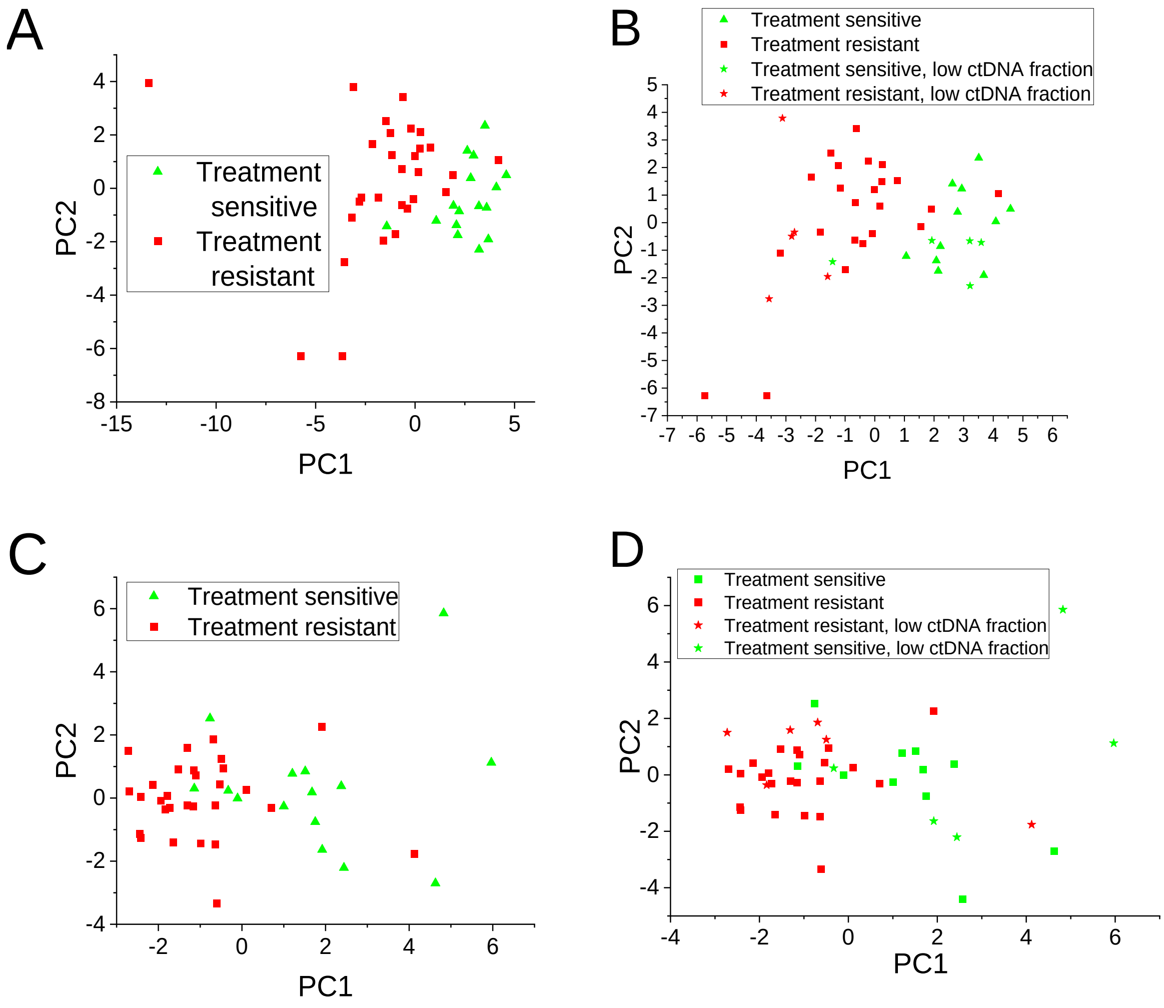


Figure 4. PCA of nucleosome occupancy in cancer-sensitive regions for SCLC patients showing different response to treatment. A) PCA based on regions that lost nucleosomes. B) PCA based on regions that lost nucleosomes with samples that came from low ctDNA fraction indicated as star-shaped points. C) PCA based on regions that gained nucleosomes. D) PCA based on regions that gained nucleosomes with samples that came from low ctDNA fraction indicated as star-shaped points.

3.3. Age-classification based on nucleosome occupancy changes in ageing-sensitive regions. The motivation to combine the study of cancer and ageing in the current work is that age as a comorbidity can affect the effectiveness of PCA-based cancer prediction. Teo *et al*. (2019) reported sequenced cfDNA from a group of 12 individuals in three age groups: 25 years old (three healthy individuals), 70 years old (three healthy individuals) and 100 years old (three healthy and three unhealthy individuals as defined by lack of autonomy in the original study).

Figure 5A shows PCA based on nucleosome occupancy in regions that lost nucleosome occupancy in 100-year-olds compared to 25-year-olds. Here I used the threshold of 0.2 for defining “stable” nucleosomes and 0.99 for lost/gained nucleosomes which produced 22 and 271 regions respectively. The three age groups cluster distinctly from each other, mostly along the PC2 axis. Healthy and unhealthy centenarians seem to cluster together for the most part, with the exception of one sample which is well within the 70 year olds’ cluster. The same is seen with regions that were gained in 100 year olds compared to 25 year olds (Figure 5B). This may be due to the difference in “biological” age as opposed to chronological age, which was the parameter by which these groups were classified.

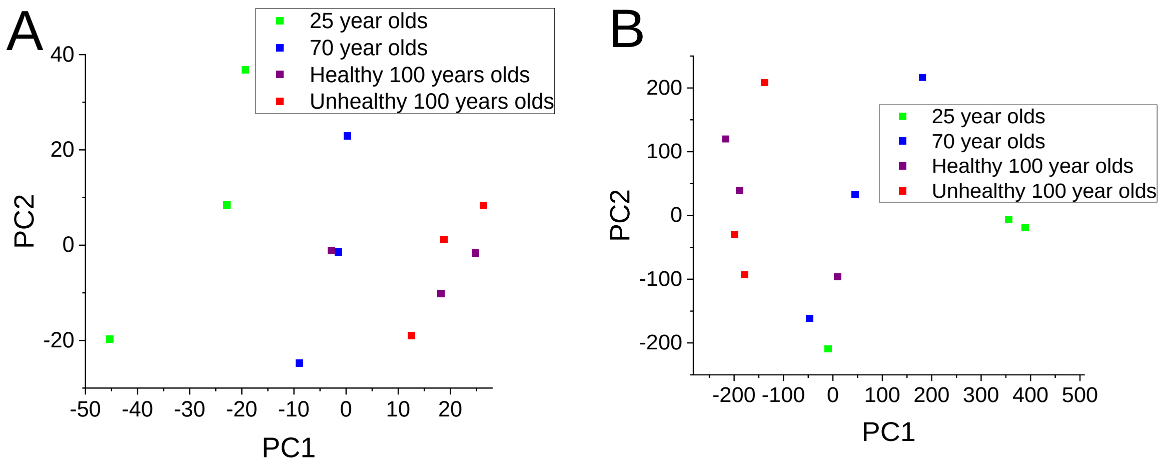


Figure 5. PCA distinguishes three age groups from Teo *et al*. A) PCA based on regions that lost nucleosomes in 100 year olds compared to 25 year olds. B) PCA based on regions that gained nucleosomes in 100 year olds compared to 25 year olds.

3.4 cfDNA sample classification based on the nucleosome repeat length (NRL).

The following method is based on the idea to calculate the average inter-nucleosome distances (a parameter known as the nucleosome repeat length, NRL), which can then be compared between a physiological condition of interest and a control state.

3.4.1. Assessing the heterogeneity of NRLs in healthy cfDNA. I took openly available healthy cfDNA data from Snyder *et al*. (2016), two of which were taken from individuals and one pooled from an unknown number of healthy volunteers. Table 2 shows peak values and NRL for the three samples. The average NRL value is 193.3 with maximum 1.9 bp difference. Interestingly, the NRL of the pooled healthy sample does not deviate from the two individual samples.

Table 2. Peak positions and NRL values for control cfDNA samples from Snyder *et al*. (2016).

Peak number	Peak position, bp		
	Sample SRR2129993	**Sample SRR2130050**	**Sample SRR2130051**
1	191.2	189.2	189.2
2	383.4	383.4	379.4
3	575.6	575.6	571.6
4	771.8	769.8	763.8
5	966	962	962
6	1158.2	1160.2	1148.15
7	1352.35	1354.35	1340.3
NRL	**193.7 ± 0.2**	**194.1 ± 0.3**	**192.2 ± 0.5**

3.4.2. NRL analysis in paired healthy/cancer breast tissues and corresponding cfDNA. In the sequencing experiments performed in the Teif lab, University of Essex (Jacob *et al*., 2022), we considered 4 breast cancer patients. MNase-seq and MNase-assisted H3 ChIP-seq done for each patient for paired healthy and tumour tissue samples, and in addition for two of these patients we sequenced cfDNA (one of these was included in two replicates). Here I used these data to calculate NRL in subsets of genomic regions of interest including Alu repeats and LINE1 repeats as well as individual chromosomes. The aim of this analysis was to check whether NRL in one of these types of genomic regions can be used as a better marker for diagnostics. Using smaller parts of the genome as opposed to the whole genome would also reduce the cost of diagnostic assays.

3.4.3. NRL inside Alu repeats. In the following analysis I applied the NRL method to breast cancer samples defined above, targeting the regions that surround Alu repeats. Table 3 shows the NRL values calculated inside those regions. Figure 6 compares the results between healthy tissue, cancerous tissue and cfDNA. The H3 ChIP-seq sample for cancerous tissue in patient 1690 appears to be an outlier (184.2 bp), compared to the average of ~190 bp in other cancer samples. Even the error range of 1.4 bp may not be sufficient to explain this by lower sequencing coverage compared to other samples. Overall, the difference between averaged NRL inside Alu repeats in normal and cancerous tissues is 9 bp and about 8 bp between healthy tissue and cfDNA, which is expected due to contribution of ctDNA.

Table 3. NRL values inside genomic regions surrounding Alu repeats, calculated for nucleosome positioning in paired tumour/normal breast tissues from BRC patients (Jacob *et al.*, 2022).

Sample ID	**Normal**	**N Error**	**Tumour**	**T Error**	**cfDNA ID**	**CfDNA**	**cfDNA Error**
311 MNase-seq	198.6	0.6	190.6	0.5	138 cfDNA	191.5	0.1
138 MNase-assisted H3 ChIP-seq	194.55	1.3	189.75	0.9	1670 cfDNA 1	193.7	0.7
138 MNase-seq	189.2	0.1	190	0.5	1670 cfDNA 2	189.6	0.6
1670 MNase-assisted H3 ChIP-seq	205.1	1.1	196.5	0.7			
1670 MNase-seq	205.1	1.3	192	1.1			
1690 MNase-assisted H3 ChIP-seq	206.4	8	184.2	1.4			
1690 MNase-seq	199.4	1.3	190.4	1			

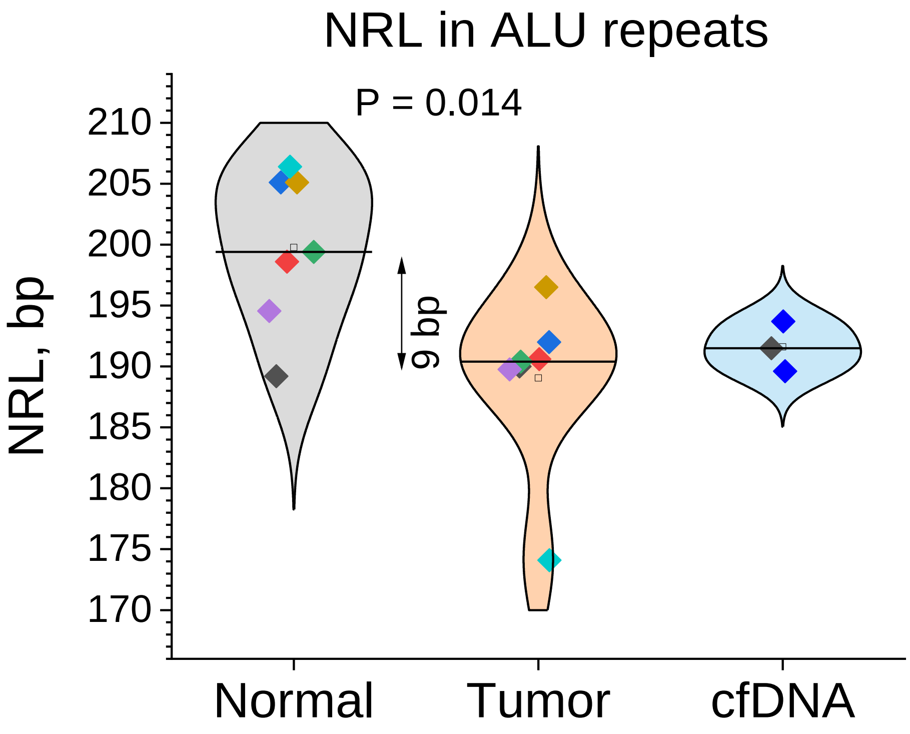


Figure 6. NRL values calculated for individual samples inside regions surrounding Alu repeats. The squares represent average NRL values for a breast cancer patient with each patient having their own colour; horizontal lines – median values. Paired-sample t-test p-value is indicated on the figure.

3.4.4. *NRL inside LINE1 repeats.* Then I applied the same method to genomic regions surrounding LINE1 repeats. Table 4 shows NRL values calculated inside those regions. The shortening of NRL in tumour tissue is again observed but is not as pronounced. Figure 7 compares the results between healthy tissue, cancerous tissue and cfDNA. Compared to regions surrounding ALU repeats, there appears to be less distinction and more overlap between cancer and healthy tissue, but cfDNA NRL values appear to be closer to tumour NRL values.

Table 4. NRL inside regions enclosing LINE1 repeats.

sample ID	Normal	N Error	Tumour	T Error	cfDNA ID	cfDNA	cfDNA Error
138 MNase-seq	189.8	1.3	189.5	1.2	138 cfDNA	191	0.5
311 MNase-seq	193.7	1	187.3	1.6	1670 cfDNA-1	191.4	0.9
1670 MNase-seq	198.4	0.9	192.5	0.9	1670 cfDNA-2	190.5	0.6
1690 MNase-seq	192.3	0.4	192.2	1.75			
138 MNase-assisted H3 ChIP-seq	191.2	0.6	189	1.1			
1670 MNase-assisted H3 ChIP-seq	197.1	1	191.2	1.1			
1690 MNase-assisted H3 ChIP-seq	198.5	1	191.6	1.9			

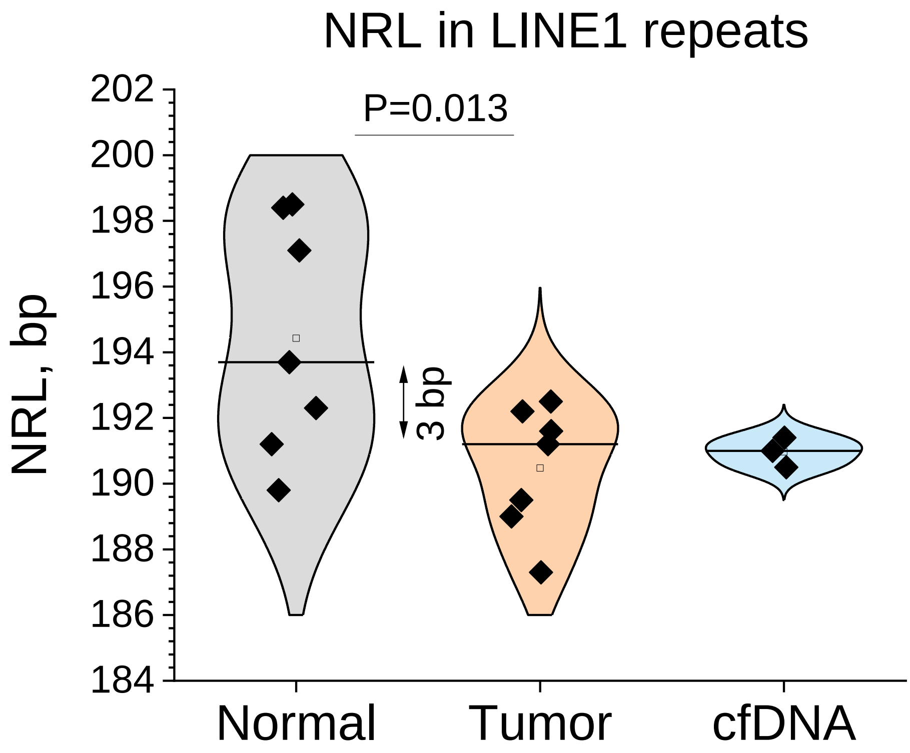


Figure 7. NRL values calculated for individual samples inside regions surrounding LINE1 repeats. The horizontal lines are median values. Paired-sample t-test p-value is indicated on the figure.

3.4.5. Chromosome-wide nucleosome reorganisation in cancer. We hypothesised that some chromosomes may have a greater effect on NRL shortening than others. Figure 8A shows chromosome 7 in the tumour sample of patient 311 inside broad regions (see Methods 2.12). The rationale behind isolating individual chromosomes is to look for the effect of DNA sequence repeats on NRL in cancer vs healthy tissue samples. Chromosome 7 (in cyan) can be visibly distinguished as more clearly defined, with sharper peaks separated by smaller distances compared to other chromosomes. Figure 8B compares chromosome 7 in normal and tumour samples from this patient calculated using the same broad peaks of increased nucleosome occupancy. The peaks of the tumour sample are sharper and are visibly separated by shorter distances than those of the healthy sample. Figures 8C and D show NRL values for healthy (197.75 bp) versus tumour sample (171.8 bp) which is a significant difference of 25.95 bp. With this magnitude it is possible that chromosome 7 alone may make a large contribution to the genome-wide shortening effect seen in 311 tumour sample.

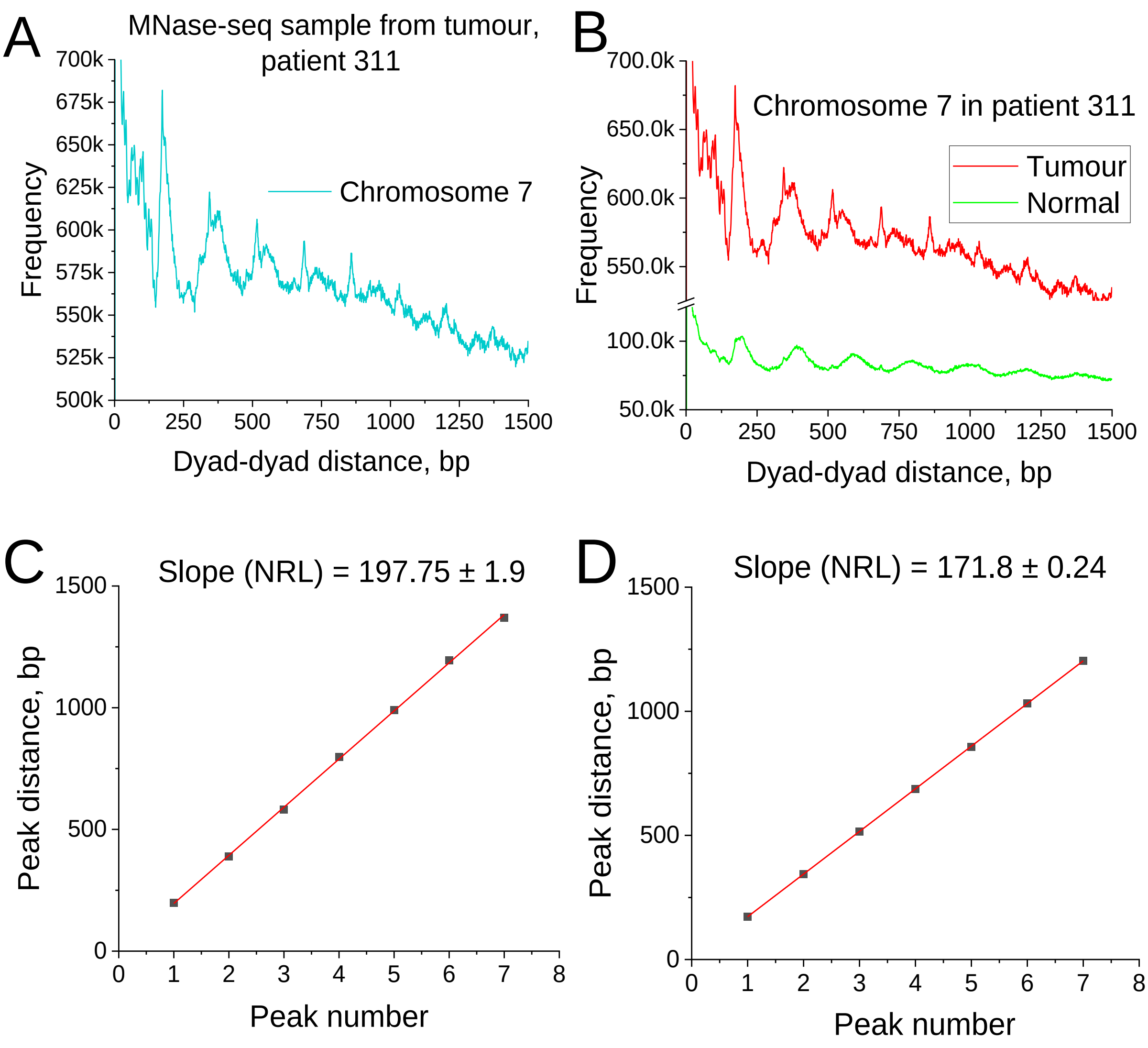
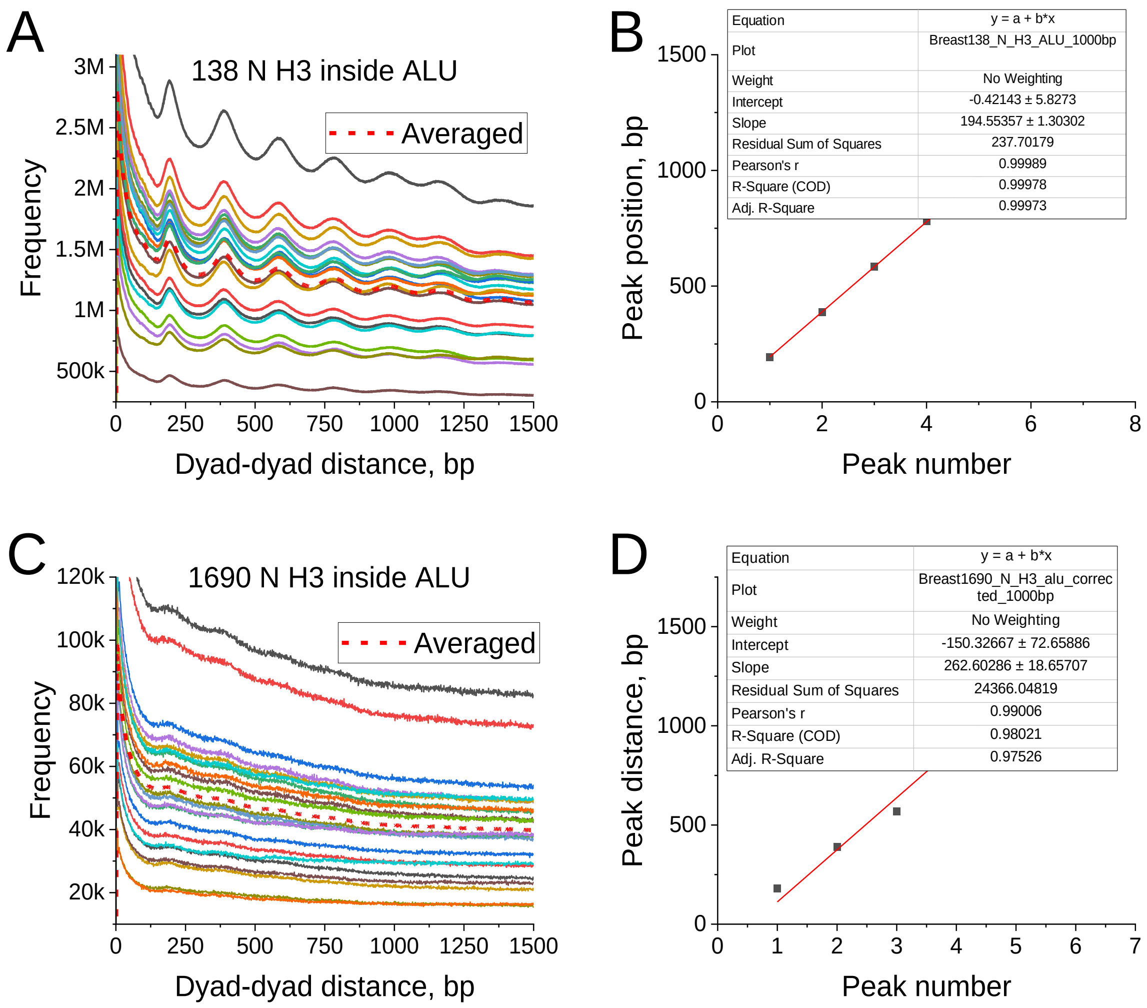


Figure 8. NRL calculation for different chromosomes inside broad regions. A) NRL profiles for individual chromosomes in patient 311, chromosome 7 in cyan has sharper peaks that are positioned differently to the rest. B) Comparison between

normal and breast cancer for chromosome 7. C) and D) NRL calculation for chromosome 7 for healthy (C) and tumour samples (D).

3.4.6. Limitations of the NRL analysis. The accuracy of NRL prediction was limited by the quality of the data provided. Figure 9A shows a sample with sequencing coverage of sufficient depth to discern the seven NRL peaks by eye, in contrast with Figure 9C where a sample with more shallow coverage is shown, where most peaks are hardly discernible and therefore are difficult to pick. Figure 9B shows a better linear fit between the peaks and a more accurate value with error of 1.3 bp. Figure 9D shows a worse fit and a greater uncertainty in the error value of 18.7 bp. This renders the sample unusable unless the outlying points are removed.



Equation	y = a + b*x
Plot	Breast138_N_H3_ALU_1000bp
Weight	No Weighting
Intercept	-0.42143 ± 5.8273
Slope	194.55357 ± 1.30302
Residual Sum of Squares	237.70179
Pearson's r	0.99989
R-Square (COD)	0.99978
Adj. R-Square	0.99973

Equation	y = a + b*x
Plot	Breast1690_N_H3_alu_corrected_1000bp
Weight	No Weighting
Intercept	-150.32667 ± 72.65886
Slope	262.60286 ± 18.65707
Residual Sum of Squares	24366.04819
Pearson's r	0.99006
R-Square (COD)	0.98021
Adj. R-Square	0.97526

Figure 9. NRL calculation inside regions surrounding Alu repeats. A) Nucleosome dyad-dyad distance distribution calculated for all chromosomes in a sample sequenced with a large sequencing depth allow easy identification of peak positions. B) Linear fit through individual peak locations from (A) showing NRL value with low standard error. C) Nucleosome dyad-dyad distance distribution calculated for all chromosomes in a sample of a low sequencing depth makes it difficult to distinguish peak locations. D) Linear fit through individual peak positions from panel (C) has too large standard error and does not allow to determine NRL in this case.

3.5. Age-classification based on the cfDNA dataset of Peneder et al. Next, I analysed a larger dataset of cfDNA from 22 healthy individuals aged 24-50 reported by Peneder *et al*. I defined 100-bp age-sensitive regions that lost or gained nucleosome occupancy in people >40 years old in comparison with 30-year-olds. In this analysis I defined lost-nucleosome and gained-nucleosome regions using NucTools with a threshold of 0.5 for the relative difference between samples within one condition for defining “stable” regions and 0.8 for the relative difference between samples in two conditions for defining regions that lost/gained occupancy. The meaning of the thresholds for the relative occupancy change is similar to the role of fold-change thresholds in the enrichment analysis. This resulted in 5710 regions which gained nucleosome occupancy (two sample t-test $P< 10^{-100}$) (which gave 34 gene names when intersected with promoters) and 367 regions that lost occupancy (two sample t-test $P = 2.24 \times 10^{-32}$) (which gave 60 gene names when intersected with promoters). To clarify the shape of the aggregate nucleosome occupancy profiles, I calculated nucleosome occupancy profiles around gained-nucleosome regions (Figure 10). The regions that gained nucleosomes in the group of older people are expected to have higher occupancy around the centre of those regions in older people in comparison with younger people. This is indeed the case on this figure where the 40-year-old group (in red) shows increased occupancy compared to 30-year-olds (in black), and the intermediate age group (31-39 years old) is in the middle of the two.

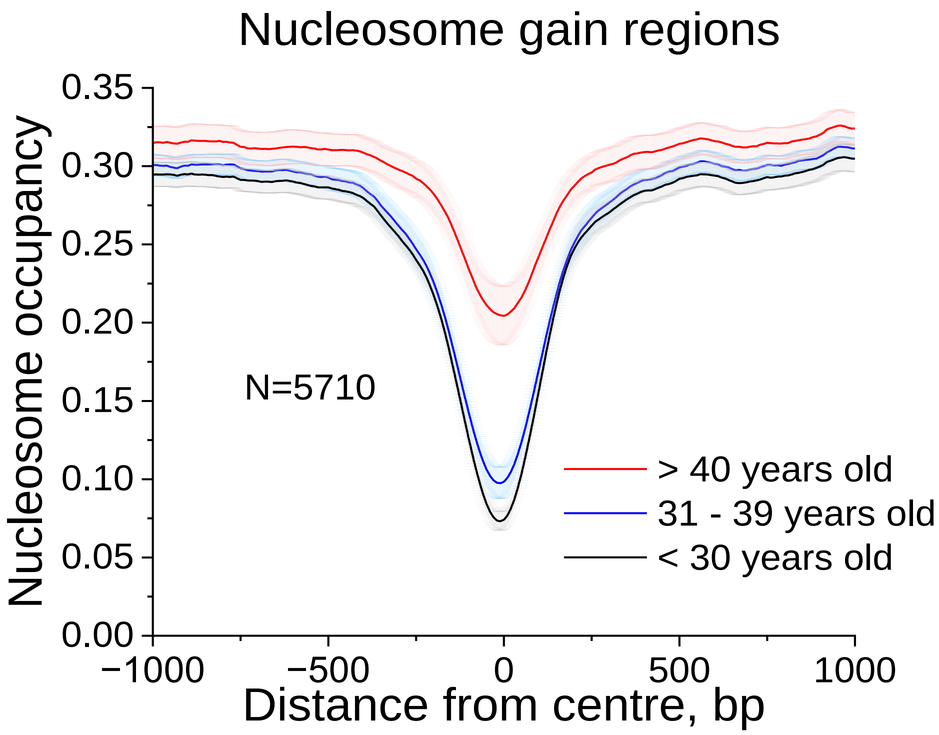


Figure 10. Nucleosome occupancy profiles around centres of regions that gained nucleosome occupancy in ageing. The black line is <30 years old, blue 31-39 years old, and red >40 years old people.

3.5.1. PCA distinguishes three age groups. PCA was based on three age groups: under 30 years old, 31-39 years old and over 40 years old. The first and the last groups were used for defining 100-bp lost-nucleosome and gained-nucleosome regions that correspondingly lost or gained nucleosome occupancy upon ageing.

Figure 11A shows results for PCA based on lost-nucleosome regions and Figure 11B shows results for PCA based on gained-nucleosome regions. The analysis based on gained-nucleosome regions achieved better clustering where the intermediate group (31-39 years old) clustered clearly between the defining groups of under 30 and over 40 years of age. In the analysis based on lost-nucleosome regions under 30 and over 40-year-old groups also cluster separately but the intermediate group is mixed with the under 30 group. Some of the outliers on both panels were labelled to see if the proximity of the age to one group or another would explain this behaviour, but there does not seem to be a connection.

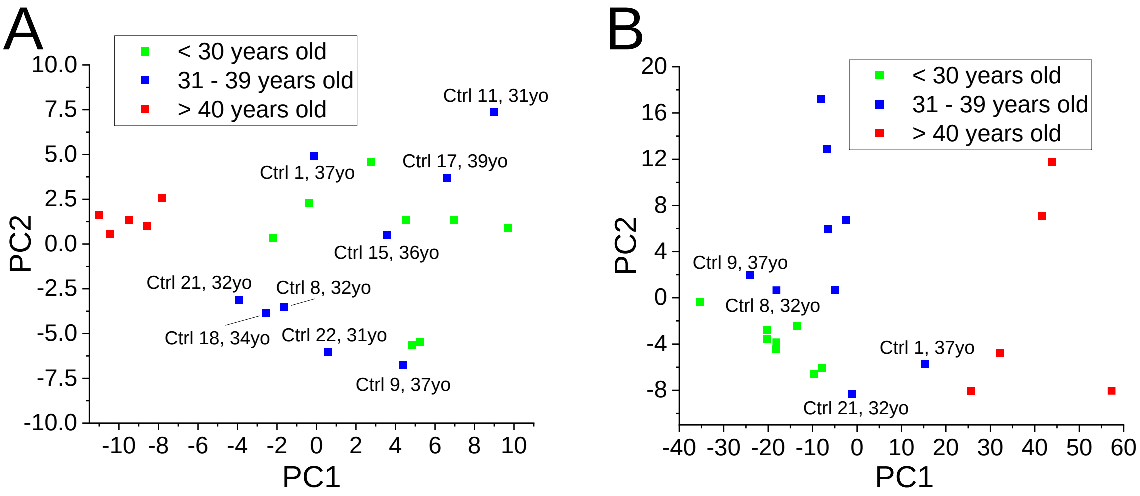


Figure 11. PCA based on cfDNA occupancy in ageing-sensitive regions for the cohort of healthy people from the study of Peneder *et al*. A) PCA based on regions that lost nucleosomes in >40 age group compared to <30 age group. B) PCA based on regions that gained nucleosomes in >40 age group compared to <30 age group. Some of the outliers are labelled.

3.5.2. Genes overlapping with regions that lost and gained nucleosomes in ageing based on data from Peneder et al. I intersected genes with regions that lost/gained nucleosomes and reported all gene names obtained as a result of this intersection. Table 5 gives the list of genes where the promoters happened to be in the range of genomic regions that lost or gained nucleosomes.

Table 5. The list of genes that overlapped with regions that lost or gained nucleosomes in the >40 years age group compared to <30 age group based on cfDNA from Peneder *et al*. (2021).

Genes that lost nucleosomes in ageing	Genes that gained nucleosomes in ageing
PRAMEF5	MIR4419A
PRAMEF22	MIR761
PRAMEF10	SPTA1
PRAMEF33	MIR215
DESI2.	MIR194-1
LARP4B	LOC102724719
FAM25C	C11orf40
FAM25G	LOC103312105
PRR20B	MIR331
PRR20C	SPIC
PRR20D	ELF1
PRR20E	SNORD116-27
PRR20A	BCAR4
LINC01070	MINK1
RAB2B	TMEM259
RAB2B	RNU6-2
TOX4	BIRC8
APH1B	USP40
GALNS	PLCB4
TRAPPC2L	CBX7
COG1	TP63
ELOA3B	IL21
ELOA3D	FYB1
ELOA3B	LRRC70
NDUFA11	IPO11-LRRC70
VMAC	MCTP1
INAFM1	LINC02148
INAFM1	IL5
PAXBP1	RFX3
C21orf62-AS1	GLIS3-AS1
FAM212A	IFNA14
P2RY1	IFNA2
CLPS	ECM2
LOC401357	LOC101928437
FAM90A7P	

DEFB103A DEFB103B VCP SPATA31C1 GAGE12F GAGE12D GAGE12C GAGE12G GAGE12E GAGE6 GAGE12I GAGE12F GAGE12D GAGE12C GAGE12G GAGE12E GAGE6 GAGE12I GAGE12D GAGE12C GAGE12H GAGE12E CSAG3 MAGEA2 SLC10A3	

3.5.3. Gene ontology analysis for ageing data from Peneder et al. Figure 12 provides a summary of gene ontology terms related to the genes that lost (A) or gained (B) nucleosomes in ageing based on the Peneder *et al*. (2021) data.

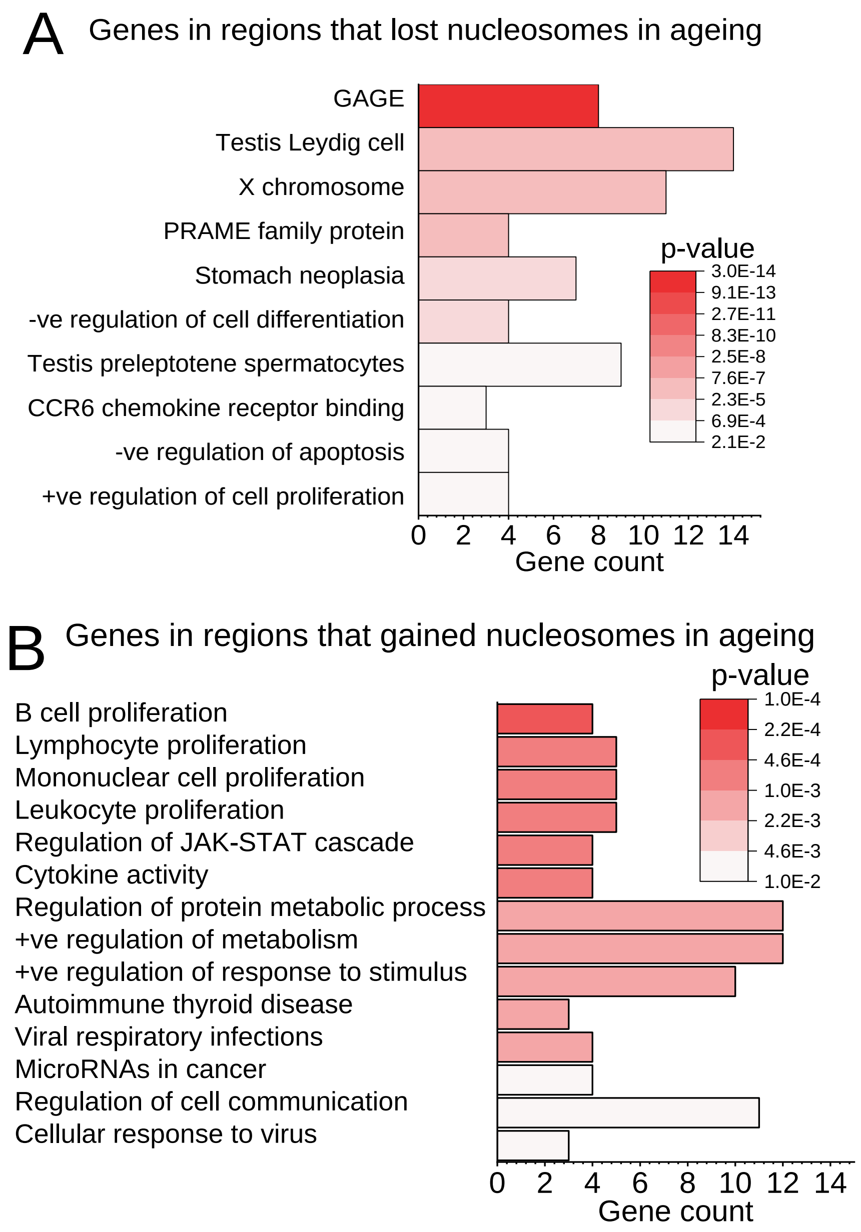
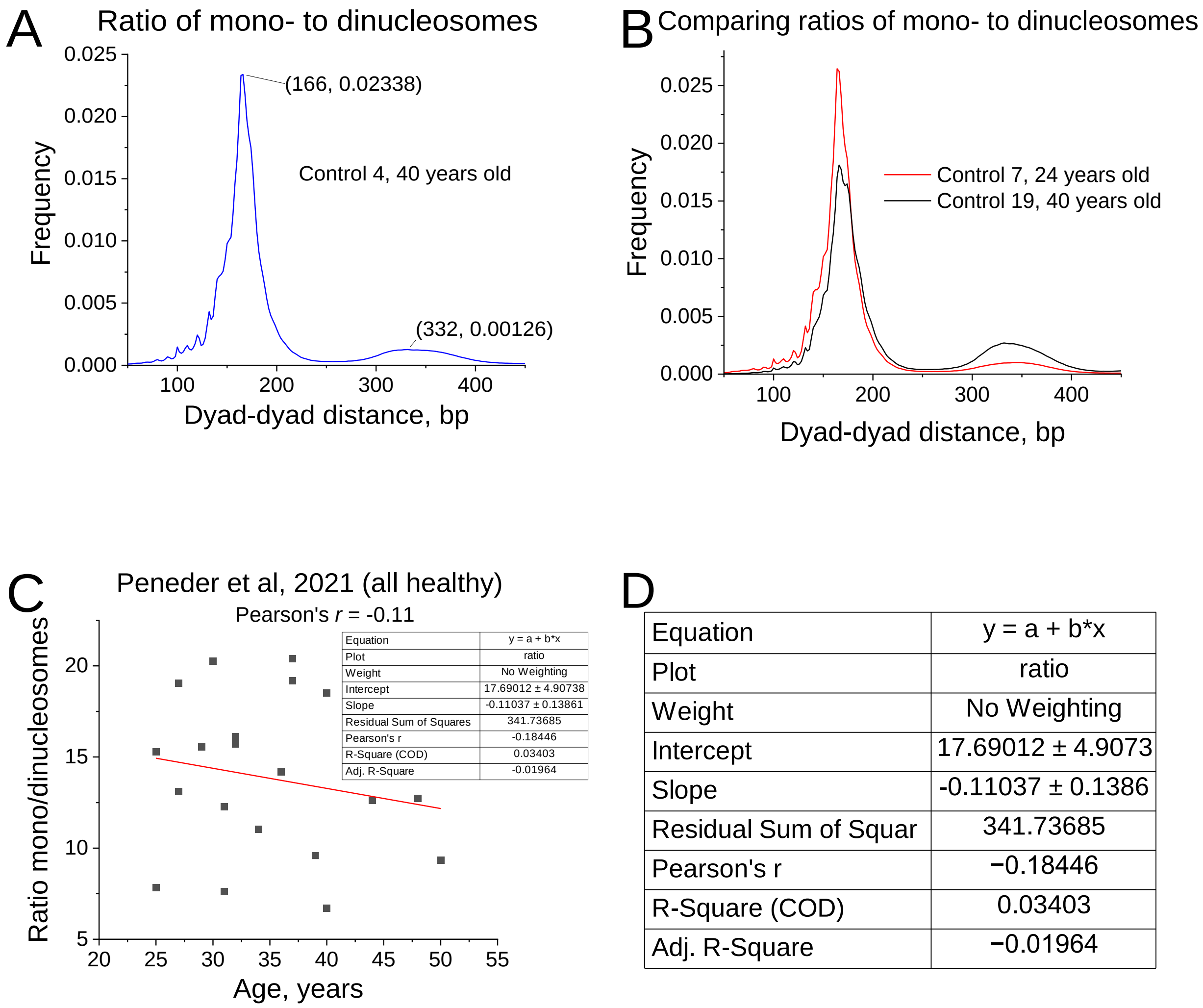


Figure 12. Gene ontology of genes with differential nucleosome occupancy based on cfDNA data from Peneder *et al*. A) Genes that lost nucleosomes in the older group compared to the younger group. B) Genes that gained nucleosome occupancy in ageing.

3.5.4. Distribution of cfDNA fragment sizes reflects biological age. Next, we looked for differences in DNA fragment size distribution in plasma cfDNA using data from

Peneder *et al*. (2021). These differences may arise because chromatin digestion mechanisms are expected to change with age.

Figure 13A shows the method we used to determine differences in fragment size. The first peak of the fragment size distribution, with fragment size ~166 bp, arises from the abundance of mononucleosomal cfDNA fragments, i.e. fragments that were wrapped around a single nucleosome which was then was digested on both sides. The second peak at ~332 bp, came from a less abundant group of dinucleosomal fragments, that were wrapped around two nucleosomes. Mono- and dinucleosomal fragments have different origins therefore the proportion of mono- to dinucleosomal fragments is a good indicator of changes in chromatin digestion. Figure 13B illustrates a difference in such a ratio between a younger and an older individual. Control 7, 24 years old, has a higher abundance of mononucleosomes and fewer dinucleosomal fragments compared to control 19 of 40 years old, which shows the opposite effect. To enable a comparison of all 22 individuals, the ratios of frequencies of mono- to dinucleosomes were calculated for each sample and plotted together on Figure 13C. The linear regression across of all individual-person ratios shows a decrease with age, meaning that the effect illustrated on Figure 13B is applicable to the entire cohort. Pearson's r correlation value is -0.11 meaning a negative correlation.



Equation	y = a + b*x
Plot	ratio
Weight	No Weighting
Intercept	17.69012 ± 4.90738
Slope	-0.11037 ± 0.13861
Residual Sum of Squares	341.73685
Pearson's r	-0.18446
R-Square (COD)	0.03403
Adj. R-Square	-0.01964

Equation	y = a + b*x
Plot	ratio
Weight	No Weighting
Intercept	17.69012 ± 4.9073
Slope	-0.11037 ± 0.1386
Residual Sum of Squar	341.73685
Pearson's r	−0.18446
R-Square (COD)	0.03403
Adj. R-Square	−0.01964

Figure 13. cfDNA fragment size distribution based on Peneder *et al*. data. A) Method for calculating the ratios of mono- to dinucleosomes. B) Comparison of the mono-to-dinucleosome ratio between a younger (24 years old) and an older individual (40 years old). C) Scatter plot of mono-to-dinucleosome ratios for the 22 people from this study. D) Pearson's correlation coefficient r = -0.11 based on the scatter plot from panel (C).

3.5.5. NRL in Peneder et al. Then I calculated NRL for each patient to see if inter-nucleosome distance changes with age. Table 6 shows a summary of the NRL values calculated for each patient with error and age in years.

Table 6. Summary of NRL values calculated for each patient, error and age in years

Sample ID	Age, years	NRL, bp	Error, bp
Ctrl_7	24	192.1	0.2
Ctrl_6	25	194	1
Ctrl_16	25	192.1	0.7
Ctrl_12	27	191	0.4
Ctrl_14	27	195.3	0.7
Ctrl_10	29	194.4	0.8
Ctrl_13	30	193.2	0.3
Ctrl_20	30	192.1	0.4
Ctrl_11	31	195.1	1
Ctrl_22	31	192	0.8
Ctrl_8	32	193	1
Ctrl_21	32	192	0.6
Ctrl_18	34	191.15	0.6
Ctrl_15	36	194.1	0.8
Ctrl_1	37	194.05	0.4
Ctrl_9	37	192.5	0.4
Ctrl_17	39	192.9	0.6
Ctrl_4	40	195.1	0.8
Ctrl_19	40	190.6	0.6
Ctrl_2	44	195.1	1.1
Ctrl_3	48	194.1	0.7
Ctrl_5	50	194.8	1.1

In order to visualise and assess these data effectively, I plotted the NRL values and calculated a Pearson correlation value (Figure 14). There is some significant correlation shown by r =0.32.

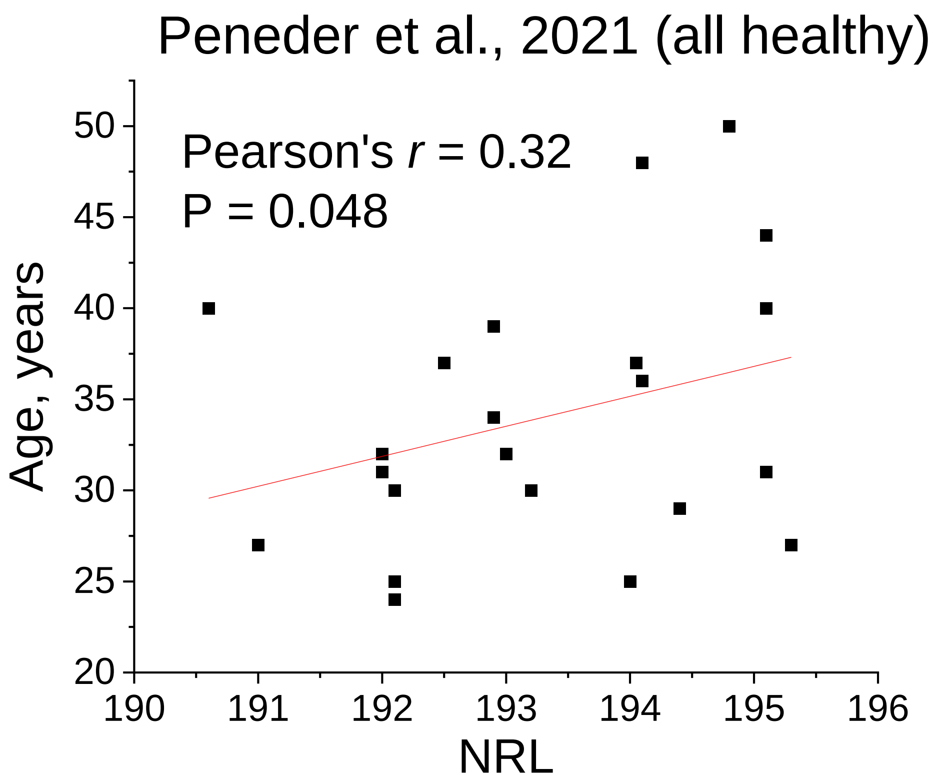


Figure 14. A scatter plot of NRL values for 22 people from the Peneder *et al*. study, showing NRL as a function of age. The Pearson correlation coefficient r = 0.32 and the corresponding linear fit is shown on the figure.

3.6. Nucleosome positioning in B-cells around regions of ageing-sensitive changes of DNA methylation. Several studies investigated the link between ageing and DNA methylation as discussed in Literature review. Here I used the dataset of genomic regions that changed DNA methylation with ageing provided by the group of Prof Leonard Schalkwyk and compared the nucleosome profiles in healthy B-cells and cancer cells from CLL around these ageing-associated differentially methylated regions to see if there is a distinction.

I calculated nucleosome positioning around CpG sites associated with ageing in healthy B-cells. The results (Figure 15A and B) showed reduced occupancy at CpG sites that are positively associated with ageing and increased occupancy at sites negatively associated with ageing.

I also compared nucleosome occupancy around CpG sites associated with ageing in two types of leukemic blood cells from chronic lymphocytic leukaemia, M-CLL in IGHV-mutated CLL subtype and U-CLL in IGHV-unmutated CLL subtype. The results (Figure 15C and D) show the same effect as with healthy B-cells to a different extent – M-CLL features highest occupancy of the three in both cases, whereas U-CLL shows the lowest and healthy B-cells are in the middle.

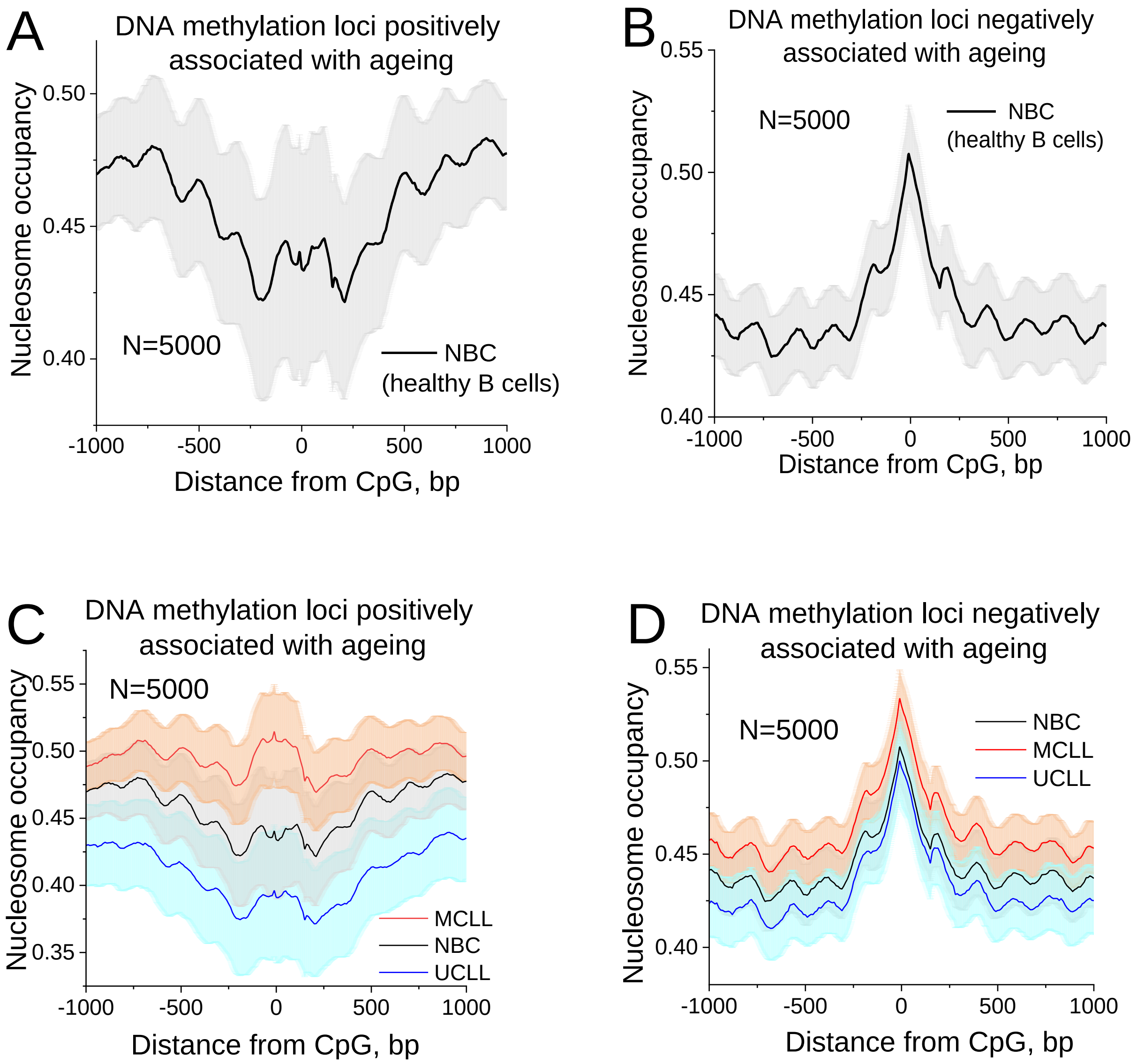


Figure 15. Nucleosome occupancy profiles around ageing-sensitive CpGs in B-cells from peripheral blood of healthy people (NBC), and patients with IGHV-mutated CLL (M-CLL) and IGHV-unmutated CLL (U-CLL). A) CpGs positively associated with ageing in healthy B-cells. B) CpGs negatively associated with ageing in healthy B-cells. C and D) The comparison of nucleosome occupancy in positively ageing-associated loci (C) and negatively associated loci (D).

4. Discussion

This study explored the potential use of cfDNA-inferred nucleosome positioning as a diagnostic marker using whole-genome cfDNA sequencing data from multiple studies. Our results showed pronounced differences in chromatin structure in cancer versus normal samples and between different age groups. On the methodological side, we used two main different approaches for sample classification: one based on the changes of nucleosome occupancy in cancer-sensitive regions, and another based on the changes in spacing between nucleosomes. For the latter, we observed a global cancer-specific decrease in in the nucleosome repeat length (NRL), which was pronounced in repetitive genomic regions that were included in this analysis.

***4.1. cfDNA sample classification based on cancer-specific nucleosome loss/gain regions*.** We established a procedure to distinguish samples using principal component analysis of cfDNA occupancy in cancer-sensitive regions.
First of all, I applied our stratification method based on regions that lost or gained nucleosome occupancy genome-wide to cfDNA data from breast cancer patients and healthy volunteers reported in the study of Snyder *et al*. (2016). PCA based on both "lost-nucleosome" and "gained-nucleosome" regions showed distinct clustering between healthy people and two different cancer subtypes (Figures 2-4). This means that nucleosome positioning as inferred from cfDNA is able to distinguish breast cancer patients from healthy individuals based on nucleosome positioning genome-wide.
Then I applied this method to small cell lung cancer (SCLC) data from our collaborators (Takahashi *et al*., 2022), which is a longitudinal study with three time points (pre-treatment, post-treatment and disease progression). This dataset also

contains information on patient sensitivity to treatment, which I used in my analysis. In both tasks of distinguishing time points and distinguishing treatment sensitivity, clear differential clustering in SCLC was achieved and in both cases the regions that lost nucleosome occupancy performed better than the regions that gained occupancy. This effect is not explained by greater number of regions that lost nucleosome occupancy. In addition, the elimination of outliers in both cases of regions that lost occupancy was more effective than with the regions that gained occupancy. Also, when applying our method to SCLC data classification according to treatment stages, the disease progression stage was not included in the training model set but clustered clearly between the pre-treatment and post-treatment clusters. Therefore, this method might be able to predict SCLC progression on data that was not included in model training.

***4.2. cfDNA sample classification based on the changes in spacing between nucleosomes in cancer*.** Then we asked whether NRL changes may be the factor behind our observations of nucleosome positioning changes. NRL is known to change during cell differentiation and has been studied in that context for some time (Singh and Mueller-Planitz, 2021). Its relationship with cancer has been suggested since 1976 (Compton, Bellard and Chambon, 1976) and had been addressed again in recent publications (Willcockson *et al*., 2021; Yusufova *et al*., 2021). In our project (Jacob *et al*., 2022) we performed the first study to assess nucleosome positioning maps in tumour and healthy tissue, and cfDNA from the same patients. Breast tissue samples for healthy individuals are normally not available unless they are undergoing some cosmetic breast surgery, in which case the tissues may contain different types of breast cells which would not be comparable with those that we are investigating, therefore the acquisition of such samples was not easily attainable for this project. Using this dataset, we observed genome-wide shortening of NRL, which is distinct from the local effects at regulatory regions at TF binding sites (van der Pol and Mouliere, 2019; Teif *et al*., 2017; Clarkson *et al*., 2019) or gene promoters (Snyder *et al*., 2016; Valouev *et al*., 2011) which face changes in occupancy related to transcription. This effect is also different from the effect of the shift to shorter cfDNA fragments in cancer compared to healthy controls (Underhill *et al*., 2016). Plasma cfDNA derived from foetal tissues was also distinguished as shorter compared to maternal cfDNA, which was the implication for non-invasive prenatal testing (Chan *et al*., 2004). The sequencing of the shorter fragments specifically for the purposes of detecting ctDNA was suggested by Mouliere *et al*. (2018). Interestingly, my results show that cfDNA of older people is enriched with shorter cfDNA fragments. Therefore, patient's age is an important factor to take into account in cfDNA-based diagnostics. Some studies suggest that repetitive genomic regions in cfDNA are overrepresented (Grabuschnig *et al*., 2020; Bronkhorst *et al*., 2018), which is why we assessed Alu and LINE1 repeats. Indeed, we found that NRL changes in repetitive regions are particularly pronounced.

***4.3. Implications of the NRL analysis*.** NRL shortening in cancer has very promising potential implications in the clinic, for example in the development of cfDNA nucleosomics-based liquid biopsy that offers greater sensitivity and cost-effectiveness than the techniques that are currently in use. It could be that the determination of the cells of origin may be completely scrapped from patient classification, which makes this method a lot more straightforward and useful in clinical practice.

It is interesting that differences at the level of individual nucleosomes may derive from large-scale changes, for example the shrinking of breast cancer cells or their nuclei (Han *et al*., 2020) and also fluctuations in chromatin flexibility which is known to decrease with disease progression (Stowers *et al*., 2019). It must also be mentioned that tumour cells appear to have features similar to those of stem cells, such as shorter NRL compared to differentiated cells (Teif *et al*., 2012). This may have an association with cancer cell de-differentiation (Friedmann-Morvinski and Verma, 2014). From the mechanics point of view, decreased NRL is known to alter the topology of nucleosome fibers to more deformable and prone for macroscopic self-association (Zhurkin and Norouzi, 2021). Therefore, the shortening of NRL in breast cancer might be related to large-scale chromatin reshuffling but the underlying causes for this are not clear. It is worth noting that the NRL approach appears to be resilient in relation to the extent of MNase digestion.

***4.4. cfDNA analysis in the context of ageing*.** Next, we hoped to investigate whether differences in nucleosome occupancy as observed in cfDNA throughout the ageing process can provide the basis for an “ageing clock” that would be able to predict the age of an individual similarly to the DNA methylation age clock (Horvath, 2013).

4.4.1. *cfDNA analysis based on regions of ageing-specific nucleosome occupancy loss/gain*. This method allowed me to distinguish different age groups based on the datasets of Teo *et al*. (2016) and Peneder *et al*. (2021). In the former study, I performed pairwise comparison of cfDNA nucleosome positioning landscapes in 25-year-olds and centenarians to identify 100-bp regions throughout the genome with drastically changed cfDNA occupancy between the two age groups. Then these regions were used for patient stratification with principal component analysis (PCA). It is important to note that the 70-year-old group was not used when defining regions which lost or gained occupancy however it is clustered distinctly from 25- and 100-year-olds along the y-axis with principle component 1 which represents age, closer to the 100 year old group. Both healthy and unhealthy centenarians are clustered together, meaning that chronological age as opposed to medical condition are the defining factor in this method.

4.4.2. Improving sample size. Teo *et al*. (2019) presented data from 12 individuals, so I expanded this analysis to include another 22 healthy individuals from a study of Peneder *et al*. (2021). This analysis allowed a distinction between three age groups: <30, 31-39 and >40 years old. Importantly, the 31–39-year-old group was not included in the model but clustered clearly separately and in between the other groups, especially in the analysis based on nucleosome-gain regions. Then I decided to explore the biological processes associated with these regions, so I extracted gene names by intersecting the coordinates of the regions of nucleosome loss or gain with promoters. It is interesting that the 5710 regions that gained occupancy gave 34 gene names which is few compared to 60 gene names given by 367 regions that lost occupancy. It could be that in the case of regions that gained occupancy, the non-coding regions have a greater impact on the effect that we observe.

4.4.3. The biology of ageing as suggested by nucleosome occupancy changes. The regions that gained occupancy appear to be associated with immunological processes. This included in particular the response to viral and/or respiratory infections, which is interesting given that senescence is associated with immune dysregulation, including the involvement of B cells (LeMaoult *et al*., 2006). The JAK pathway, which also specifically appeared in my earlier analysis, is known to be more active in ageing cells compared to non-senescent cells as it reduces senescence-related tissue dysfunction (Xu *et al*., 2016). This has the opposite effect to another relevant term – microRNAs are thought to inhibit cell proliferation and encourage cell senescence (Liu *et al*., 2014; Saeidimehr *et al*., 2016). The regions that lost occupancy are predominantly associated with the GAGE family, which are linked to melanoma and the function of healthy testes and are located on chromosome X. Since a few of the other terms are relevant to the testes and none seem to be directly relevant to melanoma, this may be relevant to the age-related decline in some the functions of the testes.

Koohy *et al*. (2018) profiled genome-wide chromatin characteristics and gene expression in B cells in mice. They found that most genes show a generally preserved level of expression with some changes at microRNA encoding genes. In particular, they observed downregulation of genes related to the insulin-like growth factor signalling pathway (Irs1) while Let-7 microRNA expression was upregulated. Our results allow to make different observations since Irs1 and Let-7 did not appear in our gene list with the parameters that we used. There is one similarity that some of the genes implicated in both studies come from the same families - those are Rab (Ras-related protein family from the RAS oncogene superfamily), genes that encode components of the TRAPP (transport protein particle) multi-subunit complex and spermatogenesis-associated (SPATA) gene family. It is interesting to see Spata6 identified as differentially expressed in pro-B cells in nuclear RNA-seq in the context of ageing given that a significant proportion of gene names in regions that lost occupancy that we have identified are associated with testicular function, and to see that in both studies the SPATA genes are downregulated. This may suggest general loss of fertility in males, which is expected (Zitzmann, 2013).

There is some consensus with genes located in the regions which gained occupancy and individual genes identified as upregulated by Koohy *et al*. - those genes are ELF1, PLCB4 and MCTP1.

- ELF1 (E74 Like ETS Transcription Factor 1) is normally expressed quite highly in B-cells and some other haematopoietic lineages, it encodes an E26 transformation-specific transcription factor and is associated with retinoblastoma and intramuscular haemangioma. It plays a role in the regulation of gene expression and IL-2 activation and signalling.
- PLCB4 (Phospholipase C Beta 4) plays an important role in the function of the retina and the heart, and sweet taste signalling.
- MCTP1 (Multiple C2 And Transmembrane Domain Containing 1) are a family of transmembrane proteins that bind Ca^{2+} but not phospholipids, therefore acting as a unique calcium sensor which stabilises release of neurotransmitters, induces and maintains presynaptic homeostatic plasticity (Shin *et al*., 2005).

In addition, quite a high proportion of the genes found within regions that increased nucleosome occupancy are miRNA genes - Koohy *et al*. found different sets of miRNA genes both upregulated and downregulated, although the exact gene names listed were not the same as the ones we observed in regions that gained occupancy.

One of the reasons for the difference of my observations from those of Koohy *et al*. is that plasma cfDNA is composed of a mixture of DNA shed by various cells of different origin (Snyder *et al*., 2016), the largest proportion of which is of haematopoietic descent. It is possible that in these results we are presented with a different scope of general ageing represented by a mixture of tissues and cell types as opposed to a single tissue or cell type. For example, it is not surprising that none of genes from the GAGE family, which were prevalent in our results, were detected by Koohy *et al*. since they did not study cells of the testes or melanoma, although it is interesting that some members of the SPATA family were detected by RNA-seq in B cells. It is possible that the downregulation of Irs1, upregulation of Let-7, differential regulation of genes coding for components of IGF/mTOR signalling and other genes implicated by Koohy *et al*. would be observed if the signal was not overpowered by other differentially regulated genes from other tissues that may be more prominent in our analysis. These genes could potentially be detected if the stringency of the thresholds when defining regions that lost or gained occupancy was reduced. This shows that cfDNA is a useful biomarker for studying the general effects of ageing as opposed to sampling a single tissue.

4.5. cfDNA analysis based on the changes in spacing between nucleosomes in ageing. Then I applied the NRL method to the ageing data from Peneder *et al*. (2021). Thankfully the datasets had sufficient sequencing coverage for a reasonable prediction with standard errors 1.1bp or less. We observed a rather small but significant (with Pearson's r =0.32) increase in NRL from ages 24 to 50. It is possible that this pattern might change in a different age group – it might be insightful to calculate NRL values for a large cohort of elderly individuals, such as centenarians as in Teo *et al*. (2019) but with larger sample size.

4.6. cfDNA fragment size distribution in ageing**.** We hypothesised that cfDNA fragment sizes might differ throughout the ageing process and possibly explain the differences in nucleosome positioning, so we applied a simple measure of the ratio of mono- to dinucleosomal fragments to see if there is a relationship between the level of nuclease digestion and age. The results show a decrease in the ratio of mono- to dinucleosomal fragments with age, so the distribution of cfDNA fragment sizes could potentially serve as a predictor of biological age. It is possible that this effect is observed due to the slight over-representation of the younger group in the cohort – there are 8 individuals of ages 30 and under, 9 aged 31-39 and 5 aged 40 and over. However, if this effect is disproved by replication with a larger cohort, the reduction in cfDNA fragment size in the older groups may represent a decrease in the length of nucleosome-protected DNA in older individuals as nucleosomes become more susceptible to digestion by apoptotic nucleases, while the increase of NRL may be interpreted as a dramatic loss of nucleosomes with age. These results are potentially of great importance in the future of nucleosomics-based liquid biopsy which is currently being incorporated into clinical practice and which does not consider patient age yet. This method is promising as a tool for stratification of age groups which might help the implementation of cfDNA-based diagnostics for cancer and other conditions in the clinic.

4.7. Assessing the effect of DNA methylation and nucleosome repositioning**.** Then we assessed the interplay between nucleosome positioning in cancer, age and

DNA methylation. The effect of DNA methylation on nucleosome stability and nucleosome positioning landscapes has been studied before (Teif *et al*., 2014). Most of the CpG dinucleotides in regions that lost nucleosome occupancy tend to be unmethylated and in comparison, a lot of the CpGs in regions that gained nucleosome occupancy tend to be methylated (Teif *et al*., 2014; Wiehle *et al*., 2019). The loss of nucleosome occupancy in largely unmethylated CpGs and a gain of nucleosome occupancy in mouse embryonic stem cells was attributed to CTCF binding sites that are cell type-specific (Teif *et al*., 2014). To assess whether DNA methylation might have a role in this effect, I calculated nucleosome occupancy profiles around CpG sites associated with ageing in healthy B-cells. The results showed reduced occupancy at CpG sites that are positively associated with ageing and increased occupancy at sites negatively associated with ageing. In addition, I compared nucleosome occupancy around CpG sites associated with ageing in two types of leukemic blood cells from patients with two subtypes of chronic lymphocytic leukaemia, mCLL and uCLL. As a result, I obtained qualitatively similar nucleosome occupancy profiles.

***4.8. Conclusion and future directions*.** Overall, this study contributed towards the development of diagnostic methods based on nucleosome positioning. I assessed nucleosome positioning in cancer and ageing and it appears that older people have longer NRL and shorter cfDNA fragments, while cancer patients have shorter NRL and shorter cfDNA fragments. Overall, these findings prove that methods in cfDNA nucleosomics can be used to stratify patients into better informed clinically relevant categories and improve diagnostic approaches in cancer, management and monitoring by liquid biopsy. This method is applicable to other cancer types beyond breast cancer and lung cancer investigated here as main examples and has a potential for pan-cancer identification. More research and larger cohorts with more cancer types and subtypes with longitudinal metadata are needed to uncover the full potential of this diagnostic method. The connection between nucleosome positioning in ageing and cancer is poorly understood and is a barrier to the development of better diagnostic methods for cancer based on cfDNA nucleosomics, which are potentially very promising for clinical use.

5. Bibliography.

Berkowitz, E. M., Sanborn, A. C., & Vaughan, D. W. (1983). Chromatin structure in neuronal and neuroglial cell nuclei as a function of age. *J Neurochem,* ***41***(2), 516-523. doi:10.1111/j.1471-4159.1983.tb04769.

Bezdan D, Grigorev K, Meydan C et al. (2020). Cell-free DNA (cfDNA) and Exosome Profiling from a Year-Long Human Spaceflight Reveals Circulating Biomarkers, *iScience*, **23**:101844.

Bronkhorst, A. J., Wentzel, J. F., Ungerer, V., Peters, D. L., Aucamp, J., de Villiers, E. P., Holdenrieder, S. and Pretorius, P. J. (2018) Sequence analysis of cell-free DNA derived from cultured human bone osteosarcoma (143B) cells. *Tumor Biology*, **40**, 1010428318801190.

Capri M, Moreno-Villanueva M, Cevenini E, Pini E, Scurti M, Borelli V, Palmas MG, Zoli M, Schön C, Siepelmeyer A, Bernhardt J, Fiegl S, Zondag G, de Craen AJ, Hervonen A, Hurme M, Sikora E, Gonos ES, Voutetakis K, Toussaint O, Debacq-Chainiaux F, Grubeck-Loebenstein B, Bürkle A, Franceschi C. (2015). MARK-AGE population: From the human model to new insights. *Mechanisms of Ageing and Development*, **151**, 13– 17. doi:10.1016/j.mad.2015.03.010

Castagne, R., Gares, V., Karimi, M., Chadeau-Hyam, M., Vineis, P., Delpierre, C., … Lifepath, C. (2018). Allostatic load and subsequent all-cause mortality: Which biological markers drive the relationship? Findings from a UK birth cohort. *European Journal of Epidemiology*, **33**(5), 441– 458. doi:10.1007/s10654-018-0364-1

Chan, K. A., Zhang, J., Hui, A. B., Wong, N., Lau, T. K., Leung, T. N., Lo, K.-W., Huang, D. W. and Lo, Y. D. (2004) Size distributions of maternal and fetal DNA in maternal plasma. *Clinical chemistry*, **50**, 88-92.

Chandrananda D, Thorne NP, Bahlo M. (2015). High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA, *BMC Med Genomics*. **8**:29.

Che, H., Jatsenko, T., Lenaerts, L. *et al.* Pan-Cancer Detection and Typing by Mining Patterns in Large Genome-Wide Cell-Free DNA Sequencing Datasets, *Clinical Chemistry*, **68**:9, 1164–1176, doi:10.1093/clinchem/hvac095

Che, H., Jatsenko, T., Lannoo, L. *et al.* Machine learning-based detection of immune-mediated diseases from genome-wide cell-free DNA sequencing datasets. *NPJ Genom. Med.* **7**, 55 (2022b). doi:10.1038/s41525-022-00325-w

Chen, B. H., Marioni, R. E., Colicino, E., Peters, M. J., Ward-Caviness, C. K., Tsai, P. C., … Horvath, S. (2016). DNA methylation-based measures of biological age: Meta-analysis predicting time to death. *Aging*, **8**(9), 1844– 1865. doi:10.18632/aging.101020

Chen, Y., Bravo, J. I., Son, J. M., Lee, C., & Benayoun, B. A. (2020). Remodeling of the H3 nucleosomal landscape during mouse aging. *Transl Med Aging,* ***4***, 22-31. doi:10.1016/j.tma.2019.12.003

Clark, S. C., Chereji, R. V., Lee, P. R., Fields, R. D., & Clark, D. J. (2020). Differential nucleosome spacing in neurons and glia. *Neurosci Lett, **714***, 134559. doi:10.1016/j.neulet.2019.134559

Clarkson C.T., Deeks E.A., Samarista R., Mamayusupova H., Zhurkin V.B., Teif V.B. (2019) CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. *Nucleic Acids Res* **47**, 11181-11196.

Compton JL, Bellard M, Chambon P. (1976). Biochemical evidence of variability in the DNA repeat length in the chromatin of higher eukaryotes. *Proc Natl Acad Sci U S A* **73**, 4382-4386.

Diehl, F., Schmidt, K., Choti, M. *et al.* (2008). Circulating mutant DNA to assess tumor dynamics. *Nat Med* **14**, 985–990. doi:10.1038/nm.1789

Feser, J., Truong, D., Das, C., Carson, J. J., Kieft, J., Harkness, T., & Tyler, J. K. (2010). Elevated histone expression promotes life span extension. *Mol Cell, **39***(5), 724-735. doi:10.1016/j.molcel.2010.08.015

Franceschi, C., & Campisi, J. (2014). Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, **69**(Suppl 1), S4– 9. doi:10.1093/gerona/glu057

Franceschi, C., Salvioli, S., Garagnani, P., de Eguileor, M., Monti, D., & Capri, *M. (2017). Immunobiography and the heterogeneity of immune responses in the elderly: A* focus on inflammaging and trained immunity. *Frontiers in Immunology*, **8**, 982. doi:10.3389/fimmu.2017.00982

Frenel JS, Carreira S, Goodall J et al. (2015). Serial Next-Generation Sequencing of Circulating Cell-Free DNA Evaluating Tumor Clone Response To Molecularly Targeted Drug Administration, *Clin Cancer Res*, **21**:4586-4596.

Friedmann-Morvinski D, Verma IM. (2014). Dedifferentiation and reprogramming: origins of cancer stem cells. *EMBO Rep* **15**, 244-253.

Gaubatz, J., Ellis, M., & Chalkley, R. (1979). Nuclease digestion studies of mouse chromatin as a function of age. *J Gerontol, **34***(5), 672-679. doi:10.1093/geronj/34.5.672

Grabuschnig, S., Soh, J., Heidinger, P., Bachler, T., Hirschböck, E., Rodriguez, I. R., Schwendenwein, D. and Sensen, C. W. (2020). Circulating cell-free DNA is predominantly composed of retrotransposable elements and non-telomeric satellite DNA. *Journal of biotechnology*, **313**, 48-56.

Han DSC, Ni M, Chan RWY et al. (2020). The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB, *Am J Hum Genet*, **106**:202-214.

Heinz S, Benner C, Spann N et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, *Mol Cell,* **38**:576-589.

Heitzer E, Auinger L, Speicher MR. (2020). Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living, *Trends Mol Med*, **26**:519-528.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, **14**(10), R115. doi:10.1186/gb-2013-14-10-r115

Hu, Z., Chen, K., Xia, Z., Chavez, M., Pal, S., Seol, J. H., . . . Tyler, J. K. (2014). Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. *Genes Dev, **28***(4), 396-408. doi:10.1101/gad.233221.113

Ignatiadis M, Sledge GW, Jeffrey SS. (2021). Liquid biopsy enters the clinic - implementation issues and future challenges, *Nat Rev Clin Oncol*. **18**(5):297-312. doi: 10.1038/s41571-020-00457-x.

Im YR, Tsui DWY, Diaz LA Jr, Wan JCM. (2021). Next-Generation Liquid Biopsies: Embracing Data Science in Oncology. *Trends Cancer*. **7**(4):283-292. doi: 10.1016/j.trecan.2020.11.001.

Ishimi, Y., Kojima, M., Takeuchi, F., Miyamoto, T., Yamada, M., & Hanaoka, F. (1987). Changes in chromatin structure during aging of human skin fibroblasts. *Exp Cell Res, **169***(2), 458-467. doi:10.1016/0014-4827(87)90206-0

Jacob D. R., Guiblet W. M., Mamayusupova H., Shtumpf M., Gretton S., Correa C., Dellow E., Ruje L., Ciuta I., Agrawal S. P., Shafiei N., Vainshtein Y., Clarkson C. T., Thorn G. J., Sohn K, Pradeepa M. M., Chandrasekharan S., Klenova E., Zhurkin V. B. and Teif V. B., (2022) Dramatic nucleosome reorganisation in breast cancer tissues. (Under review).

Jylhava, J., Kotipelto, T., Raitala, A., Jylha, M., Hervonen, A., & Hurme, M. (2011). Aging is associated with quantitative and qualitative changes in circulating cell-free DNA: The vitality 90+ study. *Mechanisms of Ageing and Development*, **132**(1–2), 20– 26. doi:10.1016/j.mad.2010.11.001

Jylhava, J., Nevalainen, T., Marttila, S., Jylha, M., Hervonen, A., & Hurme, M. (2013). Characterization of the role of distinct plasma cell-free DNA species in age-associated inflammation and frailty. *Aging Cell*, **12**(3), 388– 397. doi:10.1111/acel.12058

Kitzman JO, Snyder MW, Ventura M et al. (2012). Noninvasive whole-genome sequencing of a human fetus, *Sci Transl Med*, **6**;4(137):137ra76. doi: 10.1126/scitranslmed.3004323.

Koohy H, Bolland DJ, Matheson LS, Schoenfelder S, Stellato C, Dimond A, Várnai C, Chovanec P, Chessa T, Denizot J, Manzano Garcia R, Wingett SW, Freire-Pritchett P, Nagano T, Hawkins P, Stephens L, Elderkin S, Spivakov M, Fraser P, Corcoran AE, Varga-Weisz PD. (2018). Genome organization and chromatin analysis identify transcriptional downregulation of insulin-like growth factor signaling as a hallmark of aging in developing B cells. *Genome Biol*. **5**;19(1):126. doi: 10.1186/s13059-018-1489-y.

Kustanovich A, Schwartz R, Peretz T et al. (2019). Life and death of circulating cell-free DNA, *Cancer Biol Ther*, **20**:1057-1067.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, 1-10.

Lee, T., Rawding, P. A., Bu, J., Hyun, S., Rou, W., Jeon, H., Kim, S., Lee, B., Kubiatowicz, L. J., Kim, D., Hong, S., Eun, H. (2022). Machine-Learning-Based Clinical Biomarker Using Cell-Free DNA for Hepatocellular Carcinoma (HCC). *Cancers*, **14**(9):2061. doi:10.3390/cancers14092061

LeMaoult J, Szabo P, Weksler ME. (1997). Effect of age on humoral immunity, selection of the B-cell repertoire and B-cell development. *Immunol Rev*. **160**:115-26. doi: 10.1111/j.1600-065x.1997.tb01032.x.

Liu MC, Oxnard GR, Klein EA et al. (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA, *Annals of Oncology*, **31**:745-759.

Liu G, Sun Y, Ji P, Li X, Cogdell D, Yang D, et al. (2014). MiR506 suppresses proliferation and induces senescence by directly targeting the CDK4/6-FOXM1 axis in ovarian cancer. *J Pathol*, **233**(3):308–318.

Lu, R. J., Taylor, S., Contrepois, K., Kim, M., Bravo, J. I., Ellenberger, M., . . . Benayoun, B. A. (2021). Multi-omic profiling of primary mouse neutrophils predicts a pattern of sex and age-related functional regulation. *Nat Aging, **1***(8), 715-733. doi:10.1038/s43587-021-00086-8

Mallm J.-P., Iskar M., Ishaque N., Klett L.C., Kugler S.J., Muino J.M., Teif V.B., Poos A.M., Großmann S., Erdel F., Tavernari D., Koser S.D., Schumacher S., Brors B., König R., Remondini D., Vingron M., Stilgenbauer S., Lichter P., Zapatka M., Mertens D., Rippe K. (2019) Linking aberrant chromatin features in chronic lymphocytic leukemia to transcription factor networks. *Mol Syst Biol* **15**, e8339

Mandel P, Metais P. (1948). Les acides nucleiques du plasma sanguine chez l'homme, *C R Seances Soc Biol Fil*, **142**:241-243.

Mouliere F, Mair R, Chandrananda D et al. (2018). Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. *EMBO Mol Med*, **10**.

Mouliere F, Chandrananda D, Piskorz AM et al. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*, **10**.

Miura, Y., & Endo, T. (2016). Glycomics and glycoproteomics focused on aging and age-related diseases – Glycans as a potential biomarker for physiological alterations. *Biochimica et Biophysica Acta*, **1860**(8), 1608– 1614. doi:10.1016/j.bbagen.2016.01.013

Ng H, Havervall S, Rosell A, Aguilera K, Parv K, von Meijenfeldt FA, Lisman T, Mackman N, Thålin C, Phillipson M. (2021). Circulating Markers of Neutrophil Extracellular Traps Are of Prognostic Value in Patients With COVID-19. *Arterioscler Thromb Vasc Biol*. **41**(2):988-994. doi: 10.1161/ATVBAHA.120.315267.

Neocleous, A. C., Nicolaides, K. H. and Schizas C. N. (2016). First Trimester Noninvasive Prenatal Diagnosis: A Computational Intelligence Approach. *IEEE Journal of Biomedical and Health Informatics*, **20**:5, 1427-1438. doi: 10.1109/JBHI.2015.2462744.

Pegoraro, G., Kubben, N., Wickert, U., Gohler, H., Hoffmann, K., & Misteli, T. (2009). Ageing-related chromatin defects through loss of the NURD complex. *Nat Cell Biol, **11***(10), 1261-1267. doi:10.1038/ncb1971

Peneder P, Stütz AM, Surdez D, Krumbholz M, Semper S, Chicard M, Sheffield NC, Pierron G, Lapouble E, Tötzl M, Ergüner B, Barreca D, Rendeiro AF, Agaimy A, Boztug H, Engstler G, Dworzak M, Bernkopf M, Taschner-Mandl S, Ambros IM, Myklebost O, Marec-Bérard P, Burchill SA, Brennan B, Strauss SJ, Whelan J, Schleiermacher G, Schaefer C, Dirksen U, Hutter C, Boye K, Ambros PF, Delattre O, Metzler M, Bock C, Tomazou EM. (2021). Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nat Commun*, **12**(1):3230. doi: 10.1038/s41467-021-23445-w.

Piroeva, K. V., McDonald, C., Xanthopoulos, C., Fox, C., Clarkson, C. T., Mallm, J.-P., Vainshtein, Y., Ruje, L., Klett, L. C., Stilgenbauer, S., Mertens, D., Kostareli, E., Rippe, K. and Teif, V. B. (2022). Nucleosome repositioning in chronic lymphocytic leukaemia. (Under review).

Saeidimehr S, Ebrahimi A, Saki N, Goodarzi P, Rahim F. (2016). MicroRNA-Based Linkage between Aging and Cancer: from Epigenetics View Point. *Cell J*, **18**(2):117-26. doi: 10.22074/cellj.2016.4303.

Sen, P., Shah, P., Nativio, R., Berger, S. L. (2016). Epigenetic Mechanisms of Longevity and Aging. *Cell*, **166**(4):822-839. doi:10.1016/j.cell.2016.07.050

Serpas L, Chan RWY, Jiang P et al. (2019). Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc Natl Acad Sci U S A*, **116**:641-649.

Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. (2021). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (update). *Nucleic Acids Res*, **23**;50(1):216–21. doi: 10.1093/nar/gkac194.

Shin OH, Han W, Wang Y, Südhof TC. (2005). Evolutionarily conserved multiple C2 domain proteins with two transmembrane regions (MCTPs) and unusual Ca2+ binding properties. *J Biol Chem*. **14**;280(2):1641-51. doi: 10.1074/jbc.M407305200.

Shu, Y., Wu, X., Tong, X. *et al.* (2017). Circulating Tumor DNA Mutation Profiling by Targeted Next Generation Sequencing Provides Guidance for Personalized Treatments in Multiple Cancer Types. *Sci Rep* **7**, 583. doi:10.1038/s41598-017-00520-1

Singh, A. K. and Mueller-Planitz, F. (2021) Nucleosome positioning and spacing: from mechanism to function. *Journal of Molecular Biology*, **433**, 166847.

Song C-X, Yin S, Ma L et al. (2017). 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Research*, **27**:1231-1242.

Smith, T. A., Vallis, Y., Neary, D., & Itzhaki, R. F. (1989). Characteristics of lymphocyte chromatin from Alzheimer's disease patients and from young and old normal individuals. *Gerontology, **35***(5-6), 268-274. doi:10.1159/00021303

Snyder MW, Kircher M, Hill AJ et al. (2016). Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, **164**:57-68.

Stowers RS*, et al.* (2019). Matrix stiffness induces a tumorigenic phenotype in mammary epithelium through changes in chromatin accessibility. *Nat Biomed Eng* **3**, 1009-1019.

Takahashi, N., Pongor, L., Agrawal, S. P., Shtumpf, M., Rajapakse, V. N., Shafiei, A., Schultz, C., Kim, S. H., Roame, D., Carter, P., Zhang, Y., Vilimas, R., Nichols, S., Desai, P., Figg, D., Bagheri, M., Teif, V. B., Thomas, A. (2022) Concordance of somatic genome and inferred gene expression between circulating tumor DNA and matched small cell lung cancer. (Under review).

Teif V.B., Beshnova D.A., Marth C., Vainshtein Y., Mallm J.-P., Höfer T. and Rippe K. (2014). Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Research*. **24**, 1285-1295.

Teif V.B, Jan-Philipp Mallm, Tanvi Sharma, David B. Mark Welch, Karsten Rippe, Roland Eils, Jörg Langowski, Ada L. Olins & Donald E. Olins (2017) Nucleosome repositioning during differentiation of a human myeloid leukemia cell line. *Nucleus*, **8**:2, 188-204. doi: 10.1080/19491034.2017.1295201

Teif, V. B., Vainshtein, Y., Caudron-Herger, M., Mallm, J.-P., Marth, C., Höfer, T. and Rippe, K. (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology*, **19**, 1185-1192.

Teo YV, Capri M, Morsiani C et al. (2019). Cell-free DNA as a biomarker of aging. *Aging Cell*, **18**:e12890.

Quinlan, A. R. (2014) BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, **47**, 11.12. 1-11.12. 34.

Vainshtein, Y., Rippe, K. and Teif, V. B. (2017) NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. *BMC genomic*s, **18**, 1-13.

Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z. and Sidow, A. (2011) Determinants of nucleosome organization in primary human cells. *nature*, **474**, 516-520.

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**:350-368.

Volik S, Alcaide M, Morin RD et al. (2016). Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies. *Mol Cancer Res*, **14**:898-908.

Underhill HR, Kitzman JO, Hellwig S et al. (2016). Fragment Length of Circulating Tumor DNA. *PLoS Genet*, **12**:e1006162.

United Nations, D. o. E. a. S. A., Population Division (2017). World population prospects: The 2017 revision, key findings and advance tables. Working Paper No. ESA/P/WP/248.

Ungerer, V., Bronkhorst, A.J., Van den Ackerveken, P. *et al.* (2021).Serial profiling of cell-free DNA and nucleosome histone modifications in cell cultures. *Sci Rep* **11**, 9460 doi:10.1038/s41598-021-88866-5

Ulz P, Perakis S, Zhou Q et al. (2019). Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat Commun*, **10**:4666.

Wan N, Iinberg D, Liu TY et al. (2019). Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer*, **19**:832.

Wiehle L., Thorn G.J., Raddatz G., Clarkson C.T., Rippe K., Lyko F., Breiling A., Teif V.B. (2019). DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Research* **29**, 750-761.

Willcockson MA*, et al.* (2021). H1 histones control the epigenetic landscape by local chromatin compaction. *Nature* **589**, 293-298.

Williamson R. (1970). Properties of rapidly labelled deoxyribonucleic acid fragments isolated from the cytoplasm of primary cultures of embryonic mouse liver cells. *J Mol Biol*, **51**:157-168.

Wong SL, Demers M, Martinod K et al. (2015). Diabetes primes neutrophils to undergo NETosis, which impairs wound healing. *Nat Med*, **21**:815-819.

Xu M, Tchkonia T and Kirkland JL. (2016). Perspective: Targeting the JAK/STAT pathway to fight age-related dysfunction. *Pharmacol Res* **111**:152-154. doi: 10.1016/j.phrs.2016.05.015.

Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. (2019). GTRD: a database on gene transcription regulation-2019 update. *Nucleic Acids Res* **47**(1):100-105. doi: 10.1093/nar/gky1128.

Yusufova N*, et al.* (2021). Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. *Nature* **589**, 299-305.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9):137. doi: 10.1186/gb-2008-9-9-r137.

Zhurkin VB, Norouzi D. (2021). Topological polymorphism of nucleosome fibers and folding of chromatin. *Biophys J* **120**, 577-585.

Zitzmann M. (2013). Effects of age on male fertility. *Best Pract Res Clin Endocrinol Metab*. **27**(4):617-28. doi: 10.1016/j.beem.2013.07.004.

Zongza, V., & Mathias, A. P. (1979). The variation with age of the structure of chromatin in three cell types from rat liver. *Biochem J, **179***(2), 291-298. doi:10.1042/bj1790291.

Zviran A, Schulman RC, Shah M et al. (2020). Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med* **26**:1114-1124.