Chromosoma
https://doi.org/10.1007/s00412-021-00766-9

ORIGINAL ARTICLE

NucPosDB: a database of nucleosome positioning in vivo and nucleosomics of cell-free DNA

Mariya Shtumpf[1] · Kristan V. Piroeva[1] · Shivam P. Agrawal[1] · Divya R. Jacob[1] · Vladimir B. Teif[1]

Received: 24 November 2021 / Revised: 24 November 2021 / Accepted: 20 December 2021

Abstract

Nucleosome positioning is involved in many gene regulatory processes happening in the cell, and it may change as cells differentiate or respond to the changing microenvironment in a healthy or diseased organism. One important implication of nucleosome positioning in clinical epigenetics is its use in the "nucleosomics" analysis of cell-free DNA (cfDNA) for the purpose of patient diagnostics in liquid biopsies. The rationale for this is that the apoptotic nucleases that digest chromatin of the dying cells mostly cut DNA between nucleosomes. Thus, the short pieces of DNA in body fluids reflect the positions of nucleosomes in the cells of origin. Here, we report a systematic nucleosomics database — NucPosDB — curating published nucleosome positioning datasets in vivo as well as datasets of sequenced cell-free DNA (cfDNA) that reflect nucleosome positioning in situ in the cells of origin. Users can select subsets of the database by a number of criteria and then obtain raw or processed data. NucPosDB also reports the originally determined regions with stable nucleosome occupancy across several individuals with a given condition. An additional section provides a catalogue of computational tools for the analysis of nucleosome positioning or cfDNA experiments and theoretical algorithms for the prediction of nucleosome positioning preferences from DNA sequence. We provide an overview of the field, describe the structure of the database in this context, and demonstrate data variability using examples of different medical conditions. NucPosDB is useful both for the analysis of fundamental gene regulation processes and the training of computational models for patient diagnostics based on cfDNA. The database currently curates ~ 400 publications on nucleosome positioning in cell lines and in situ as well as cfDNA from > 10,000 patients and healthy volunteers. For open-access cfDNA datasets as well as key MNase-seq datasets in human cells, NucPosDB allows downloading processed mapped data in addition to the regions with stable nucleosome occupancy. NucPosDB is available at https://generegulation.org/nucposdb/.

Keywords cfDNA · Cell-free DNA · Liquid biopsy · Nucleosome positioning · Nucleosomics

Background

Genomic nucleosome positions are non-random and unique for each cell, reflecting many biological processes that require the access of regulatory molecules to the DNA (e.g. reviewed in Clarkson et al. 2019; Baldi et al. 2020; Parmar and Padinhateeri 2020)). Previously, we assembled a comprehensive collection of experimental datasets of nucleosome positioning across many organisms and cell lines as well as software tools for the analysis and prediction of nucleosome positioning (Teif 2016). After the initial focus on nucleosome positioning in organisms such as yeast (Yuan et al. 2005; Ioshikhes et al. 2006; Segal et al. 2006), many studies focused on human cells (Schones et al. 2008; Valouev et al. 2011; Gaffney et al. 2012; Kundaje et al. 2012; Diermeier et al. 2014; Ho et al. 2014; Teif et al. 2017; Mallm et al. 2019). Furthermore, more recently, the field has moved towards clinical applications of nucleosome positioning to cell-free DNA (cfDNA), as will be explained below. There is a strong need for an integrative database that connects both fundamental and clinically focused "nucleosomics". Here, we report a systematic database, called *NucPosDB*, which integrates classical nucleosome positioning studies with a new direction of nucleosome positioning landscapes reconstructed from cfDNA from human patients.

The shift of the focus of the research from fundamental roles of nucleosome positioning in gene regulation

✉ Vladimir B. Teif
vteif@essex.ac.uk

[1] School of Life Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

Published online: 21 January 2022

to patient diagnostics is happening due to the fact that nucleosome positioning can provide a valuable diagnostic marker offering unique features not available in other clinical tests. There are two main arguments for this. Firstly, the timescale of the change of nucleosome positioning landscape is comparable to the timing of gene activation or the cell cycle (Schones et al. 2008; Teif et al. 2012) which is between the quick changes of gene expression and concentrations of disease-related small molecules and the much slower changes reflected by DNA mutations or aberrant methylation happening in cancer (Dawson and Kouzarides 2012; Pich et al. 2018; Li and Luscombe 2020) (Fig. 1A). Thus, differences in nucleosome positioning can be in principle suitable for monitoring a patient's response to therapy in this intermediate time range. While very informative, determining genome-wide nucleosome positioning maps in tumour tissues of cancer patients would be an expensive and invasive procedure. Here, the second argument comes into play: luckily, nucleosome positioning in tissues is directly reflected in cfDNA circulating in blood and other body liquids. This is because nucleases, which shred the chromatin of dying cells to form what later becomes cfDNA, preferentially cut the DNA between nucleosomes (Chandrananda et al. 2015; Kustanovich et al. 2019; Serpas et al. 2019; Han et al. 2020; Heitzer et al. 2020) (Fig. 1B). Since the half-life of cfDNA in blood is about 15 min (Volik et al. 2016), cfDNA extracted at any given time point represents a very recent snapshot of nucleosome positioning in the cells of origin.

Medical tests based on cfDNA are sometimes called "liquid biopsies" because this promising approach allows avoiding tissue biopsy in the case of solid tumours (Volik et al. 2016; Wan et al. 2017; Peng et al. 2021; Ignatiadis et al. 2021; Lo et al. 2021). The history of cfDNA research can be traced back to 1944 when it was first reported (Mandel and Metais 1948). cfDNA source was correctly interpreted as the products of apoptotic cleavage of chromatin subunits as early as 1970 (Williamson 1970; Henikoff and Church 2018). However, the active use of cfDNA for medical purposes using next-generation sequencing (NGS) started only in the recent years (Ignatiadis et al. 2021) with many diverse applications ranging from prenatal testing (Kitzman et al. 2012; Sun et al. 2018), cancer (Frenel et al. 2015; Phallen et al. 2017; Cristiano et al. 2019; Zviran et al. 2020), ageing (Teo et al. 2019), inference of patterns of gene expression (Snyder et al. 2016; Ulz et al. 2016) and transcription factor binding (Ulz et al. 2019), to even monitoring astronaut's health on spaceflights (Bezdan et al. 2020). While the field of liquid biopsies is expanding dramatically, it is still in the search of methods balancing sensitivity and cost (Abbosh et al. 2017; Wan et al. 2019; Peng et al. 2021).

Historically, the first class of genomics-based cfDNA diagnostic methods relied on mutation analysis (Frenel et al. 2015; Abbosh et al. 2017; Dudley and Diehn 2021; Zviran et al. 2020). Related approaches involve analyses of gene fusions (Palande et al. 2020) or copy number variations (CNVs) (Mouliere et al. 2018b). In all these cases, assay sensitivity critically depends on the sequencing depth as well

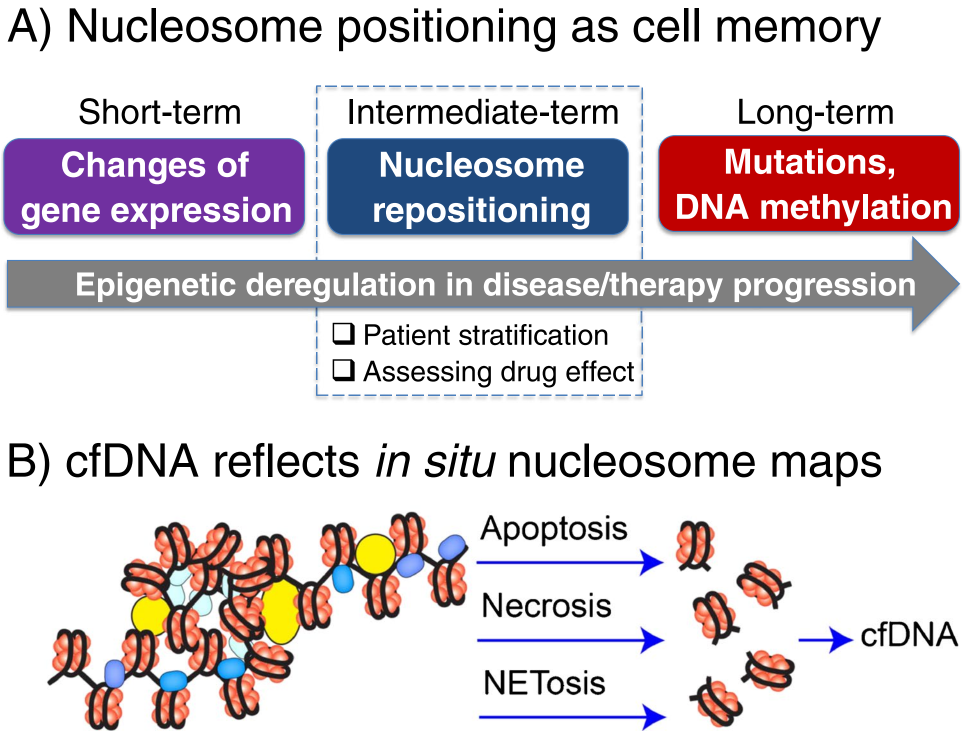

A) Nucleosome positioning as cell memory
Short-term
Changes of gene expression
Intermediate-term
Nucleosome repositioning
Long-term
Mutations, DNA methylation
Epigenetic deregulation in disease/therapy progression
❑ Patient stratification
❑ Assessing drug effect
B) cfDNA reflects *in situ* nucleosome maps
Apoptosis
Necrosis
NETosis
cfDNA

Fig. 1 The motivation for the use of nucleosome positioning in situ and cfDNA as a diagnostic marker. A) Nucleosome positioning acts as the cell memory at intermediate timescales between faster changes of gene expression and reaction metabolites and long-term changes such as the accumulation of mutations and changes of DNA methylation. B) cfDNA extracted from blood plasma or other body liquids reflects the nucleosome positioning landscape in the cells of origin. This is because enzymes that shred chromatin into pieces in processes such as apoptosis, necrosis or NETosis preferentially cut DNA between nucleosomes

as on the abundance of cfDNA derived from tumour cells (ctDNA) which usually correlates with the severity/stage of disease (Abbosh et al. 2017; van der Pol and Mouliere 2019; Zviran et al. 2020). In fact, a recent report showed that elevated cfDNA levels correlate with all-cause mortality (Kananen et al. 2020). Thus, many assays use cfDNA concentration as a marker of disease severity without sequencing.

However, if the detection method is based on few genomic regions that are not represented in cfDNA, then even increasing the sequencing depth would not help the diagnostics. To overcome this problem, it is possible to base cfDNA analysis on a larger number of genomic regions with more subtle epigenetic changes, hence, departing from the idea of mutation analysis and focusing the analysis, for example, on changes in DNA methylation (Shen et al. 2018; Erger et al. 2020; Liu et al. 2020; Nassiri et al. 2020) or hydroxymethylation (Song et al. 2017) of multiple genomic locations that reflect disease-specific changes in the cells of origin. cfDNA methylomics is being actively used in a growing number of applications. The main challenge with this class of approaches is that the detection of DNA modifications requires at least moderate sequencing depth which drives up the cost of the assay. In addition, changes in DNA modifications (as well as DNA sequence) accumulate at a long-term timescale and may not be prevalent at the onset of disease or as a response to therapy (see Fig. 1A). To address these problems, one can consider assays that are based on the detection of smaller changes at a larger number of genomic loci. The most straightforward solution is to look at nucleosome positioning *per se,* which is reflected in cfDNA localisation patterns.

New types of liquid biopsy tests based on nucleosome positioning-inspired analysis of cfDNA are sometimes termed “fragmentomics” and “nucleosomics” (Im et al. 2021). Fragmentomics analyses have been focused on the distribution of sizes of cfDNA fragments (Snyder et al. 2016; Underhill et al. 2016; Mouliere et al. 2018a; Sun et al. 2018; Markus et al. 2021; Guo et al. 2020; Zukowski et al. 2020) as well as the nucleotide patterns at their cut sites (Chandrananda et al. 2015). Sizes of cfDNA fragments reflect the contributions of different biological processes such as apoptosis, necrosis and NETosis. For example, apoptotic enzymes tend to cut out DNA fragments which are slightly smaller than mononucleosomal DNA (Serpas et al. 2019; Han et al. 2020). Such short cfDNA fragments tend to be enriched in cancer patients (van der Pol and Mouliere 2019). On the other hand, ultra-long cfDNA fragments may result from NETosis — a process in which neutrophils release nets of chromatin called neutrophil extracellular traps (NETs) in order to catch and destroy pathogens (Kustanovich et al. 2019). Such long cfDNA fragments can be associated with NETosis in different types of inflammation, for example, in diabetes (Wong et al. 2015) and COVID-19 (Ng et al. 2021). Necrotic cell death is also usually associated with longer DNA fragments (> 10 kb) (Kustanovich et al. 2019). Thus, each type of cell death has its distinct pattern of cfDNA size distribution. cfDNA size may also differ for different body fluids, e.g. urine usually harbours shorter cfDNA than blood plasma (van der Pol and Mouliere 2019). The situation is further complicated by the fact that cell senescence opposes cfDNA release (Rostami et al. 2020). Several studies in fragmentomics suggested using a simple ratio of the amount of short/long cfDNA fragments as an estimate of ctDNA/cfDNA fraction (Mouliere et al. 2018a; van der Pol and Mouliere 2019) but, given the complexity of different cell death pathways mentioned above, it is not always easily interpretable. We will show below that even within a narrow group of medical conditions, the distribution of cfDNA sizes is quite heterogeneous. Another type of fragmentomics analysis is based on the fact that DNA nucleases have different sequence preferences (Serpas et al. 2019; Han et al. 2020) and therefore the distribution of nucleotide patterns at the ends of the cfDNA fragments may provide valuable diagnostic information (van der Pol and Mouliere 2019).

cfDNA nucleosomics is very promising since it eliminates the need of specific genomic markers and pre-set hypotheses about the underlying medical condition, and the bottleneck is now on the computational side. Recent studies have used machine learning to distinguish the cells of origin or perform binary classification healthy/cancer based on cfDNA patterns in gene promoters (Snyder et al. 2016; Wan et al. 2019) or cfDNA density in megabase-size genomic windows (Cristiano et al. 2019). Another successful approach combined several features in the PCA analysis including the amplitude of cfDNA oscillations with 10-bp periodicity, gene copy number variation and the relative abundances of cfDNA fragments with sizes in certain ranges (Mouliere et al. 2018a). One of the directions actively pursued by the cfDNA community is creating targeted sequencing assays based on nucleosomics of a small number of genomic regions — as small as just 6 regions in a recent publication (Zhu et al. 2021). The smaller the number of regions in the targeted nucleosomics assays, the better. However, this also has to be balanced with the sensitivity and ability to recognise more than one medical condition. Currently, the “holy grail” of liquid biopsies — the ability to diagnose an arbitrary medical condition — is still far from reach. Notably, achieving this aim requires access to as many as possible published cfDNA datasets to train the models. Few web sites started appearing that allow visualisation and download of a limited number of cfDNA datasets (Yu et al. 2020; Zheng et al. 2021), but a centralised resource which systematically collects cfDNA datasets from the dozens (and increasing towards hundreds) of currently available cfDNA publications is desperately needed. Here, we have developed such a resource — NucPosDB — which aims to curate all published datasets of sequenced cfDNA, nucleosome positioning maps in vivo and software for nucleosomics analysis. NucPosDB also intends to provide our integrative analysis to quantify the genome in terms of regions with differential nucleosome occupancy and stability (Vainshtein et al. 2017), connecting cfDNA and nucleosome maps in healthy (Schones et al. 2008; Gaffney et al. 2012) and cancer cells (Mallm et al. 2019).

Construction and content

Database structure NucPosDB curates open- and restricted-access datasets of nucleosome positioning in vivo and sequenced cfDNA, as well as computational software for cfDNA/nucleosome positioning analysis and modelling. The structure of the database is summarised in Fig. 2. It contains the following sections: (1) nucleosome positioning in vivo, (2) sequenced cfDNA, (3) database of regions in the human genome with stable nucleosome occupancy for a given condition, and the repository of software for nucleosomics, further separated into three subsections devoted to (4) analysis of nucleosome maps in vivo, (5) prediction of nucleosome formation preferences based on DNA sequence and (6) cfDNA-specific analysis.

The section of nucleosome positioning in vivo contains datasets from > 250 publications in > 16 biological species, dominated by *Saccharomyces cerevisiae* (28.6%), *Mus musculus* (25.9%), *Homo sapiens* (20.1%) and *Drosophila melanogaster* (14.3%). Figure 3 demonstrates relative abundances of different model organisms used for nucleosome positioning analysis. This section of the database features more than 18 experimental techniques, dominated by MNase-seq, complemented by methods such as histone H3 ChIP-seq, MH-seq, MPE-seq, MiSeq, NOME-seq and RED-seq (detailed in our previous publications Teif 2016; Teif and Clarkson 2019) as well as newer techniques based on long single-molecule reads, Nanopore-seq (Baldi et al. 2018) and Fiber-seq (Stergachis et al. 2020), and nucleosome-scale mapping of 3D genome contact, Micro-C (Hsieh et al. 2015). Techniques such as ATAC-seq, which map nucleosomes only in a limited number of "open" genomic locations, are currently not included in NucPosDB.

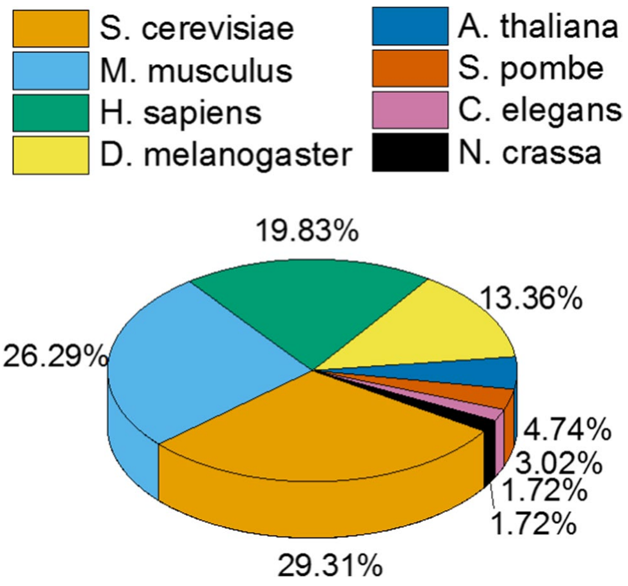


Fig. 3 The distribution of nucleosome positioning datasets across different biological species

The repository of sequenced cfDNA represents a recent addition to NucPosDB and currently features more than 75 studies. cfDNA processing is complicated by the fact that many datasets dealing with patient data have restricted access, e.g. where the raw data is stored in the European Nucleotide Archive (ENA) or the database of Genotypes and Phenotypes (dbGaP). The application for access to each such dataset is considered individually by the corresponding

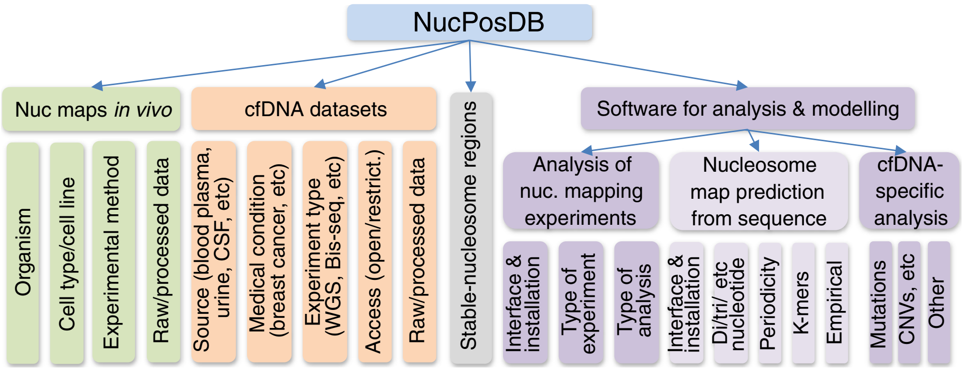


Fig. 2 The structure of NucPosDB containing six major sections (listed left to right in the scheme): (1) nucleosome maps measured in vivo in different cell types, (2) sequenced cfDNA datasets, (3) regions with stable nucleosome occupancy in the human genome for different conditions based on (1) and (2), (4) software for analysis of nucleosome mapping experiments, (5) software for predicting preferences of nucleosome formation from the DNA sequence and (6) software for cfDNA-specific analysis

data access committee, and the time required to receive regulatory approval may reach several months. On the other hand, when the raw data is stored in databases such as GEO, such datasets are available without restrictions. NucPosDB curates both open-access and restricted-assess datasets, but only open-access datasets are supplied with the processed data including the locations of all mapped nucleosomes and stable-nucleosome regions (see below). Table 1 shows examples of cfDNA datasets from NucPosDB that have no access restrictions. cfDNA datasets included in NucPosDB can be browsed by organism (e.g. human, mouse or dog). Currently, the majority of cfDNA datasets included in NucPosDB are of human origin. For patients, it is possible to select medical condition (currently around 50 conditions), source of cfDNA (blood, e.g. serum/plasma, cerebrospinal liquid or urine), experimental method (at present 12 methods) and access type (restricted or not).

A special NucPosDB section is devoted to the regions of the human genome with stable nucleosome occupancy. It contains condition-specific coordinates of genomic locations where nucleosome occupancy has low relative standard deviation across all samples within the same condition. This is defined with NucTools (Vainshtein et al. 2017) using a window-based approach as detailed below and arranged in tab-separated BED files with the following columns: chromosome, region start, region end, normalised nucleosome occupancy, standard deviation, relative deviation. In addition, for a number of open-access cfDNA entries, our database provides access to the uniformly processed BED files with locations of all mapped nucleosomes (based on paired-end cfDNA reads). We have mapped these cfDNA reads to the human genome assemblies hg19 and hg38 as detailed below. These were further processed with NucTools (Vainshtein et al. 2017) to generate tab-separated files with the following columns: chromosome, fragments start, fragment end, fragment size. Each patient sample has been processed separately. The links from the interactive database tables lead to the file repository with directories separated by publication and further split into different medical conditions.

The repository of software for analysis of nucleosome positioning experiments currently contains 31 entries representing different classes of software ranging from nucleosome array visualisers and nucleosome peak callers to predictors of specific parameters such as the nucleosome repeat length (Vainshtein et al. 2017). The repository of algorithms for prediction of DNA sequence-dependent affinity of nucleosome octamer currently contains 23 entries, as described previously (Teif 2016; Teif and Clarkson 2019). The repository of software specific for the analysis of cfDNA currently includes 32 entries.

Data collection and curation The datasets were searched in NCBI GEO as well as in peer-reviewed publications and preprints from bioRxiv and medRxiv servers. Initial search was conducted using the keywords “nucleosome positioning”, “MNase-seq” and “cfDNA”. Further relevant studies were extracted through publication chaining. Over 300 papers reporting relevant datasets and software were arranged into five sections: nucleosome maps in vivo, cfDNA datasets, computational tools for nucleosome positioning analysis, DNA sequence-based modelling and cfDNA analysis. The criterion for the dataset inclusion was the ability to reconstruct a nucleosome positioning profile with single-nucleotide resolution based on a given dataset. Dataset

Table 1 Example open-access datasets from NucPosDB reporting whole genome sequencing of cfDNA

Description	Medical conditions	*N* patients
Generation of highly biomimetic quality control materials for non-invasive prenatal testing based on enzymatic digestion of matched mother–child cell lines (Zhang et al. 2019)	Prenatal testing	2
Sequencing of cfDNA derived from the plasma of individuals of different ages (Teo et al. 2019)	Ageing	12
Very short mitochondrial DNA fragments and heteroplasmy in human plasma (Zhang et al. 2016)	Sepsis, tissue transplantation	7
Cell-free DNA comprises an in vivo, genome-wide nucleosome footprint that informs its tissue(s)-of-origin (Snyder et al. 2016)	Healthy, lupus, Crohn’s disease, colitis, cancer	60
Cell-free DNA provides a good representation of the tumour genome despite its biased fragmentation patterns (Ma et al. 2017)	Cancer	5
The next-generation sequencing (NGS) technologies related assessments of circulating tumour DNA (ctDNA) in both primary brain tumours and metastatic brain tumours (Liang et al. 2020)	cancer	28
WGS of human pooled plasma cfDNA sampled from GI diseased individuals (PRJEB1791)	Healthy, cancer, inflammatory bowel disease	24
Decoding the evolutionary response to prostate cancer therapy by plasma genome sequencing (Ramesh et al. 2020)	Cancer	23

reporting methods such as ChIP-seq and microarrays were normally excluded unless the corresponding publications provided specific nucleosome positioning analysis. ATAC-seq was excluded since it maps nucleosomes only in a limited number of "open" genomic locations. cfDNA datasets were included when they were obtained using any variation of a sequencing technique that involves whole-genome or targeted sequencing and thus allows partial or complete reconstruction of nucleosome profiles. This includes methods determining DNA methylation and hydroxymethylation, but not microarray-based techniques/assays.

User interface The user interface of NucPosDB is realised in PHP. The search and keywords selection is currently enabled with the help of the TablePress plugin for WordPress (tablepress.org, author Tobias Bäthge, Magdeburg, Germany). Datasets can be searched by typing a query into the search box or using multiple-item selection in drop-down menus such as "Organism" and "Experiment type". Additionally, the repository of cfDNA datasets contains drop-down menus "Medical condition", "cfDNA source" and "Access" (open/restricted). The interactive tables with experimental datasets can be ordered or sub-selected by the combination of the following criteria: "Description" (typically includes the title of the original publication and a clickable link), "Organism", "Cell type" (only in the section nucleosome maps in vivo), "Experiment type/method", "Raw data" and "Processed data". The cfDNA repository allows additional selection/ordering criteria: "Medical condition", "cfDNA source", "Access" (open/restricted) and "Number of patients".

Data processing The calculation of the histogram of DNA fragment size distribution was carried out using R. The calculation of nucleotide frequencies was performed with HOMER (Heinz et al. 2010). Raw paired-end reads were aligned to the human genomes hg19 and hg38 using Bowtie (Langmead et al. 2009), reporting only uniquely aligned reads with up to two mismatches. Normalised nucleosome occupancy was calculated genome-wide with 100-bp windows by dividing the average nucleosome occupancy in a given window by the average chromosome-wide nucleosome occupancy. Stable-nucleosome regions were determined with NucTools with 100 bp sliding window and the threshold 0.5 applied to the relative deviation of nucleosome occupancy across all samples with a given condition, as described previously (Vainshtein et al. 2017). The relative deviation was defined as the ratio of the standard deviation to the normalised nucleosome occupancy in a given window.

Utility and discussion

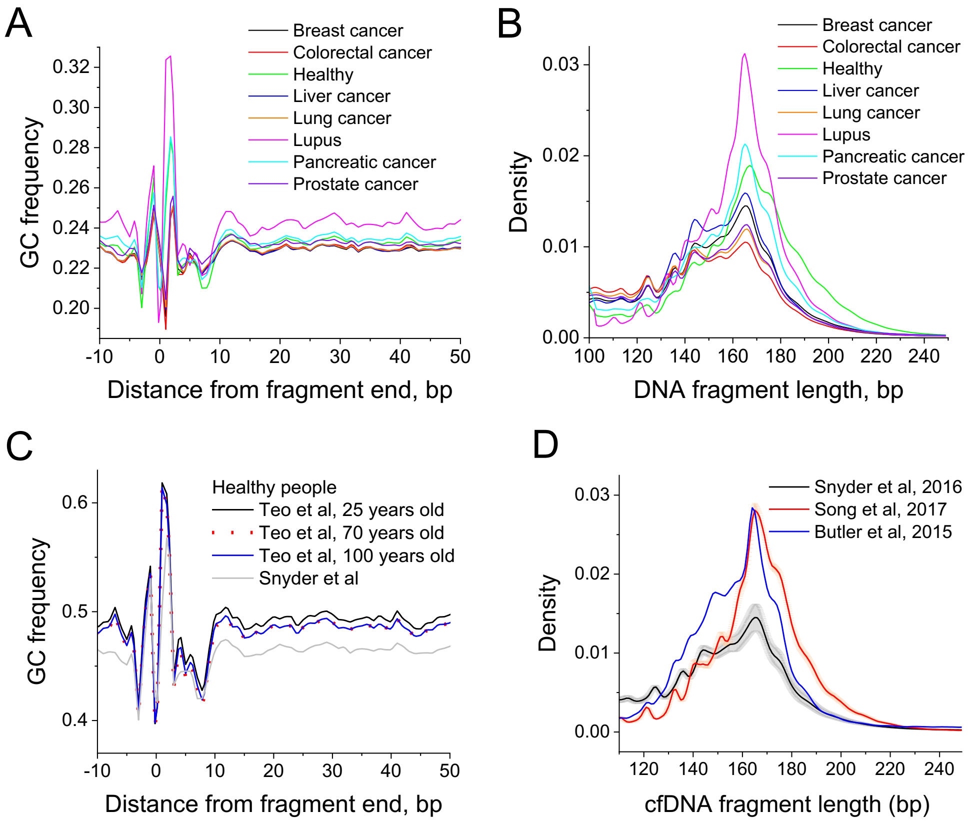
One of the main purposes of having a centralised repository of nucleosome positioning/cfDNA datasets is to be able to assess the data heterogeneity within conditions and the variability between different conditions and experimental protocols. While a systematic analysis of such variability of all datasets in NucPosDB is beyond the scope of the current work, let us demonstrate the typical distributions of two basic characteristics of cfDNA, namely the GC content and the DNA fragment sizes.

Firstly, let us consider the nucleotide frequency as a function of the distance from the cfDNA fragment end (Fig. 4A). This type of analysis is motivated by previous findings that endogenous nucleases have distinct preferences for DNA cut sites, and these preferences are different from artificial cut sites observed in MNase-seq experiments (Serpas et al. 2019; van der Pol and Mouliere 2019; Han et al. 2020). Apoptosis in different types of cancer may involve the same set of nucleases; therefore, based on this metric, different types of cancer may not be easily distinguishable from each other. Indeed, this is what we observe for the distribution of GC frequencies near cfDNA fragment ends in Fig. 4A. On the other hand, different biological processes such as NETosis may employ a different combination of enzymes; thus, it may be possible to distinguish medical conditions that are characterised by inflammation (inflammation triggers NETosis). Indeed, Fig. 4A shows that nucleotide profiles of cfDNA from patients with lupus (systemic inflammation) differ quite significantly from those in cancer or healthy controls.

Next, let us consider the distributions of DNA fragment sizes. Previous studies reported that cancer cfDNA appears to have shorter fragments that are more strongly digested (Snyder et al. 2016; Underhill et al. 2016; Mouliere et al. 2018a; Sun et al. 2018; Markus et al. 2021; Guo et al. 2020; Zukowski et al. 2020). Our results do show differences in cfDNA fragment size distributions, most notably for lupus (Fig. 4B). The difference of cfDNA in lupus from cancer and healthy samples may be explained by the different DNA digestion processes undergoing in this systemic inflammatory condition (Fig. 1B). However, special care is required to normalise the data and take into account different protocols (e.g. the lupus samples in the study considered above may have been clinically processed in a different way than the cancer samples).

The differences in the experimental protocols used in different labs for cfDNA processing as well as comorbidities of patients may play major roles in the data interpretation. To demonstrate this, Fig. 4 C and D compare samples from different subgroups of healthy people. Figure 4 C shows that the average GC content of cfDNA extracted

Fig. 4 Aggregate characteristics of cfDNA datasets across different medical conditions (A, B) and ages of healthy people (C, D). A GC content as a function of the distance from the end of cfDNA fragment (Snyder et al. 2016). B Distribution of lengths of cfDNA fragments (Snyder et al. 2016). C GC content as a function of the distance from the end of cfDNA fragment for 25-, 70- and 100-year-old people (Teo et al. 2019), compared with pooled healthy people from another study (Snyder et al. 2016). D Differences of cfDNA fragment sizes for cfDNA of breast cancer patients collected in three different studies (Snyder et al. 2016; Song et al., 2017 and Butler et al., 2015)



for whole-genome sequencing by different methods differs dramatically. In one case, Teo et al. collected cfDNA from three age groups (25, 70 and 100 years old) and the differences of GC profiles between these age groups are pretty minor (Teo et al. 2019). On the other hand, in another group of healthy people where sequencing was performed by the method of Snyder et al., the average GC content is about 3% lower (Snyder et al. 2016). Such difference may lead to biased representation of different types of genomic regions and needs to be taken into account when comparing datasets across different laboratories. Indeed, Fig. 4D shows that the distribution of cfDNA fragment sizes varies quite substantially between datasets reported by three different labs even when all of these refer to the same condition (breast cancer in this example), and when samples within one lab's dataset are consistently similar to each other. This probably reflects differences in experimental protocols and needs to be taken with special care when performing nucleosomics analysis for cancer diagnostics. Similar care is needed when comparing MNase-seq datasets obtained in different laboratories, because it is known that parameters such as the degree of chromatin digestion greatly affect nucleosome maps due to differential sensitivity of partially unwrapped nucleosomes to digestion level (Teif et al. 2014; Chereji et al. 2016; Ramachandran et al. 2017). In such situations, it may be helpful to adjust clinically relevant analyses taking into account the locations of regions with stable nucleosome occupancy in a given condition as reported by NucPosDB.

Finally, the examples shown above demonstrate that the development of a robust clinical diagnostics based on cfDNA nucleosomics will require many datasets across different laboratories and types of wet lab assays. This is where NucPosDB may be particularly helpful, allowing the use of data from more than 10,000 patients.

Conclusions

NucPosDB offers a user-friendly interface and curates published in vivo nucleosome positioning datasets including > 18 types of experimental techniques in > 16 different species and distinct cell types, supplemented with the repository curating cfDNA datasets for more than 10,000 patients as well as the software packages for "nucleosomics" analysis. For many open-access datasets, we also provide systematically calculated condition-specific stable-nucleosome regions which are useful in comparison between different conditions. In the future, NucPosDB can serve as a centralised resource for the nucleosomics community, providing a platform for the annotation of cfDNA datasets and storage of processed data required for training models for patient diagnostics with liquid biopsies.

Author contribution MS developed the repository of cfDNA, curated the repository of nucleosome positioning in vivo and performed cfDNA data analysis. KVP contributed to the curation of the repository of nucleosome positioning in vivo and data analysis. SPA compiled the repository of software for cfDNA analysis and contributed to cfDNA analysis. DRJ contributed to cfDNA analysis. VBT designed and supervised the project, performed programming and drafted the manuscript.

Funding This work was supported by the Wellcome Trust 200733/Z/16/Z to VBT and the Genetics Society fellowships to MS and KVP.

Data Availability NucPosDB is available at https://generegulation.org/nucposdb/.

Declarations

Competing interests The authors declare no competing interests.

References

Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, Le Quesne J, Moore DA, Veeriah S, Rosenthal R et al (2017) Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature 545:446–451

Baldi S, Korber P, Becker PB (2020) Beads on a string-nucleosome array arrangements and folding of the chromatin fiber. Nat Struct Mol Biol 27:109–118

Baldi S, Krebs S, Blum H, Becker PB (2018) Genome-wide measurement of local nucleosome array regularity and spacing by nanopore sequencing. Nat Struct Mol Biol 25:894–901

Bezdan D, Grigorev K, Meydan C, Pelissier Vatter FA, Cioffi M, Rao V, MacKay M, Nakahira K, Burnham P, Afshinnekoo E et al. 2020. Cell-free DNA (cfDNA) and exosome profiling from a year-long human spaceflight reveals circulating biomarkers. iScience **23**: 101844.

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

Chereji RV, Kan TW, Grudniewska MK, Romashchenko AV, Berezikov E, Zhimulev IF, Guryev V, Morozov AV, Moshkin YM (2016) Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster. Nucleic Acids Res 44:1036–1051

Clarkson CT, Deeks EA, Samarista R, Mamayusupova H, Zhurkin VB, Teif VB (2019) CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. Nucleic Acids Res 47:11181–11196

Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, Jensen SO, Medina JE, Hruban C, White JR et al (2019) Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570:385–389

Dawson MA, Kouzarides T (2012) Cancer epigenetics: from mechanism to therapy. Cell 150:12–27

Diermeier S, Kolovos P, Heizinger L, Schwartz U, Georgomanolis T, Zirkel A, Wedemann G, Grosveld F, Knoch TA, Merkl R et al (2014) TNFalpha signalling primes chromatin for NF-kappaB binding and induces rapid and widespread nucleosome repositioning. Genome Biol 15:536

Dudley JC, Diehn M (2021) Detection and diagnostic utilization of cellular and cell-free tumor DNA. Annu Rev Pathol 16:199–222

Erger F, Norling D, Borchert D, Leenen E, Habbig S, Wiesener MS, Bartram MP, Wenzel A, Becker C, Toliat MR et al (2020) cfNOMe - a single assay for comprehensive epigenetic analyses of cell-free DNA. Genome Med 12:54

Frenel JS, Carreira S, Goodall J, Roda D, Perez-Lopez R, Tunariu N, Riisnaes R, Miranda S, Figueiredo I, Nava-Rodrigues D 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

Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK (2012) Controls of nucleosome positioning in the human genome. PLoS Genet 8:e1003036

Guo J, Ma K, Bao H, Ma X, Xu Y, Wu X, Shao YW, Jiang M, Huang J (2020) Quantitative characterization of tumor cell-free DNA shortening. BMC Genomics 21:473

Han DSC, Ni M, Chan RWY, Chan VWH, Lui KO, Chiu RWK, Lo YMD (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, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (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

Henikoff S, Church GM (2018) Simultaneous discovery of cell-free DNA and the nucleosome ladder. Genetics 209:27–29

Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A, Tolstorukov MY, Appert A et al (2014) Comparative analysis of metazoan chromatin organization. Nature 512:449–452

Hsieh TH, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ (2015) Mapping nucleosome resolution chromosome folding in yeast by micro-C. Cell 162:108–119

Ignatiadis M, Sledge GW, Jeffrey SS (2021) Liquid biopsy enters the clinic - implementation issues and future challenges. Nat Rev Clin Oncol 18:297–312

Im YR, Tsui DWY, Diaz LA Jr, Wan JCM (2021) Next-generation liquid biopsies: embracing data science in oncology. Trends Cancer 7:283–292

Ioshikhes IP, Albert I, Zanton SJ, Pugh BF (2006) Nucleosome positions predicted through comparative genomics. Nat Genet 38:1210–1215

Kananen L, Hurme M, Jylha M, Harkanen T, Koskinen S, Stenholm S, Kahonen M, Lehtimaki T, Ukkola O, Jylhava J (2020) Circulating cell-free DNA level predicts all-cause mortality independent of other predictors in the Health 2000 survey. Sci Rep 10:13809

Kitzman JO, Snyder MW, Ventura M, Lewis AP, Qiu R, Simmons LE, Gammill HS, Rubens CE, Santillan DA, Murray JC et al. 2012. Noninvasive whole-genome sequencing of a human fetus. Sci Transl Med **4**: 137ra176.

Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A (2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Res 22:1735–1747

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

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Li C, Luscombe NM (2020) Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. Nat Commun 11:1363

Liang J, Zhao W, Lu C, Liu D, Li P, Ye X, Zhao Y, Zhang J, Yang D. 2020. Next-generation sequencing analysis of ctDNA for the detection of glioma and metastatic brain tumors in adults. Front Neurol **11**.

Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Liu MC, Oxnard GR, Klein EA, Smith D, Richards D et al (2020) Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol 31:745–759

Lo YMD, Han DSC, Jiang P, Chiu RWK. 2021. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science **372**.

Ma X, Zhu L, Wu X, Bao H, Wang X, Chang Z, Shao YW, Wang Z. 2017. Cell-free DNA provides a good representation of the tumor genome despite its biased fragmentation patterns. PLoS One **12**: e0169231.

Mallm JP, Iskar M, Ishaque N, Klett LC, Kugler SJ, Muino JM, Teif VB, Poos AM, Grossmann S, Erdel F et al. 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

Markus H, Zhao J, Contente-Cuomo T, Stephens MD, Raupach E, Odenheimer-Bergman A, Connor S, McDonald BR, Moore B, Hutchins E et al (2021) Analysis of recurrently protected genomic regions in cell-free DNA found in urine. Sci Trans Med 13:eaaz3088

Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, Mair R, Goranova T, Marass F, Heider K et al. 2018a. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med **10**.

Mouliere F, Mair R, Chandrananda D, Marass F, Smith CG, Su J, Morris J, Watts C, Brindle KM, Rosenfeld N. 2018b. Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. EMBO Mol Med **10**.

Nassiri F, Chakravarthy A, Feng S, Shen SY, Nejad R, Zuccato JA, Voisin MR, Patil V, Horbinski C, Aldape K et al (2020) Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. Nat Med 26:1044–1047

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 41:988–994

Palande V, Detroja R, Gorohovski A, Glass R, Flueh C, Kurtz M, Perez S, Shay DR, Siegal T, Frenkel-Morgenstern M (2020) A liquid biopsy platform for detecting gene-gene fusions as glioma diagnostic biomarkers and drug targets. *bioRxiv* 2020.02.25.963975

Parmar JJ, Padinhateeri R (2020) Nucleosome positioning and chromatin organization. Curr Opin Struct Biol 64:111–118

Peng X, Li H-D, Wu F-X, Wang J (2021) Identifying the tissues-of-origin of circulating cell-free DNAs is a promising way in noninvasive diagnostics. Brief Bioinform 22:bbaa060

Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, Anagnostou V, Fiksel J, Cristiano S, Papp E et al. 2017. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med **9**.

Pich O, Muinos F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. 2018. Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. Cell **175**: 1074–1087 e1018.

Ramachandran S, Ahmad K, Henikoff S. 2017. Transcription and remodeling produce asymmetrically unwrapped nucleosomal intermediates. Mol Cell **68**: 1038–1053 e1034.

Ramesh N, Sei E, Tsai PC, Bai S, Zhao Y, Troncoso P, Corn PG, Logothetis C, Zurita AJ, Navin NE (2020) Decoding the evolutionary response to prostate cancer therapy by plasma genome sequencing. Genome Biol 21:162

Rostami A, Lambie M, Yu CW, Stambolic V, Waldron JN, Bratman SV. 2020. Senescence, necrosis, and apoptosis govern circulating cell-free DNA release kinetics. Cell Rep **31**: 107830.

Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132:887–898

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J (2006) A genomic code for nucleosome positioning. Nature 442:772–778

Serpas L, Chan RWY, Jiang P, Ni M, Sun K, Rashidfarrokhi A, Soni C, Sisirak V, Lee WS, Cheng SH 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

Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, Zuzarte PC, Borgida A, Wang TT, Li T et al (2018) Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 563:579–583

Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J (2016) Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell 164:57–68

Song C-X, Yin S, Ma L, Wheeler A, Chen Y, Zhang Y, Liu B, Xiong J, Zhang W, Hu J et al (2017) 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. Cell Res 27:1231–1242

Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA (2020) Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science 368:1449–1454

Sun K, Jiang P, Wong AIC, Cheng YKY, Cheng SH, Zhang H, Chan KCA, Leung TY, Chiu RWK, Lo YMD (2018) Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. Proc Natl Acad Sci U S A 115:E5106-e5114

Teif VB (2016) Nucleosome positioning: resources and tools online. Brief Bioinform 17:745–757

Teif VB, Clarkson CT (2019) Nucleosome positioning. In: Encyclopedia of Bioinformatics and Computational Biology, (ed. S Ranganathan, et al.). Academic Press, Oxford, pp 308–317

Teif VB, Beshnova DA, Vainshtein Y, Marth C, Mallm JP, Hofer T, Rippe K (2014) Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. Genome Res 24:1285–1295

Teif VB, Mallm JP, Sharma T, Mark Welch DB, Rippe K, Eils R, Langowski J, Olins AL, Olins DE (2017) Nucleosome repositioning during differentiation of a human myeloid leukemia cell line. Nucleus 8:188–204

Teif VB, Vainshtein Y, Caudron-Herger M, Mallm JP, Marth C, Hofer T, Rippe K (2012) Genome-wide nucleosome positioning during embryonic stem cell development. Nat Struct Mol Biol 19:1185–1192

Teo YV, Capri M, Morsiani C, Pizza G, Faria AMC, Franceschi C, Neretti N. 2019. Cell-free DNA as a biomarker of aging. Aging Cell **18**: e12890.

Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, Wolfler A, Zebisch A, Gerger A, Pristauz G et al (2019) Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun 10:4666

Ulz P, Thallinger GG, Auer M, Graf R, Kashofer K, Jahn SW, Abete L, Pristauz G, Petru E, Geigl JB et al (2016) Inferring expressed genes by whole-genome sequencing of plasma DNA. Nat Genet 48:1273–1278

Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, Baker DN, Gligorich KM, Rostomily RC, Bronner MP, Shendure J. 2016. Fragment length of circulating tumor DNA. PLoS Genet **12**: e1006162.

Vainshtein Y, Rippe K, Teif VB (2017) NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. BMC Genomics 18:158

Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, 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, Collins C (2016) Cell-free DNA (cfDNA): Clinical significance and utility in cancer shaped by emerging technologies. Mol Cancer Res 14:898–908

Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, Pacey S, Baird R, Rosenfeld N (2017) Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer 17:223–238

Wan N, Weinberg D, Liu T-Y, Niehaus K, Ariazi EA, Delubac D, Kannan A, White B, Bailey M, Bertin M 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

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, Gallant M, Wang Y, Goldfine AB, Kahn CR, Wagner DD (2015) Diabetes primes neutrophils to undergo NETosis, which impairs wound healing. Nat Med 21:815–819

Yu F, Li K, Li S, Liu J, Zhang Y, Zhou M, Zhao H, Chen H, Wu N, Liu Z et al (2020) CFEA: a cell-free epigenome atlas in human diseases. Nucleic Acids Res 48:D40–D44

Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309:626–630

Zhang R, Ding J, Gao P, Li Z, Tan P, Li J (2019) Generation of highly biomimetic quality control materials for noninvasive prenatal testing based on enzymatic digestion of matched mother-child cell lines. Clin Chem 65:761–770

Zhang R, Nakahira K, Guo X, Choi AMK, Gu Z (2016) Very short mitochondrial DNA fragments and heteroplasmy in human plasma. Sci Rep 6:36097

Zheng H, Zhu MS, Liu Y (2021) FinaleDB: a browser and database of cell-free DNA fragmentation patterns. Bioinformatics 37:2502–2503

Zhu G, Guo YA, Ho D, Poon P, Poh ZW, Wong PM, Gan A, Chang MM, Kleftogiannis D, Lau YT et al (2021) Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. Nat Commun 12:2229

Zukowski A, Rao S, Ramachandran S. 2020. Phenotypes from cell-free DNA. Open Biol **10**: 200119.

Zviran A, Schulman RC, Shah M, Hill STK, Deochand S, Khamnei CC, Maloney D, Patel K, Liao W, Widman AJ et al (2020) Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. Nat Med 26:1114–1124

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.