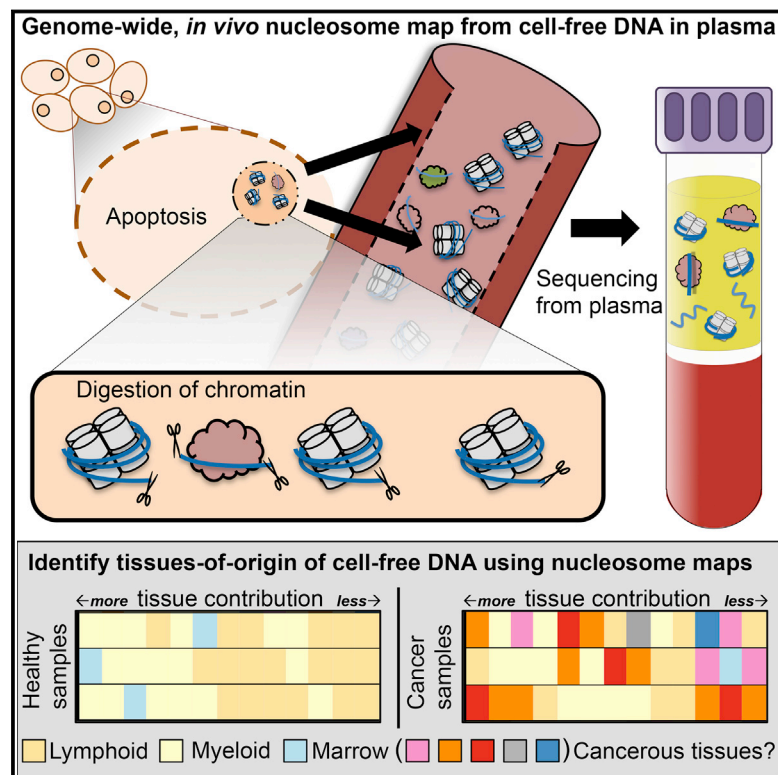# Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin

Graphical Abstract

Authors

Matthew W. Snyder, Martin Kircher,
Andrew J. Hill, Riza M. Daza,
Jay Shendure

Correspondence

shendure@uw.edu

In Brief

Deep sequencing of circulating cell-free DNA yields a dense, genome-wide map of nucleosome occupancy that enables identification of its cell-types of origin, potentially enabling the noninvasive monitoring of a much broader set of clinical conditions than currently possible.

Highlights

- Sequencing cell-free DNA yields a genome-wide map of in vivo nucleosome occupancy

- Short cell-free DNA fragments directly footprint transcription factor occupancy

- In healthy persons, these epigenetic footprints match hematopoietic lineages

- Additional contributions are detected in cancer, often aligning with the cancer type

CrossMark

CellPress

Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin

Matthew W. Snyder,[1,3] Martin Kircher,[1,3] Andrew J. Hill,[1] Riza M. Daza,[1] and Jay Shendure[1,2,*]
[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[2]Howard Hughes Medical Institute, Seattle, WA 98195, USA
[3]Co-first author
*Correspondence: shendure@uw.edu
http://dx.doi.org/10.1016/j.cell.2015.11.050

SUMMARY

Nucleosome positioning varies between cell types. By deep sequencing cell-free DNA (cfDNA), isolated from circulating blood plasma, we generated maps of genome-wide in vivo nucleosome occupancy and found that short cfDNA fragments harbor footprints of transcription factors. The cfDNA nucleosome occupancies correlate well with the nuclear architecture, gene structure, and expression observed in cells, suggesting that they could inform the cell type of origin. Nucleosome spacing inferred from cfDNA in healthy individuals correlates most strongly with epigenetic features of lymphoid and myeloid cells, consistent with hematopoietic cell death as the normal source of cfDNA. We build on this observation to show how nucleosome footprints can be used to infer cell types contributing to cfDNA in pathological states such as cancer. Since this strategy does not rely on genetic differences to distinguish between contributing tissues, it may enable the noninvasive monitoring of a much broader set of clinical conditions than currently possible.

INTRODUCTION

Cell-free DNA (cfDNA) is present in the circulating plasma, urine, and other bodily fluids of humans (Chan et al., 2003). The cfDNA comprises double-stranded DNA fragments that are overwhelmingly short (<200 base pairs [bp]) and normally at a low concentration (Fleischhacker and Schmidt, 2007). In healthy individuals, plasma cfDNA is believed to derive primarily from apoptosis of normal cells of the hematopoietic lineage, with minimal contributions from other tissues (Lui et al., 2002). The short half-life of cfDNA in the circulation (Lo et al., 1999) suggests a model of ongoing release from apoptotic cells and rapid degradation or filtration. The size distribution of cfDNA fragments bears correspondence with these origins—specifically, peaks corresponding to nucleosomes (~147 bp) and chromatosomes (nucleosome + linker histone; ~167 bp) have been noted (Fan et al., 2008; Lo et al., 2010)—and some proportion of cfDNA

may circulate as nucleosomes or chromatosomes, rather than as free DNA (Holdenrieder et al., 2005; Wimberger et al., 2011).

In the context of specific physiological conditions or disease processes, a substantial proportion of cfDNA may be derived from a different distribution of tissues than during the typical, healthy state. This fact has been exploited in recent years to achieve noninvasive diagnostics based on cfDNA composition. In pregnant women, ~10%–15% of cfDNA originates from placental trophoblasts, and cfDNA-based screening for fetal genetic abnormalities is now common in high-risk pregnancies (Chiu et al., 2008; Fan et al., 2008). In oncology, the monitoring of advanced cancers by quantifying mutations or aneuploidy in tumor-shed cfDNA is gaining traction (Diaz and Bardelli, 2014). In transplant medicine, allograft rejection events can be correlated with abnormally high levels of donor-derived cfDNA fragments contributed by the transplanted solid organ (Snyder et al., 2011).

Despite these advances, a common limitation is the requirement for genetic differences to distinguish between contributing tissues, e.g., fetus versus mother, tumor versus normal, or donor versus recipient. Conditions such as myocardial infarction (Chang et al., 2003), stroke (Rainer et al., 2003), and autoimmune disorders (Galeazzi et al., 2003) are associated with elevations in cfDNA levels, possibly consequent to tissue damage, but cannot be specifically monitored via cfDNA because of the lack of such genetic differences. Furthermore, even as mutations enable monitoring of tumor-derived cfDNA, they only weakly inform a tumor's tissue-of-origin.

We hypothesized that if cfDNA is the detritus of cell death and if the boundaries of cfDNA fragments are biased by their association with nucleosomes, then the fragmentation patterns observed in an individual's cfDNA might contain evidence of the epigenetic landscape(s) of the cells giving rise to these fragments and thus, of their tissue(s)-of-origin (i.e., a strategy that does not rely on genotypic differences between contributing cell types).

To evaluate this hypothesis, we first set out to deeply sequence cfDNA to better understand the processes that give rise to it. We use the resulting data to build a map of nucleosome occupancy that approaches saturation of the mappable human genome. By optimizing protocols to recover short fragments, we discover that the in vivo occupancies of transcription factors (TFs) such as CTCF are also directly footprinted by cfDNA. Finally, we

CrossMark

show that nucleosome spacing in regulatory elements and gene bodies, as revealed by cfDNA sequencing in healthy individuals, correlates most strongly with DNase I hypersensitivity (DHS) and gene expression in lymphoid and myeloid cell lines. To test whether we can infer additional contributing tissues in non-healthy states, we sequenced cfDNA samples from five late-stage cancer patients. The patterns of nucleosome spacing in these samples reveal additional contributions to cfDNA that correlate most strongly with non-hematopoietic tissues or cell lines, often matching the anatomical origin of the patient's cancer.

RESULTS

cfDNA Fragments Correspond to Chromatosomes and Contain Substantial DNA Damage

We prepared conventional sequencing libraries by end repair and by adaptor ligation to cfDNA fragments purified from plasma pooled from an unknown number of healthy individuals ("BH01") or a single individual ("IH01") (Figure 1A; Table S1). We sequenced these libraries to 96- and 105-fold coverage (1.5 billion and 1.6 billion fragments). The fragment length distributions have a dominant peak at ~167 bp (coincident with the length of DNA associated with a chromatosome) and ~10.4 bp periodicity in the 100–160 bp range (Figure 1B). These distributions support a model in which cfDNA fragments are preferentially protected from nuclease cleavage by association with proteins—in this case, by the nucleosome core particle (NCP) and linker histone—but where some degree of additional nicking or cleavage occurs in relation to the helical pitch of nucleosome-bound DNA (Fan et al., 2008; Lo et al., 2010). Further supporting this model is the dinucleotide composition of these fragments, which recapitulates key features of earlier studies of micrococcal nuclease (MNase)-derived, nucleosome-associated fragments (e.g., bias against A/T dinucleotides at the dyad) (Gaffney et al., 2012) and supports the notion that the NCP is symmetrically positioned with respect to the chromatosome (Harshman et al., 2013) (Figure 1C).

A prediction of this model is widespread DNA damage, e.g., single-strand nicks as well as 5′ and 3′ overhangs. During conventional library preparation, damaged as well as short double-stranded DNA molecules (Mouliere et al., 2014) may be poorly recovered. To address this, we prepared a single-stranded cfDNA library from an additional healthy individual ("IH02") using a protocol adapted from studies of ancient DNA (Figure S1A; Table S2) (Gansauge and Meyer, 2013) and sequenced it to 30-fold coverage (779 million fragments). The fragment length distribution again exhibited a dominant peak at ~167 bp but was considerably enriched for shorter fragments relative to conventional library preparation (Figures 1D, 1E, and S1B–S1D). Although all libraries exhibit ~10.4 bp periodicity, the fragment sizes are offset by 3 bp for the two methods, consistent with damaged or non-flush input molecules whose true endpoints are more faithfully represented in single-stranded libraries.

A Genome-wide Map of In Vivo Nucleosome Protection Based on Deep cfDNA Sequencing

We next asked whether the predominant local positions of nucleosomes in tissue(s) contributing to cfDNA could be inferred from the distribution of aligned fragment endpoints. Specifically, we expect that cfDNA fragment endpoints should cluster adjacent to NCP boundaries, while also being depleted on the NCP itself. To quantify this, we developed a windowed protection score (WPS), which is the number of DNA fragments completely spanning a 120 bp window centered at a given genomic coordinate minus the number of fragments with an endpoint within that same window (Figure 2A). As expected, the WPS correlates with the locations of nucleosomes within strongly positioned arrays, as mapped by other groups with in vitro methods (Gaffney et al., 2012; Valouev et al., 2011) or ancient DNA (Pedersen et al., 2014) (Figure 2B). At other sites, the WPS correlates with genomic features such as DHS sites, e.g., consistent with the repositioning of nucleosomes flanking a distal regulatory element (Figure 2C).

We applied a heuristic peak-calling algorithm to the genome-wide WPS of the BH01, IH01, and IH02 datasets to identify and score 12.6 million, 11.9 million, and 9.7 million local maxima of nucleosome protection (Figures 2 and S2A). In each sample, the mode distance between adjacent peaks is 185 bp with low variance (Figure 2D), consistent with previous analyses of the nucleosome repeat length in mammalian cells (Teif et al., 2012; Valouev et al., 2011). The positions of peak calls are concordant between samples (Figures 2E and S2B), e.g., the median (absolute) distance from a BH01 peak call to a nearest-neighbor-IH01 peak call is 23 bp overall but <10 bp for the most highly scored peaks (Figure S2C). Furthermore, the density of peak calls is constant across the spectrum of GC content (Figure S2D).

As biases introduced by nuclease specificity or library preparation might artifactually contribute to the signal of nucleosome protection, we also simulated fragment endpoints, matching for the depth, size distribution, and terminal dinucleotide frequencies (Figures S1C and S1D). We then calculated genome-wide WPS and, using the same heuristic, called 10.3 million, 10.2 million, and 8.0 million local maxima for simulated datasets matched to BH01, IH01, and IH02, respectively. Peaks from simulated datasets are associated with lower scores than peaks from real datasets (Figure S2E) and do not align well with the locations of peaks called from real datasets (Figures 2E and S2F).

We next pooled and reanalyzed data from BH01, IH01, and IH02 ("CH01"; 231-fold coverage; 3.8 billion fragments; Table S1). The resulting map of in vivo nucleosome occupancy comprises 12.9 million peaks, with higher scores and approaching saturation (Figure S2E). Considering all peak-to-peak distances below 500 bp (Figure 2F), the CH01 peaks span 2.53 gigabases.

Nucleosomes are known to be well-positioned in relation to landmarks of gene regulation, e.g., transcriptional start sites (Pedersen et al., 2014) and exon-intron boundaries (Andersson et al., 2009; Chodavarapu et al., 2010). We observe such positioning in our data as well, in relation to landmarks of transcription, translation, and splicing (Figures 3A–3D and S3). We further examined the median peak-to-peak spacing within 100 kilobase (kb) windows that had been assigned to compartment A (enriched for open chromatin) or compartment B (enriched for closed chromatin) on the basis of chromatin contact maps in a lymphoblastoid cell line (Rao et al., 2014). Nucleosomes in compartment A exhibit tighter spacing than nucleosomes in compartment B (median 187 bp [A] versus 190 bp [B]), with
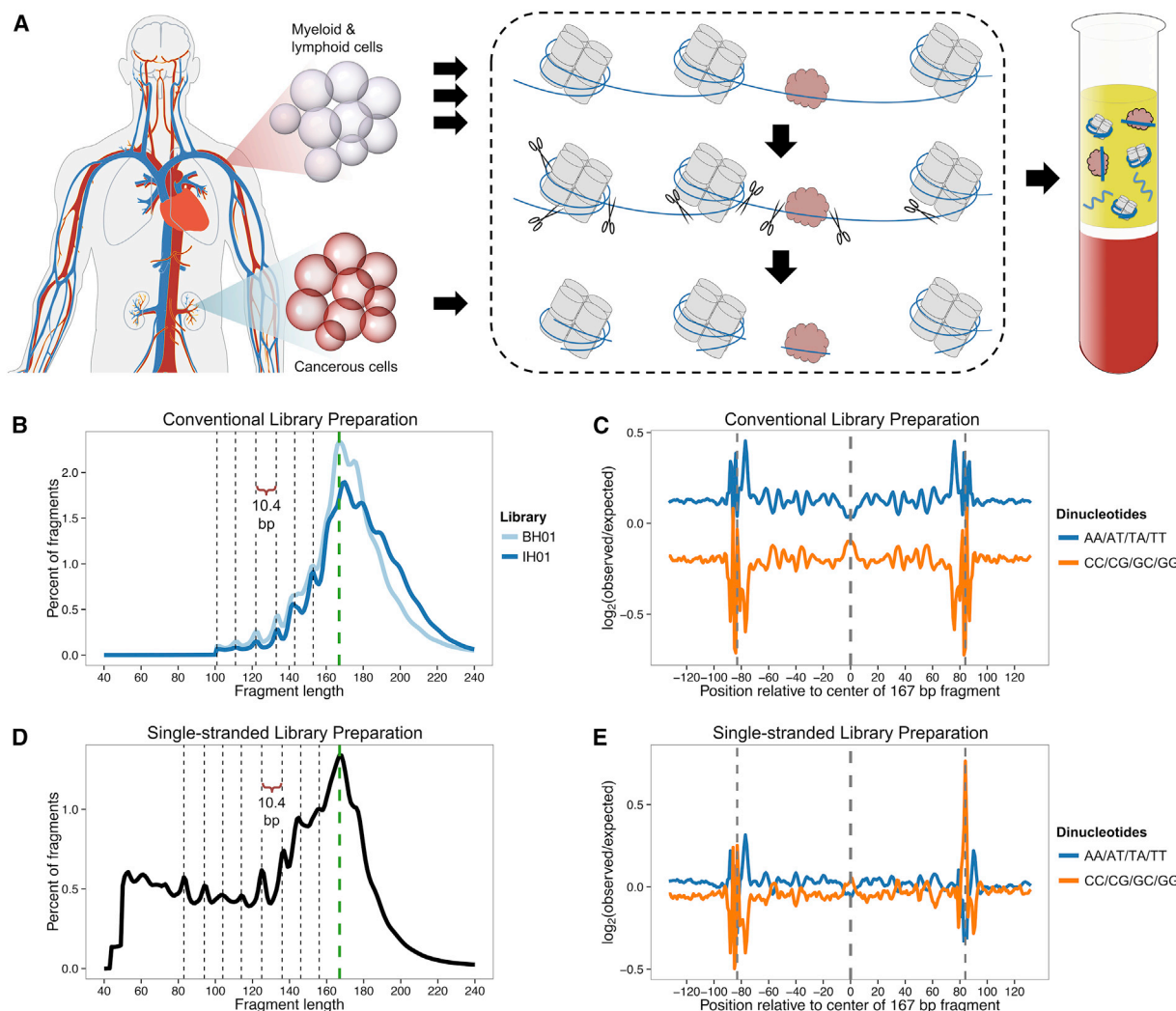
Figure 1. Origins and Characteristics of cfDNA Fragments in Human Plasma

(A) Schematic overview of cfDNA fragmentation. Apoptotic or necrotic cell death results in near-complete digestion of native chromatin. Protein-bound DNA fragments, typically associated with histones or TFs, preferentially survive digestion and are released into the circulation, while naked DNA is lost. Fragments can be recovered from peripheral blood plasma following proteinase treatment. In healthy individuals, cfDNA is primarily derived from myeloid and lymphoid cell lineages, but contributions from one or more additional tissues may be present in certain medical or physiological conditions.

(B) Fragment length of cfDNA observed with conventional sequencing library preparation, inferred from alignment of paired-end reads. A reproducible peak in fragment length at 167 bp (green dashed line) is consistent with association with chromatosomes. Additional peaks evidence ∼10.4 bp periodicity, corresponding to the helical pitch of DNA on the nucleosome core. Enzymatic end repair during library preparation removes 5′ and 3′ overhangs and may obscure true cleavage sites.

(C) Dinucleotide composition of 167 bp fragments and flanking genomic sequence in conventional libraries. Observed dinucleotide frequencies in the BH01 library were compared to expected frequencies from simulated fragments.

(D) Fragment length of cfDNA in single-stranded sequencing library preparation. No enzymatic end repair is performed to template molecules during library preparation. Short fragments of 50–120 bp are highly enriched compared to conventional libraries. While ∼10.4 bp periodicity remains, its phase is shifted by ∼3 bp.

(E) Dinucleotide composition of 167 bp fragments and flanking genomic sequence in single-stranded library IH02, calculated as in (C). The apparent difference in the background level of bias between BH01 and IH02 relate to differences between the simulations, rather than the real libraries (data not shown).

See also Figure S1 and Table S1.

further differences between subcompartments (Figure 3E). Along the length of chromosomes, no general pattern is seen, except that median nucleosome spacing drops sharply in pericentromeric regions, presumably driven by strong positioning across arrays of alpha satellites (Figures 2B and 3F).

Short cfDNA Fragments Directly Footprint CTCF and Other TFs

Previous studies of DNase I cleavage patterns identified two dominant classes of fragments: longer fragments associated with cleavage between nucleosomes and shorter fragments
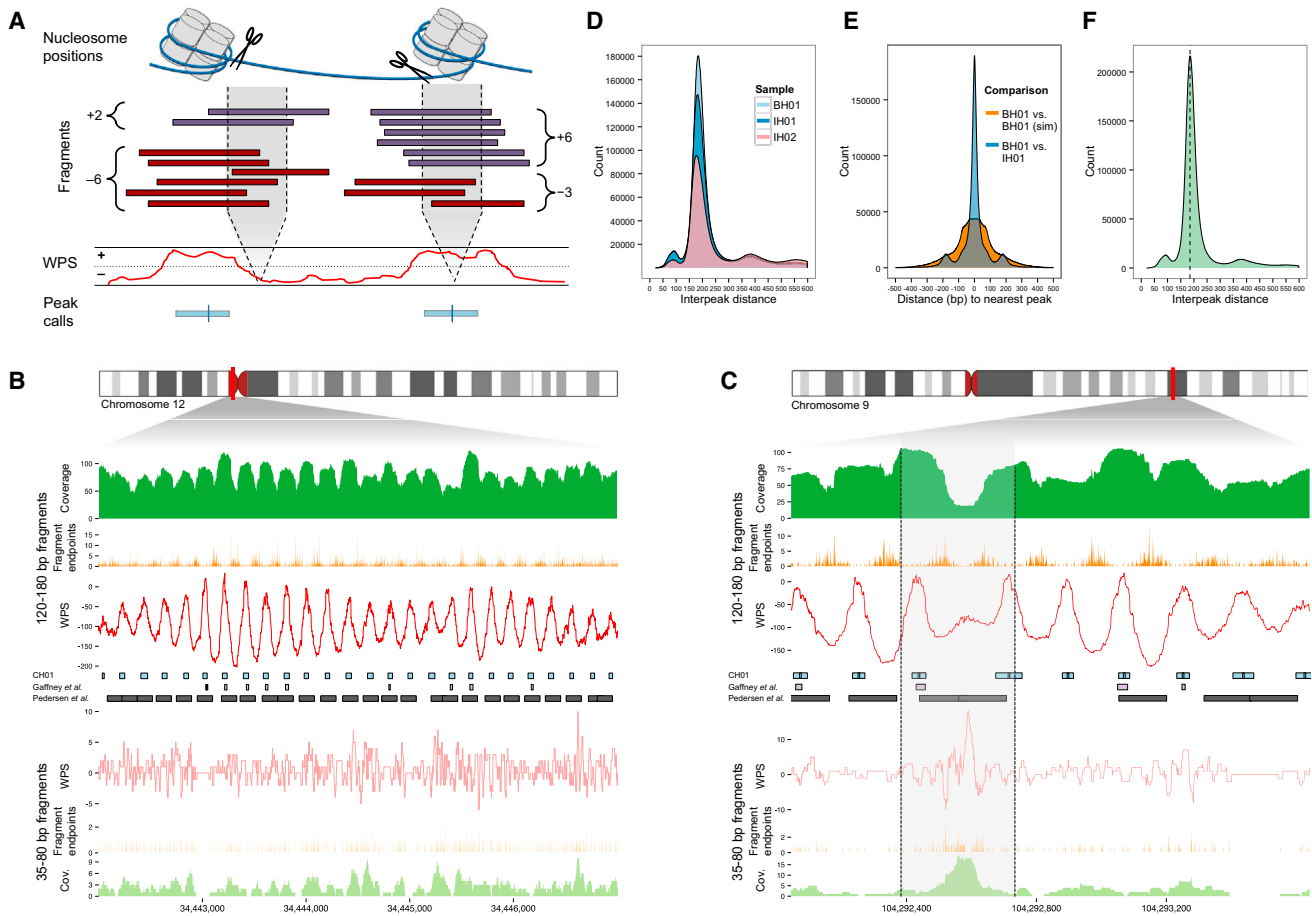
Figure 2. Genome-wide Determination of Nucleosome Positions from cfDNA Fragmentation Patterns

(A) Schematic of inference of nucleosome positioning. A per-base WPS is calculated by subtracting the number of fragment endpoints within a 120 bp window from the number of fragments completely spanning the window. High WPS values indicate increased protection of DNA from digestion; low values indicate that DNA is unprotected. Peak calls identify contiguous regions of elevated WPS.

(B) Strongly positioned nucleosomes at a well-studied alpha-satellite array. Coverage, fragment endpoints, and WPS values from sample CH01 are shown for long fragment (120 bp window; 120–180 bp fragments) or short fragment (16 bp window; 35–80 bp fragments) bins at a pericentromeric locus on chromosome 12. Nucleosome calls from CH01 (middle, blue boxes) are regularly spaced across the locus. Nucleosome calls from two published callsets (Gaffney et al., 2012; Pedersen et al., 2014) (middle, purple, and black boxes) are also displayed.

(C) Inferred nucleosome positioning around a DHS site. Coverage, fragment endpoints, WPS values, and nucleosome calls are shown as in (B). The hypersensitive region (gray shading) is marked by reduced coverage in the long fragment bin. Nucleosome calls adjacent to the DHS site are spaced more widely than typical adjacent pairs, consistent with accessibility of the intervening sequence to regulatory proteins including TFs. Coverage of short fragments, which may be associated with such proteins, is increased at the DHS site, which overlaps with several annotated TFBSs (not shown).

(D) Distances between adjacent peaks by sample. Distances are measured between adjacent peak centers.

(E) Comparison of peak calls between samples. For each pair of samples, the distances between each peak call in the sample with fewer peaks and the nearest peak call in the other sample are shown. Negative and positive numbers indicate the nearest peak is upstream or downstream, respectively.

(F) Distances between adjacent peaks, sample CH01. The dotted black line indicates the mode of the distribution (185 bp).

See also Figure S2.

associated with cleavage adjacent to transcription factor binding sites (TFBS) (Vierstra et al., 2014). To ask whether in-vivo-derived cfDNA fragments also result from two classes of sensitivity to nuclease cleavage, we partitioned sequence reads (CH01) on the basis of inferred fragment length and recalculated the WPS using long fragments (120–180 bp; 120 bp window; the same as the WPS described for nucleosome calling) or short fragments (35–80 bp; 16 bp window) separately (Figures 2B and 2C). To obtain a set of well-defined TFBSs enriched for actively bound

sites in our data, we intersected clustered FIMO predictions (Grant et al., 2011; Maurano et al., 2012) with a unified set of chromatin immunoprecipitation sequencing (ChIP-seq) peaks from Encyclopedia of DNA Elements (ENCODE) for each TF.

Consistent with observations by others (Fu et al., 2008; Pedersen et al., 2014; Teif et al., 2012), the long fraction WPS (L-WPS) supports strong organization of nucleosomes near CTCF binding sites (Figure 4A). However, we also observe a strong signal in the short fraction WPS (S-WPS) coincident with the CTCF site itself
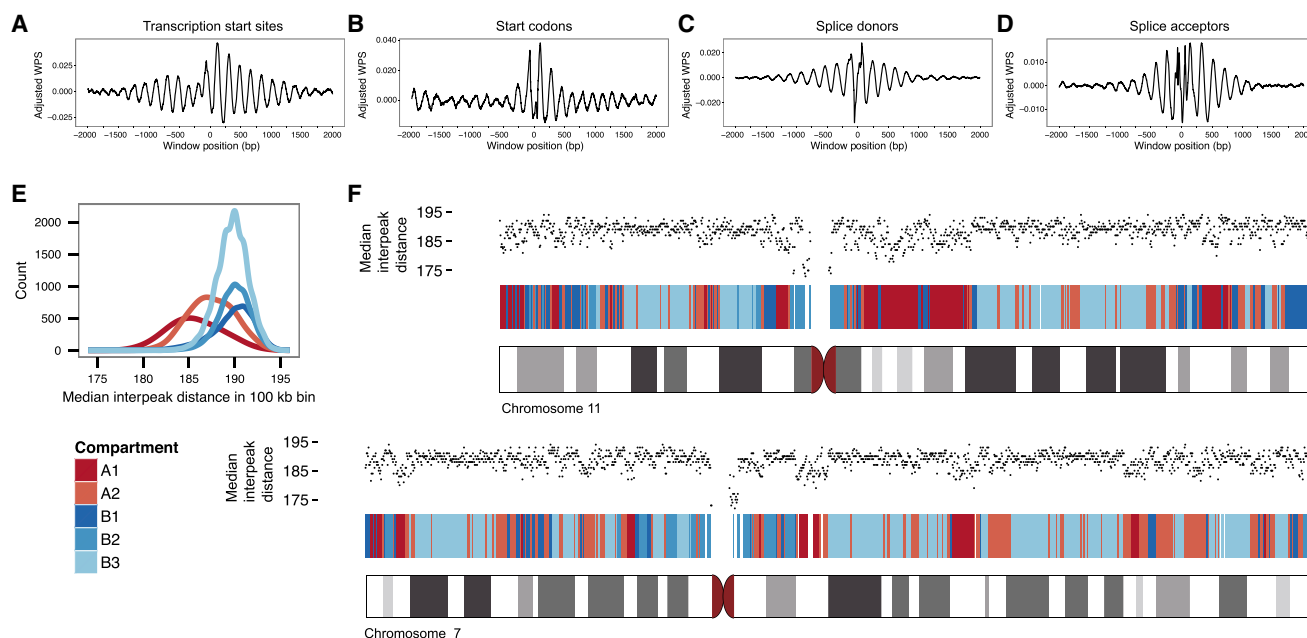
Figure 3. Nucleosome Positioning and Spacing Correlates with Genomic Features

(A) Aggregate, adjusted WPS (120 bp window) around 22,626 TSS. TSS are aligned at the 0 position after adjusting for strand and direction of transcription. Aggregate WPS is tabulated for both real data and simulated data by summing per-TSS WPS at each position relative to the centered TSS. The values plotted represent the difference between the real and simulated aggregate WPS (see Experimental Procedures for details).

(B) Aggregate, adjusted WPS around 22,626 start codons.

(C and D) Aggregate, adjusted WPS around 224,910 splice donor (C) and 224,910 splice acceptor (D) sites.

(E) Nucleosome spacing in A/B compartments. Median nucleosome spacing in non-overlapping 100 kb bins, each containing ~500 nucleosome calls, is calculated genome-wide. A/B compartment predictions, also with 100 kb resolution, are shown for GM12878. Compartments A and B are associated with open and closed chromatin, respectively.

(F) Nucleosome spacing and A/B compartments on chromosomes 7 and 11. A/B segmentation (red and blue bars) largely recapitulates chromosomal G-banding (ideograms, gray bars). Median nucleosome spacing (black dots) is calculated in 100 kb bins.

See also Figure S3.

(Ong and Corces, 2014) (Figures 4A and 4B). We stratified CTCF sites based on our confidence that they are bound in vivo. Experimentally well-supported CTCF sites exhibit substantially broader spacing between the flanking −1 and +1 nucleosomes based on the L-WPS, consistent with their repositioning upon CTCF binding (~190 bp → ~260 bp; Figures 4C and S4A). Experimentally well-supported CTCF sites also exhibit a much stronger S-WPS signal over the CTCF binding site itself (Figures 4D and S4B).

We performed similar analyses for additional TFs for which both FIMO predictions and ENCODE ChIP-seq data were available (Figure S4C). For many of these, e.g., ETS and MAFK (Figures 4E and 4F), we observe a short fraction footprint that is accompanied by periodic signal in the L-WPS, consistent with strong positioning of nucleosomes surrounding bound TFBS. Overall, these data support the view that short cfDNA fragments, which are much better recovered by the single-stranded protocol (Figures 1B and 1D), directly footprint the in vivo occupancy of DNA-bound TFs including CTCF and others.

Nucleosome Spacing Patterns Inform cfDNA Tissues-Of-Origin

We next asked whether in vivo nucleosome protection, as measured through cfDNA, could be used to infer the cell types contributing to cfDNA in healthy individuals. We examined the peak-to-peak spacing of nucleosome calls within DHS sites defined in 116 diverse biological samples (Maurano et al., 2012). Similar to bound CTCF sites (Figure 4C), we observed substantially broader spacing for nucleosome pairs within a subset of DHS sites, plausibly corresponding to sites at which the nucleosomes are repositioned by intervening TF binding in the cell type(s) giving rise to cfDNA (~190 bp → ~260 bp; Figure 5A). Indeed, the proportion of widened nucleosome spacing (~260 bp) varies considerably depending on which cell type's DHS sites are used. However, all of the cell types for which this proportion is highest are lymphoid or myeloid in origin (e.g., CD3_CB-DS17706, etc. in Figure 5A), consistent with hematopoietic cell death as the dominant source of cfDNA in healthy individuals (Lui et al., 2002).

We next reexamined the signal of nucleosome protection in the vicinity of transcriptional start sites (TSS) (Figure 3A). If we stratify based on gene expression in a lymphoid cell line, NB-4, we observe strong differences in the patterns of nucleosome protection in relation to the TSS in highly versus lowly expressed genes (Figure 5B). Furthermore, if we examine the S-WPS, we observe a clear footprint immediately upstream of the TSS whose intensity also strongly correlates with expression
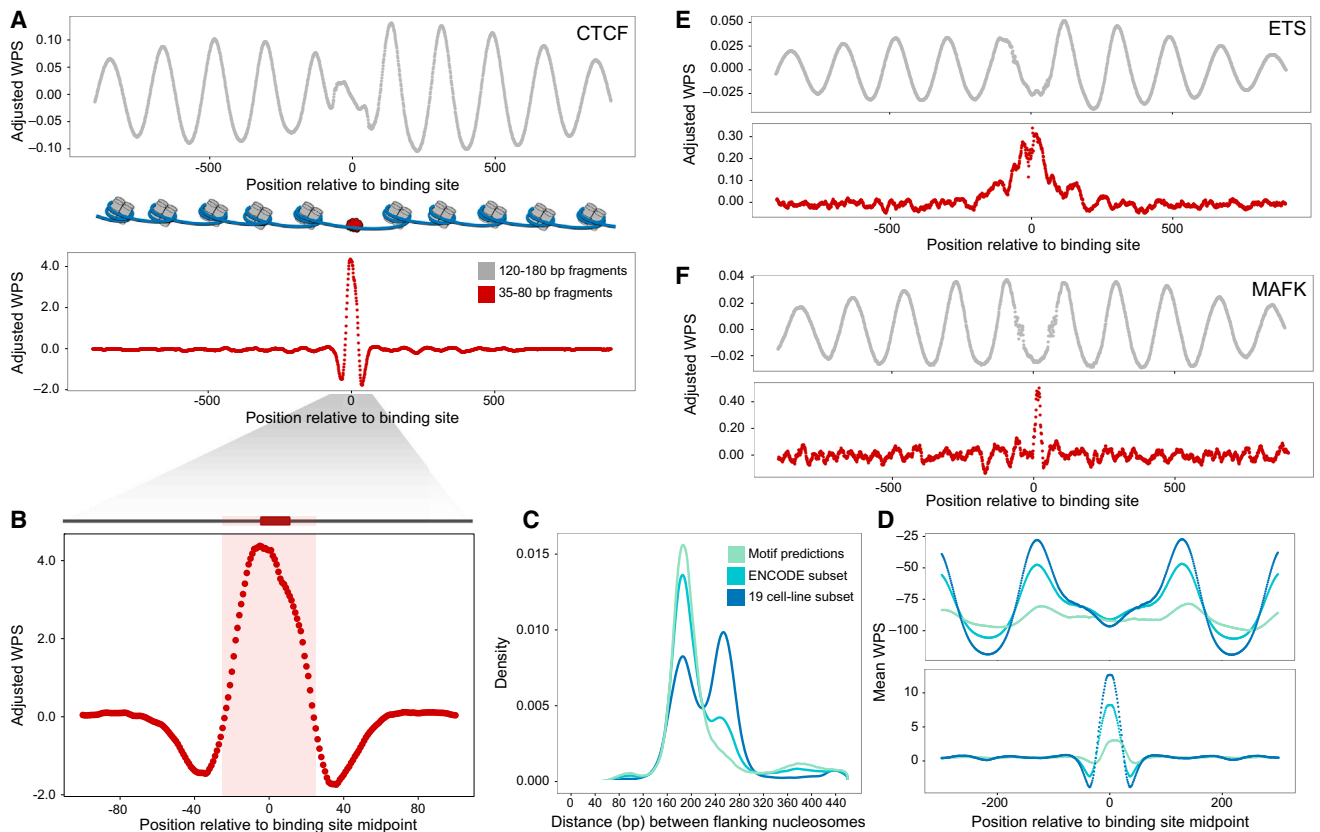
Figure 4. Short cfDNA Fragments Footprint CTCF and other TF Binding Sites

Clustered FIMO predictions were intersected with ChIP-seq data to obtain confident sets of binding site predictions for various TFs. Aggregate, adjusted WPS was calculated for both the long (120–180 bp) and short (35–80 bp) fractions of cfDNA fragments. Higher WPS values indicate greater nucleosome or TF protection, respectively.

(A) Aggregate, adjusted WPS for 518,632 predicted CTCF binding sites for the long (top) and short (bottom) cfDNA fractions. Binding of CTCF results in strong positioning of neighboring nucleosomes.

(B) Aggregate, adjusted WPS, calculated for 518,632 predicted CTCF sites as in (A) and magnified for detail, for 35–80 bp cfDNA fragments. The pink shading indicates the larger 52 bp CTCF binding motif, and the black box shows the location of the 17 bp motif used for FIMO predictions.

(C) Density of −1 to +1 nucleosome spacing around CTCF sites derived from clustered FIMO predictions (purely motif-based: 518,632 sites), a subset of these predictions overlapping with ENCODE ChIP-seq peaks (93,530 sites), and a further subset active across 19 cell lines (23,723 sites). Flanking nucleosome spacing at the least stringent set of sites (motif-based) mirrors the genome-wide average (~185 bp), while spacing at the most stringent set of sites is highly enriched for greater distances (~260 bp), consistent with active CTCF binding and repositioning of adjacent nucleosomes.

(D) Mean WPS calculated for the long (top) and short (bottom) cfDNA fractions at the sets of CTCF sites in (C).

(E and F) Aggregate, adjusted WPS calculated for both long (top) and short (bottom) cfDNA fractions at predicted binding sites for ETS (210,798 sites) (E) and MAFK (32,159 sites) (F). For both factors, short fraction WPS is consistent with TF-conferred protection of the binding site, whereas long fraction WPS evidences regular, local positioning of surrounding nucleosomes.

See also Figure S4.

(Figure 5C). This plausibly reflects footprinting of the transcription pre-initiation complex, or some component thereof, at transcriptionally active genes.

These observations support our thesis that cfDNA fragmentation patterns indeed contain signal that might be used to infer the tissue(s) or cell-type(s) giving rise to cfDNA. However, a challenge is that relatively few reads in a genome-wide cfDNA library directly overlap DHS sites and TSSs.

It was previously observed that nucleosome spacing varies between cell types as a function of chromatin state and gene expression (Teif et al., 2012; Valouev et al., 2011). In general, open chromatin and transcription are associated with a shorter nucleosome repeat length, consistent with our analyses of compartment A versus B (Figure 3E). In our peak calls, we also observe a correlation between nucleosome spacing across gene bodies and their expression levels, with tighter spacing associated with higher expression (Figure 5D; $\rho = -0.17$; n = 19,677 genes). The correlation is highest for the gene body itself, relative to adjacent regions (upstream 10 kb $\rho = -0.08$; downstream 10 kb $\rho = -0.01$). If we limit this analysis to gene bodies that span at least 60 nucleosome calls, the correlation is much stronger ($\rho = -0.50$; n = 12,344 genes).

An advantage of exploiting signals, such as nucleosome spacing across gene bodies or other domains, is that a

much larger proportion of cfDNA fragments will be informative. To test whether we might be able to detect mixtures of signals resulting from multiple cell types contributing to cfDNA, we performed fast Fourier transformation (FFT) on the L-WPS across the first 10 kb of gene bodies and on a gene-by-gene basis. The intensity of the FFT signal is correlated with gene expression at specific frequency ranges, with a maximum at 177–180 bp for positive correlation and a minimum at ~199 bp for negative correlation (Figure 5E). In performing this analysis against a dataset of 76 expression datasets for human cell lines and primary tissues (Uhlén et al., 2015), we observed that the strongest correlations are with hematopoietic lineages (Figure 5E). For example, the most highly ranked negative correlations with average intensity in the 193–199 bp frequency range for each of three healthy samples (BH01, IH01, IH02) are all to lymphoid cell lines, myeloid cell lines, or bone marrow tissue (Figure 5F; Table S3). These top correlation ranks are robust to downsampling (Figure S5A).

Nucleosome Spacing in Cancer Patients' cfDNA Identifies Non-hematopoietic Contributions

We next sought to ask whether we could detect signatures of non-hematopoietic cell types contributing to circulating cfDNA in non-healthy states. We first screened 44 plasma samples from individuals with clinical diagnoses of a variety of Stage IV cancers with light sequencing of single-stranded libraries prepared from cfDNA (Table S4; median 2.2-fold coverage; of note, with same protocol as and many in the same batch as IH02). Because matched tumor genotypes were not available, we scored each sample on two metrics of aneuploidy to identify a subset likely to contain a high proportion of tumor-derived cfDNA: first, the deviation from the expected proportion of reads derived from each chromosome (Leary et al., 2012); and second, the per-chromosome allele balance profile for a panel of common single-nucleotide polymorphisms. Based on these metrics, we sequenced single-stranded libraries derived from five individuals (with a small cell lung cancer, a squamous cell lung cancer, a colorectal adenocarcinoma, a hepatocellular carcinoma, and a ductal carcinoma in situ breast cancer) to a depth similar to that of IH02 (Table S5).

We again performed FFT on the L-WPS across gene bodies and correlated the average intensity in the 193–199 bp frequency range against the same 76 expression datasets for human cell lines and primary tissues (Uhlén et al., 2015). In contrast with the three samples from healthy individuals (where all of the top 10 and nearly all of the top 20 correlations were to lymphoid or myeloid lineages), we observed that many of the most highly ranked cell lines or tissues represent non-hematopoietic lineages, in some cases aligning with the cancer type (Figure 5F). For example, for IC17, where the patient had a hepatocellular carcinoma, the top-ranked correlation is with HepG2, a hepatocellular carcinoma cell line. For IC35, where the patient had a ductal carcinoma in situ breast cancer, the top-ranked correlation is with MCF7, a metastatic breast adenocarcinoma cell line. In other cases, the cell lines or primary tissues that exhibit the greatest change in correlation rank align with the cancer type. For example, for IC15, where

the patient had small-cell lung cancer, the largest change in correlation rank (−31) is for a small-cell lung cancer cell line (SCLC-21H), and the second largest change (−25) is for a neuroblastoma cancer cell line (SH.SY5Y). For IC20 (a lung squamous cell carcinoma) and IC35 (a colorectal adenocarcinoma), there are many non-hematopoietic cancer cell lines displacing the lymphoid/myeloid cell lines in terms of correlation rank, but the alignment of these to the specific cancer type is less clear. It is possible that the molecular profile of these patients' cancers is not well-represented among our 76 expression datasets (none are lung squamous cell carcinomas; CACO-2 is a cell line derived from a colorectal adenocarcinoma, but is highly heterogeneous [Sambuy et al., 2005]). As with samples from healthy individuals, the top correlation ranks associated with the samples from cancer patients are robust to downsampling (Figure S5A). However, in silico "dilution" of samples from cancer patients with samples from healthy patients results in proportionally lower ranks for non-hematopoietic cell lines, consistent with our expectation (Figure S5B).

DISCUSSION

We present a dense, genome-wide map of in vivo nucleosome protection inferred from plasma-borne cfDNA fragments. Although the number of peaks is essentially saturated in CH01, other metrics of quality continued to be a function of sequencing depth (Figure S2C). We therefore constructed an additional genome-wide nucleosome map based on all of the cfDNA sequencing that we have performed to date ("CA01", 14.5 billion fragments; 700-fold coverage; 13.0 M peaks). Although this map exhibits even more uniform spacing (Figure 6A) and more highly supported peak calls (not shown), we caution that it is based on cfDNA from both healthy and non-healthy individuals (Tables S1 and S5).

Our work builds directly on previous efforts to map nucleosome occupancy in human cells genome-wide (Gaffney et al., 2012; Pedersen et al., 2014; Schones et al., 2008; Teif et al., 2012; Valouev et al., 2011), but our callset is substantially more complete and uniform (Figure 6). The fragments that we observe are generated by endogenous physiological processes, avoiding the technical variation associated with in vitro MNase digestion. A limitation of our map is that the cell types that give rise to cfDNA are inevitably heterogeneous (e.g., a mixture of lymphoid and myeloid cell types in healthy individuals). Nonetheless, the map's relative completeness may facilitate a deeper understanding of the interplay of nucleosome positioning and spacing with primary sequence, epigenetic regulation, transcriptional output, and nuclear architecture.

A second goal of this study was to explore whether the nucleosome footprints contained in cfDNA fragments can be used to infer contributing cell types. Through comparisons with gene expression and regulatory site profiles, we identify the epigenetic signature of hematopoietic lineages contributing to cfDNA in healthy individuals, with plausible additional contributions from one or more non-hematopoietic tissues in a small panel of individuals with advanced cancers. For this proof of concept, we stacked the odds in our favor by focusing on individuals that appeared to have large burdens of tumor-derived DNA. However, it
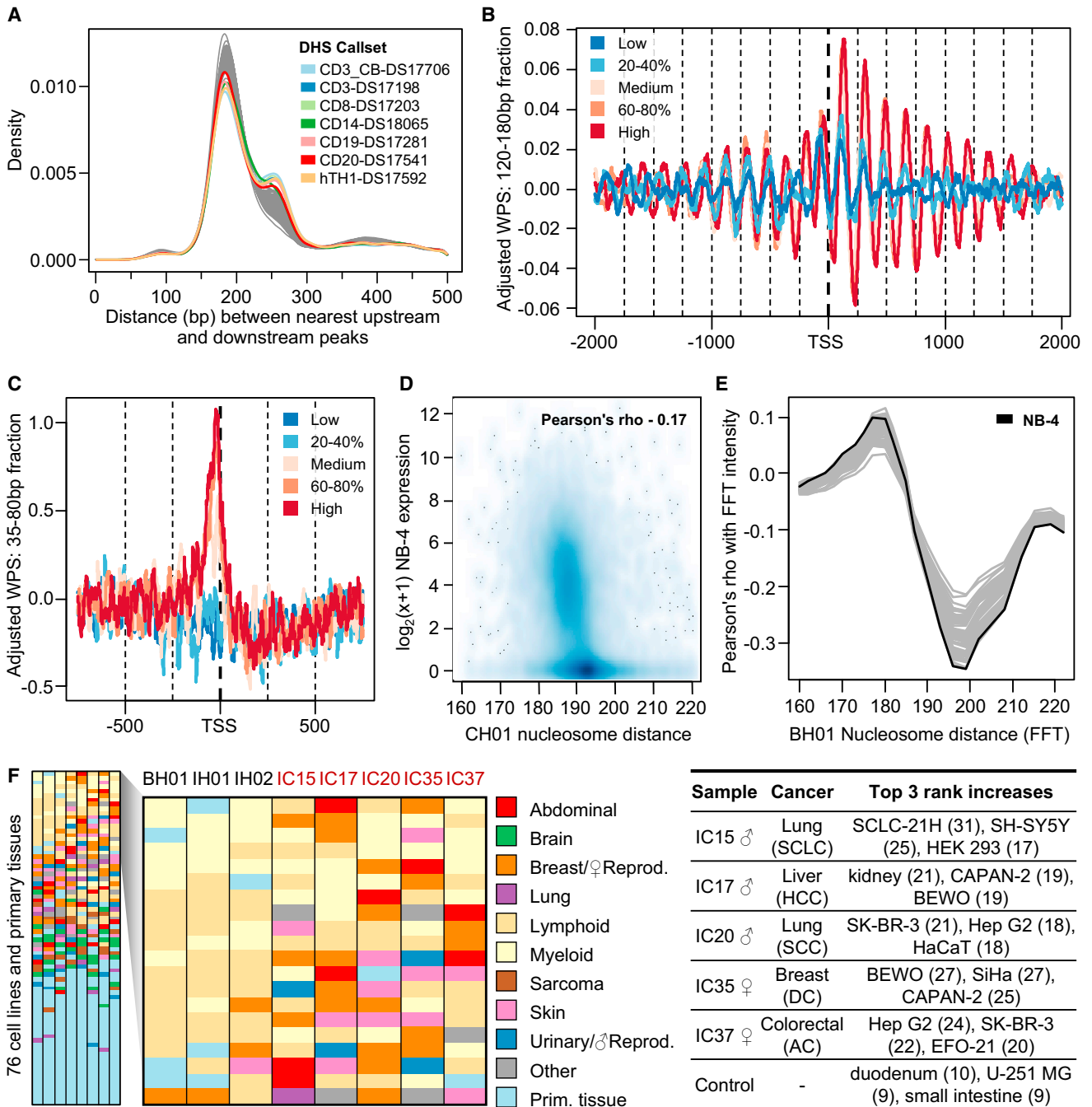
Figure 5. Inference of Mixtures of Cell Types Contributing to cfDNA in Healthy States and Cancer

(A) The distribution of nucleosome spacing for peaks flanking DHS sites in 116 callsets is bimodal, plausibly corresponding to widened nucleosome spacing at active DHS sites due to intervening TF binding (∼190 bp → 260 bp). Lymphoid or myeloid callsets have the largest proportions of DHS sites with widened nucleosome spacing, consistent with hematopoietic cell death as the dominant source of cfDNA in healthy individuals.

(B and C) Partitioning adjusted WPS scores around TSS into five gene expression bins (quintiles) defined for NB-4 (an acute promyelocytic leukemia cell line) reveals differential nucleosome spacing and positioning.

(B) Highly expressed genes show strong nucleosome phasing within the transcript body. Upstream of the TSS, −1 nucleosomes are well-positioned across expression bins, but −2 and −3 nucleosomes are well-positioned only for medium to highly expressed genes.

(C) For medium to highly expressed genes, a short fragment WPS peak is observed between the TSS and the −1 nucleosome, plausibly footprinting some or all of the transcription preinitiation complex at transcriptionally active genes.

(legend continued on next page)

should be emphasized that, in the context of cancer, our goal is not necessarily to outperform the sensitivity of mutation-based monitoring of circulating tumor DNA. Rather, we envision that a unique application of this approach may be to non-invasively classify cancers at time of diagnosis by matching the epigenetic signature of cfDNA fragmentation patterns against reference datasets corresponding to diverse cancer types. For example, this may have value for non-invasively and molecularly classifying "cancers of unknown primary," which comprise 4%–5% of all invasive cancers (Greco and Hainsworth, 2009), as well as cancers where invasive biopsies are currently required for definitive diagnosis and/or for subtyping (e.g., lung cancer).

In addition, there are a range of non-malignant conditions for which it may be valuable to explore the nucleosome and TF footprints contained in cfDNA as markers for acute or chronic tissue damage, e.g., myocardial infarction (Chang et al., 2003), stroke (Rainer et al., 2003), and autoimmune disorders (Galeazzi et al., 2003). Contributions from these tissues to cfDNA cannot be readily detected under the current paradigm of analyzing genotypic differences, which are effectively non-existent in these conditions. By contrast, the approach presented here should generalize to detecting contributions to cfDNA from any non-hematopoietic cell lineage (and, possibly, grossly aberrant contributions from hematopoietic cell lineages).

Alternative "genotype-independent" approaches for using circulating nucleic acids as markers for disease include cell-free RNA (Koh et al., 2014) and DNA methylation. Although these merit exploration, tumor-derived cell-free RNA appears to be much less stable than nucleosome-bound cfDNA (García-Olmo et al., 2013), while bisulfite sequencing libraries are challenging to robustly construct from small amounts of starting material.

A limitation of this study is the small number of samples studied (n = 8) and the relatively small size of the reference dataset of cell lines and tissues against which these samples were compared (n = 76). We anticipate that increasing the number of samples studied as well as the range of physiological states and diseases with which these samples are associated is necessary to fully evaluate the potential and limitations of this approach. Furthermore, expanding the breadth and quality of the reference datasets against which these samples are compared (e.g., directly comparing to cell-type-specific nucleo-

some maps, rather than to expression profiles) is likely to improve the ability to robustly assign and quantify contributing cell types.

Cell-free DNA has tremendous potential as a "liquid biopsy," and indeed, its use in non-invasive prenatal screening for fetal trisomies has vastly outpaced all other applications of DNA sequencing in terms of clinical uptake. In contrast with current paradigms for analyzing cfDNA, we show how the information contained in cfDNA fragmentation patterns, effectively the footprints of protein-DNA interactions, can be used to infer contributing cell types without relying on genotypic differences. To the extent that cfDNA composition is impacted by cell death consequent to malignancy, acute or chronic tissue damage, or other conditions, this method may substantially expand the range of clinical scenarios in which cfDNA sequences comprise a clinically useful biomarker.

EXPERIMENTAL PROCEDURES

Plasma Samples

Bulk human peripheral blood plasma, containing contributions from an unknown number of healthy individuals, was obtained from STEMCELL Technologies. Anonymous, individual human peripheral blood plasma from healthy donors, donors with clinical diagnosis of stage IV cancer, and donors with clinical diagnosis of autoimmune disease (Tables S1, S4, and S5) was obtained from Conversant Bio or PlasmaLab International. Plasma was stored at −80°C and thawed on the benchtop immediately before use. Cell-free DNA was purified from each sample with the QiaAMP Circulating Nucleic Acids kit (QIAGEN) as per the manufacturer's protocol. DNA was quantified with a Qubit fluorometer (Invitrogen).

Preparation of Sequencing Libraries

Conventional, double-stranded sequencing libraries were prepared with the ThruPLEX-FD or ThruPLEX DNA-seq kits (Rubicon Genomics), comprising a proprietary series of end repair, ligation, and amplification reactions. Libraries were prepared with 0.5–30.0 ng of cfDNA input and individually barcoded. Library amplification was monitored by real-time PCR and was typically terminated after four to six cycles. Single-stranded sequencing libraries were prepared with a protocol adapted from Gansauge and Meyer (2013). A detailed description can be found in Supplemental Experimental Procedures.

Sequencing and Primary Data Processing

All libraries were sequenced on HiSeq 2000 or NextSeq 500 instruments (Illumina). Details of sequencing are provided in Tables S1 and S5. Barcoded paired-end (PE) sequencing data were split, allowing up to one substitution

---

(D) Median nucleosome spacing in the transcript body is negatively correlated with gene expression in NB-4 ($\rho = -0.17$, n = 19,677 genes). Spacing in genes with low or no expression is 193 bp, while spacing in expressed genes ranges from 186 to 193 bp.

(E) To deconvolve multiple contributions, intensities from FFT quantified the specific frequency contributions in the long fragment WPS for 10 kb windows downstream of each TSS. Shown are correlation trajectories for RNA expression in 76 cell lines and primary tissues at different frequencies. Correlations are strongest for intensities in the 193–199 bp frequency range.

(F) The ranks of correlation for 76 RNA expression datasets with average intensity in the 193–199 bp frequency range for various cfDNA libraries are shown, categorized by type and listed from highest (top row) to lowest rank (bottom row). Correlation values and full cell line or tissue names are provided in Table S3. All of the strongest correlations for all three healthy samples (BH01, IH01, and IH02; first three columns) are with lymphoid and myeloid cell lines or with bone marrow. In contrast, cfDNA samples obtained from stage IV cancer patients (IC15, IC17, IC20, IC35, IC37; last five columns) show top correlations with various cancer cell lines, e.g., IC17 (hepatocellular carcinoma, HCC) showing highest correlations with HepG2 (HCC cell line), and IC35 (breast ductal carcinoma, DC) with MCF7 (metastatic breast adenocarcinoma cell line). When comparing cell line/tissue ranks observed for the cancer samples to each of the three healthy samples and averaging the rank changes, maximum rank changes are over 2-fold higher than those observed from comparing the three healthy samples with each other and averaging rank changes ("Control"). For example, for IC15 (small cell lung carcinoma, SCLC), the rank of SCLC-21H (SCLC cell line) increased by an average of 31 positions; for IC20 (squamous cell lung carcinoma, SCC), SK-BR-3 (metastatic breast adenocarcinoma cell line) increased by an average rank of 21; and for IC37 (colorectal adenocarcinoma, AC), HepG2 increased by 24 ranks.
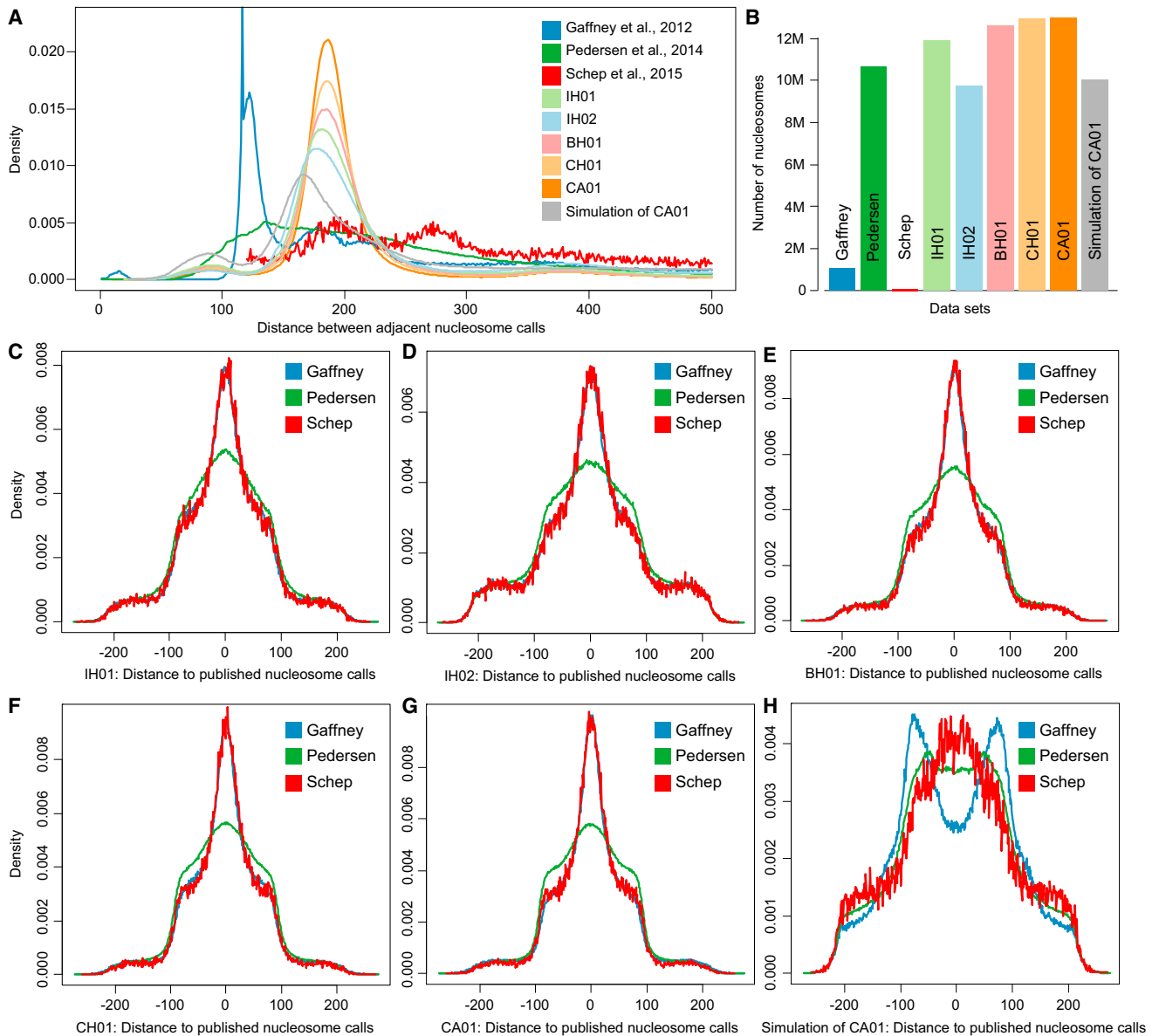
See also Figure S5 and Table S4.

Figure 6. Comparison of Nucleosome Callsets

(A) Distance between nucleosome peak calls across three published datasets (Gaffney et al., 2012; Pedersen et al., 2014; Schep et al., 2015) and calls produced in this study. Previously published callsets lack one defined mode at the canonical ~185 bp nucleosome spacing, possibly due to sparse sampling or wide call ranges. In contrast, all the nucleosome calls from cfDNA show one well-defined mode, the magnitude of which increases with the number of fragments examined. The callset produced from simulation has a lower mode (166 bp) and a wider distribution.

(B) Number of calls in each set. The densest cfDNA-derived callset contains nearly 13 million nucleosome calls.

(C–G) Comparison of peak locations between samples. For each pair of samples, the distribution of distances between each peak call in the sample with fewer peaks and the nearest peak call in the other sample is shown. Negative numbers indicate the nearest peak is upstream; positive numbers indicate the nearest peak is downstream. Concordance between callsets increases with the number of cfDNA fragments examined.

(H), As in (G), comparing other callsets to the matched simulation of CA01.

See also Table S5.

in the barcode sequence. Fragments shorter than or equal to the read length were consensus-called and adaptor-trimmed. Remaining consensus single-end reads (SR) and the individual PE reads were aligned to the human reference genome (GRCh37, 1000 Genomes phase 2 technical reference, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/) using the ALN algorithm in BWA v0.7.10 (Li

and Durbin, 2010). PE reads were further processed with BWA SAMPE to resolve ambiguous placement of read pairs or to rescue missing alignments by a more sensitive alignment step around the location of one placed read end. Aligned SR and PE data were stored in BAM format using the samtools API (Li et al., 2009). BAM files for each sample were merged across lanes and sequencing runs.

Simulations

Sequencing data were simulated procedurally to mimic observed cleavage and ligation biases and length distributions. See Supplemental Experimental Procedures for details.

Coverage, Fragment Endpoints, and Windowed Protection Scores

Fragment endpoint coordinates were extracted from BAM files with the SAMtools API. Both outer alignment coordinates of PE data were extracted for properly paired reads. Both end coordinates of SR alignments were extracted when PE data were collapsed to SR data by adaptor trimming. A fragment's coverage is defined as all positions between the two (inferred) fragment ends, inclusive of endpoints. We define the windowed protection score (WPS) of a window of size k as the number of molecules spanning the window minus those with an endpoint within the window. We assign the determined WPS to the center of the window. For 35–80 bp fragments (short fraction, S-WPS), $k = 16$; for 120–180 bp fragments (long fraction, L-WPS), $k = 120$.

Nucleosome Peak Calling

L-WPS is locally adjusted to a running median of zero in 1 kb windows and smoothed using a Savitzky-Golay filter (Savitzky and Golay, 1964) (window size 21, second-order polynomial). The L-WPS track is then segmented into above-zero regions (allowing up to five consecutive positions below zero). If the resulting region is 50–150 bp, we identify the median L-WPS value of that region and search for the maximum-sum contiguous window above the median. We report the start, end, and center coordinates of this window as the "peak," or local maximum of nucleosome protection. All calculations involving distances between peaks are based on these center coordinates. A score for each peak is determined as the distance between maximum value in the window and the average of the two adjacent L-WPS minima neighboring the region. If the identified region is 150–450 bp, we apply the same above-median contiguous window approach, but only report those windows that are 50–150 bp. For score calculation of multiple windows derived from 150–450 bp regions, we set the neighboring minima to zero. We discard regions <50 bp or >450 bp.

Analysis of TFBS, DHS Sites, and Genic Features

Features were aggregated and aligned at starting coordinates while adjusting for strand and direction of transcription. TFBS sets were obtained by filtering motif predictions with ChIP-seq peaks. For most features, L-WPS values were adjusted to account for signal observed in matched simulations. See Supplemental Experimental Procedures for details.

Gene Expression Analysis

Fragments per kilobase of transcript per million mapped reads (FPKM) gene expression (GE) values measured for 20,344 Ensembl gene identifiers in 44 human cell lines and 32 primary tissues by the Human Protein Atlas (Uhlén et al., 2015) were downloaded from http://www.proteinatlas.org/download/rna.csv.zip. Genes with three or more non-zero expression values were retained (n = 19,378 genes). The GE dataset is provided with one decimal precision for the FPKM values. Thus, a zero GE value (0.0) indicates expression in the interval (0, 0.05) Unless otherwise noted, we set the minimum GE value to 0.04 FPKM before log_2-transformation.

Fourier Transformation and Correlation with Expression

L-WPS was used to calculate periodograms of genomic regions using FFT (spec.pgram in R) with frequencies between 1/500 and 1/100 bases. See Supplemental Experimental Procedures for details. Intensity values for the 120–280 bp frequency range were determined from smooth FFT periodograms. S-shaped Pearson correlation between GE values and FFT intensities was observed around the major inter-nucleosome distance peak, along with a pronounced negative correlation in the 193–199 bp frequency range. The mean intensity in this frequency range was correlated with the average intensity with log_2-transformed GE values for downstream analysis.

ACCESSION NUMBERS

The accession number for the aligned reads of the 60 samples as well as the peak calls reported in this paper is NCBI GEO: GSE71378.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.11.050.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. 19, 1732–1741.

Chan, A.K.C., Chiu, R.W.K., and Lo, Y.M.D.; Clinical Sciences Reviews Committee of the Association of Clinical Biochemists (2003). Cell-free nucleic acids in plasma, serum and urine: a new tool in molecular diagnosis. Ann. Clin. Biochem. 40, 122–130.

Chang, C.P.-Y., Chia, R.-H., Wu, T.-L., Tsao, K.-C., Sun, C.-F., and Wu, J.T. (2003). Elevated cell-free serum DNA detected in patients with myocardial infarction. Clin. Chim. Acta 327, 95–101.

Chiu, R.W.K., Chan, K.C.A., Gao, Y., Lau, V.Y.M., Zheng, W., Leung, T.Y., Foo, C.H.F., Xie, B., Tsui, N.B.Y., Lun, F.M.F., et al. (2008). Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc. Natl. Acad. Sci. USA 105, 20458–20463.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., et al. (2010). Relationship between nucleosome positioning and DNA methylation. Nature 466, 388–392.

Diaz, L.A., Jr., and Bardelli, A. (2014). Liquid biopsies: genotyping circulating tumor DNA. J. Clin. Oncol. 32, 579–586.

Fan, H.C., Blumenfeld, Y.J., Chitkara, U., Hudgins, L., and Quake, S.R. (2008). Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc. Natl. Acad. Sci. USA 105, 16266–16271.

Fleischhacker, M., and Schmidt, B. (2007). Circulating nucleic acids (CNAs) and cancer–a survey. Biochim. Biophys. Acta 1775, 181–232.

Fu, Y., Sinha, M., Peterson, C.L., and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. 4, e1000138.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of nucleosome positioning in the human genome. PLoS Genet. *8*, e1003036.

Galeazzi, M., Morozzi, G., Piccini, M., Chen, J., Bellisai, F., Fineschi, S., and Marcolongo, R. (2003). Dosage and characterization of circulating DNA: present usage and possible applications in systemic autoimmune disorders. Autoimmun. Rev. *2*, 50–55.

Gansauge, M.-T., and Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat. Protoc. *8*, 737–748.

García-Olmo, D.C., Picazo, M.G., Toboso, I., Asensio, A.I., and García-Olmo, D. (2013). Quantitation of cell-free DNA and RNA in plasma during tumor progression in rats. Mol. Cancer *12*, 8.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Greco, F.A., and Hainsworth, J.D. (2009). Introduction: unknown primary cancer. Semin. Oncol. *36*, 6–7.

Harshman, S.W., Young, N.L., Parthun, M.R., and Freitas, M.A. (2013). H1 histones: current perspectives and challenges. Nucleic Acids Res. *41*, 9593–9609.

Holdenrieder, S., Stieber, P., Chan, L.Y.S., Geiger, S., Kremer, A., Nagel, D., and Lo, Y.M.D. (2005). Cell-free DNA in serum and plasma: comparison of ELISA and quantitative PCR. Clin. Chem. *51*, 1544–1546.

Koh, W., Pan, W., Gawad, C., Fan, H.C., Kerchner, G.A., Wyss-Coray, T., Blumenfeld, Y.J., El-Sayed, Y.Y., and Quake, S.R. (2014). Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. Proc. Natl. Acad. Sci. USA *111*, 7361–7366.

Leary, R.J., Sausen, M., Kinde, I., Papadopoulos, N., Carpten, J.D., Craig, D., O'Shaughnessy, J., Kinzler, K.W., Parmigiani, G., Vogelstein, B., et al. (2012). Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci. Transl. Med. *4*, 162ra154.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lo, Y.M., Zhang, J., Leung, T.N., Lau, T.K., Chang, A.M., and Hjelm, N.M. (1999). Rapid clearance of fetal DNA from maternal plasma. Am. J. Hum. Genet. *64*, 218–224.

Lo, Y.M.D., Chan, K.C.A., Sun, H., Chen, E.Z., Jiang, P., Lun, F.M.F., Zheng, Y.W., Leung, T.Y., Lau, T.K., Cantor, C.R., and Chiu, R.W. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci. Transl. Med. *2*, 61ra91.

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

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Mouliere, F., El Messaoudi, S., Pang, D., Dritschilo, A., and Thierry, A.R. (2014). Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. Mol. Oncol. *8*, 927–941.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Pedersen, J.S., Valen, E., Velazquez, A.M.V., Parker, B.J., Rasmussen, M., Lindgreen, S., Lilje, B., Tobin, D.J., Kelly, T.K., Vang, S., et al. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. Genome Res. *24*, 454–466.

Rainer, T.H., Wong, L.K.S., Lam, W., Yuen, E., Lam, N.Y.L., Metreweli, C., and Lo, Y.M.D. (2003). Prognostic use of circulating plasma nucleic acid concentrations in patients with acute stroke. Clin. Chem. *49*, 562–569.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Sambuy, Y., De Angelis, I., Ranaldi, G., Scarino, M.L., Stammati, A., and Zucco, F. (2005). The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics. Cell Biol. Toxicol. *21*, 1–26.

Savitzky, A., and Golay, M. (1964). Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. *36*, 1627–1639.

Schep, A.N., Buenrostro, J.D., Denny, S.K., Schwartz, K., Sherlock, G., and Greenleaf, W.J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res. *25*, 1757–1770. http://dx.doi.org/10.1101/gr.192294.115.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Snyder, T.M., Khush, K.K., Valantine, H.A., and Quake, S.R. (2011). Universal noninvasive detection of solid organ transplant rejection. Proc. Natl. Acad. Sci. USA *108*, 6229–6234.

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. Nat. Struct. Mol. Biol. *19*, 1185–1192.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

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.

Vierstra, J., Wang, H., John, S., Sandstrom, R., and Stamatoyannopoulos, J.A. (2014). Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. Nat. Methods *11*, 66–72.

Wimberger, P., Roth, C., Pantel, K., Kasimir-Bauer, S., Kimmig, R., and Schwarzenbach, H. (2011). Impact of platinum-based chemotherapy on circulating nucleic acid levels, protease activities in blood and disseminated tumor cells in bone marrow of ovarian cancer patients. Int. J. Cancer *128*, 2572–2580.