Long-read sequencing reveals aberrant fragmentation patterns and origins of circulating DNA in cancer

Benjamin P. Berman[1,2,*], Sarah A. Erdman[1,*], Jean-Valery Turatsinze[3], Justin Cayford[1], Theresa K. Kelly[1]
1. Volition America LLC, Henderson, Nevada, United States of America
2. Department of Developmental Biology and Cancer Research, The Hebrew University of Jerusalem, Jerusalem, Israel
3. Belgian Volition SRL, Parc Scientifique Crealys, Isnes, Belgium
* Equal contribution
Correspondence to: b.berman@volition.com, t.kelly@volition.com

Abstract

Circulating cell-free DNA (cfDNA), which includes tumor and immune-derived fragments, is often elevated in cancer patients relative to healthy individuals. This can be accompanied by changes in cfDNA fragmentation patterns, including fragment length distributions, fragment end sequences, and genomic context. Here, we survey fragmentation changes across 12 cancer types using Oxford Nanopore Technologies (ONT) shallow whole-genome sequencing. We confirm a hyperfragmentation pattern across a large fraction of the cancers and associate this with markers of altered DNase activity and elevation of circulating DNA and nucleosome levels. We also identify a cluster of cancers with fragments greater than 1 kilobase and distinguish these long fragments from genomic contamination based on length distribution and a DNASE1L3 fragmentation signature. Future studies using ONT sequencing will determine the prevalence and implications of this hypofragmentation phenotype across cancer.

Introduction

Circulating cell-free DNA (cfDNA) is used as a biomarker because it allows us to detect and monitor cell turnover and cell death during normal physiological processes and disease[1]. In oncology, circulating tumor DNA (ctDNA) derived from tumor cells can be detected and characterized for screening, diagnosis, treatment selection, prognosis, and monitoring [2,3]. Typically, ctDNA represents a minority of circulating cfDNA in cancer patients; the remainder primarily comes from blood cells [4]. Several cfDNA features can be used as biomarkers for early cancer detection, including total quantity, fragmentation patterns, single-nucleotide variants, copy number alterations and DNA methylation [5,6]. High levels of circulating DNA or nucleosomes have been used as a biomarker for cancer, although a significant portion of the excess cfDNA in cancer patients originates from immune cell types [4]. Cancer cfDNA remains bound to nucleosomes, allowing detection by immunoassays and immunoprecipitation to unmodified or modified histones [7–10].

DNA fragmentation patterns or “fragmentomics” can detect a number of informative features associated with cancer cfDNA [11,12], including fragment lengths [13–15], fragment end motif sequences [16,17] and their DNA methylation states [18], gene regulatory elements [19,20], and long range chromatin domain compaction in specific chromosomal domains [21]. Fragmentation patterns are shaped by the interaction of cell-type specific chromatin features and the specific DNAse enzymes that act on them intracellularly and in circulation [22].

While fragmentomic features of ctDNA are of great interest, much of our understanding comes from Illumina short-read sequencing. Recently, amplification-free sequencing of longer circulating DNA fragments in cancer has been demonstrated using Oxford Nanopore Technologies (ONT) and Pacific Biosciences sequencing [23–28]. While these technologies broadly recapitulate fragmentomic and other features associated with cancer [25], they also provide an opportunity to expand our understanding and re-examine earlier findings. Some of these studies identified cfDNA fragments much longer than those that can be identified using short-read technologies, and have linked these long fragments to fetal DNA, pre-eclampsia, and cancer [24,27,29]. In this study, we use ONT sequencing to survey a diverse set of cancer types to understand the range of fragmentomic changes that occur in cancer.

Results

H3.1 nucleosome levels and cfDNA are elevated across a set of 207 Stage III-IV cancer cases from 14 cancer types

Cell-free DNA circulates as nucleosomes rather than protein-free DNA. As such, we measured H3.1 nucleosome concentration directly from plasma (Nu.Q® assay), and extracted and quantified DNA from a common set of 207 cancer samples from 14 cancer types and 52 samples from healthy donors. The two measurements were highly correlated (R^2=0.82) (Figure 1A) and were similarly elevated in cancers relative to healthy samples (Figure 1B-G).

In order to estimate the percentage of tumor DNA, we used Copy Number Alterations (CNAs) by performing Illumina whole-genome sequencing at a genomic coverage of 5-50 million filtered read pairs (median 17 million read pairs, Supplementary Figure 1A). We used the tool ichorCNA [30], and were able to detect significant CNAs in 23 of the 207 cancer cases, with no false positives among the 52 healthy samples. Based on these, ichorCNA was able to estimate the fraction of tumor DNA in these samples samples. The tumor DNA fraction was never above 0.6, and only weakly correlated with either H3.1 nucleosome level (Figure 1H) or total DNA level (Figure 1I). By multiplying the tumor DNA fraction from ichorCNA by the total H3.1 nucleosome level, we were able to estimate the absolute “cancer” component of the H3.1 level (Figure 1J), which agreed with previous findings that high levels of circulating DNA in cancer do not primarily come from cancer cells [4]. The same was true for total DNA concentrations (Figure 1K).

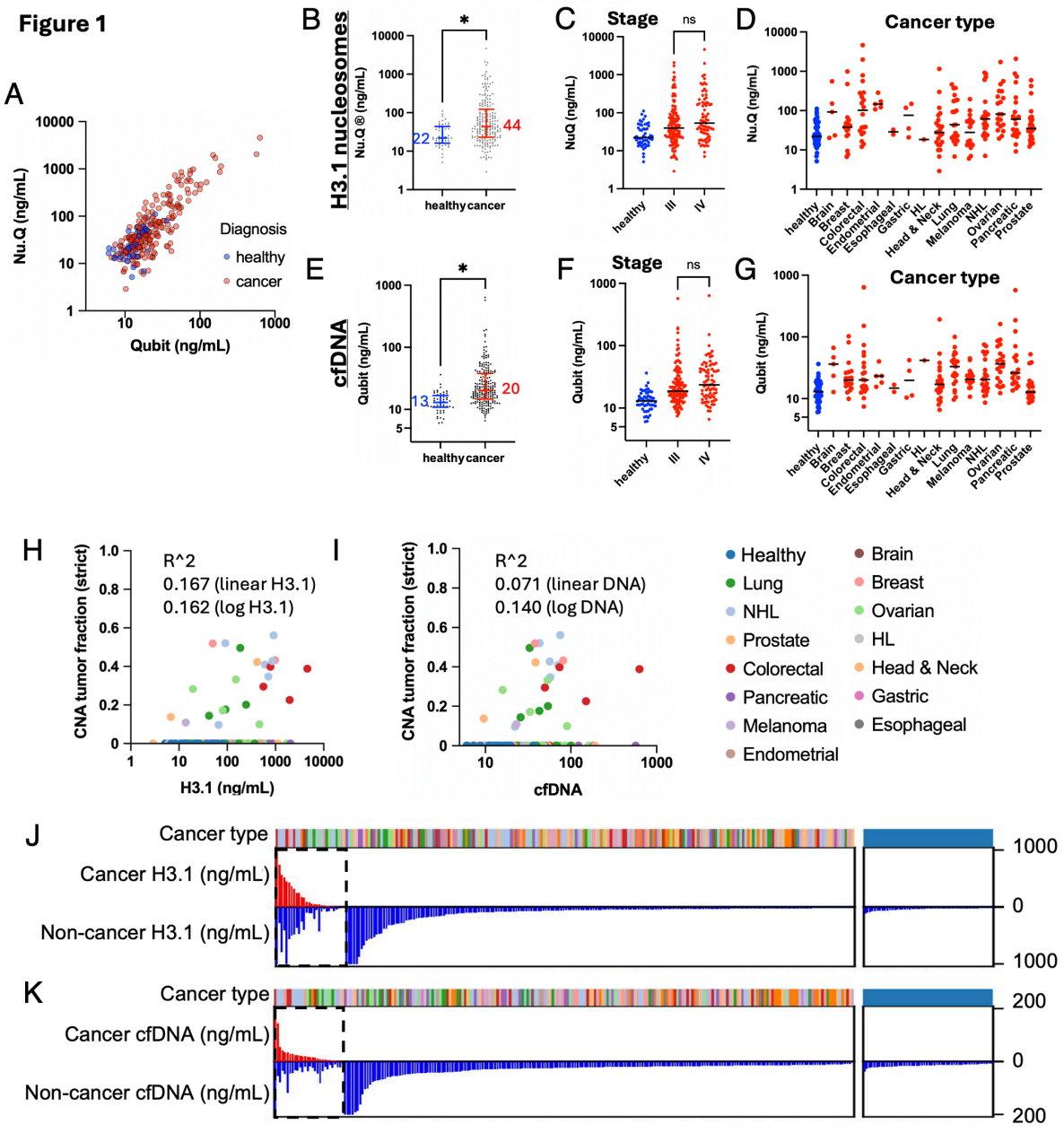


Figure 1: Comparison of total DNA and nucleosome levels vs. cancer cell fraction for 207 Stage III-IV cancer cases from 14 cancer types (A) DNA concentration (Qubit) and H3.1 nucleosome concentrations (Nu.Q®) for all samples. (B) H3.1 nucleosome concentrations by cancer status, with median values specified. H3.1 nucleosome concentrations by stage (C), and by cancer type (D). (E-G) cfDNA concentrations by cancer status (E), stage (F), and cancer type (G). (H) H3.1 nucleosome levels vs. tumor fractions estimated from ichorCNA analysis. (I) cfDNA levels vs. tumor fractions. (J) Absolute concentration of cancer-derived and non-cancer derived H3.1 nucleosomes, as determined by total concentration times tumor fraction. Cancer and non-cancer samples are shown separately, and each is sorted by the absolute concentration of cancer-derived H3.1 for samples with CNAs detected, and by non-cancer-derived concentration of H3.1 for the rest. (K) same as J but using absolute cfDNA levels. P-values in B,C,E, and F were calculated using a two-tailed unpaired t-test, * signifies $p<0.05$ ($p=0.020$ for B and $p=0.013$ for E).

Nanopore sequencing detects cancer cell of origin by CNA and methylation Cell of Origin (COO)

To survey a variety of cancer types using ONT sequencing, we selected a subset of 37 patients from 12 cancer types from the Illumina sequencing study, along with 21 healthy samples. Each sample was sequenced to a depth of 1-16 million reads (median=4.7 million) using ONT MinION 9.4.1 flowcells, LSK109 library chemistry (Supplementary Figure 1B). We first used ichorCNA, which detected cancer DNA in 16 of the 37 cancer samples (Figure 2A). The ichorCNA-estimated tumor fraction determined from ONT and Illumina based sequencing were highly concordant, although CNAs were detected in 5 additional patients using ONT, all with tumor fractions estimated at less than 0.2 (Figure 2B). This 45% increase in CNA detection was despite the higher median sequence coverage in Illumina sequencing (median 17M paired reads in Illumina vs. 5M fragments in ONT, Supplementary Figure 1).

Using methylation information from ONT reads and a methylation atlas of normal cell types, we performed cell of origin (COO) deconvolution [25,31]. This identified the correct cancer COO for 11 of the 16 CNA-positive samples, and 1 CNA-negative sample (Figure 2C-D). Importantly, the absolute fraction of the cancer derived DNA was never above 50%, consistent with CNA analysis and indicating that cancer DNA was not the major contributor of elevated cfDNA levels in the samples studied (Supplementary Figure 2). Five Acute Myeloid Leukemia samples had no CNAs but did have elevated granulocyte derived DNA, which could not be unambiguously assigned to cancer or non-cancer origin given the common myeloid lineage. The remaining 15 samples had neither detection of CNAs nor cancer derived DNA, suggesting low cancer fractions despite many having high cfDNA and H3.1 nucleosome levels (Figure 2C).

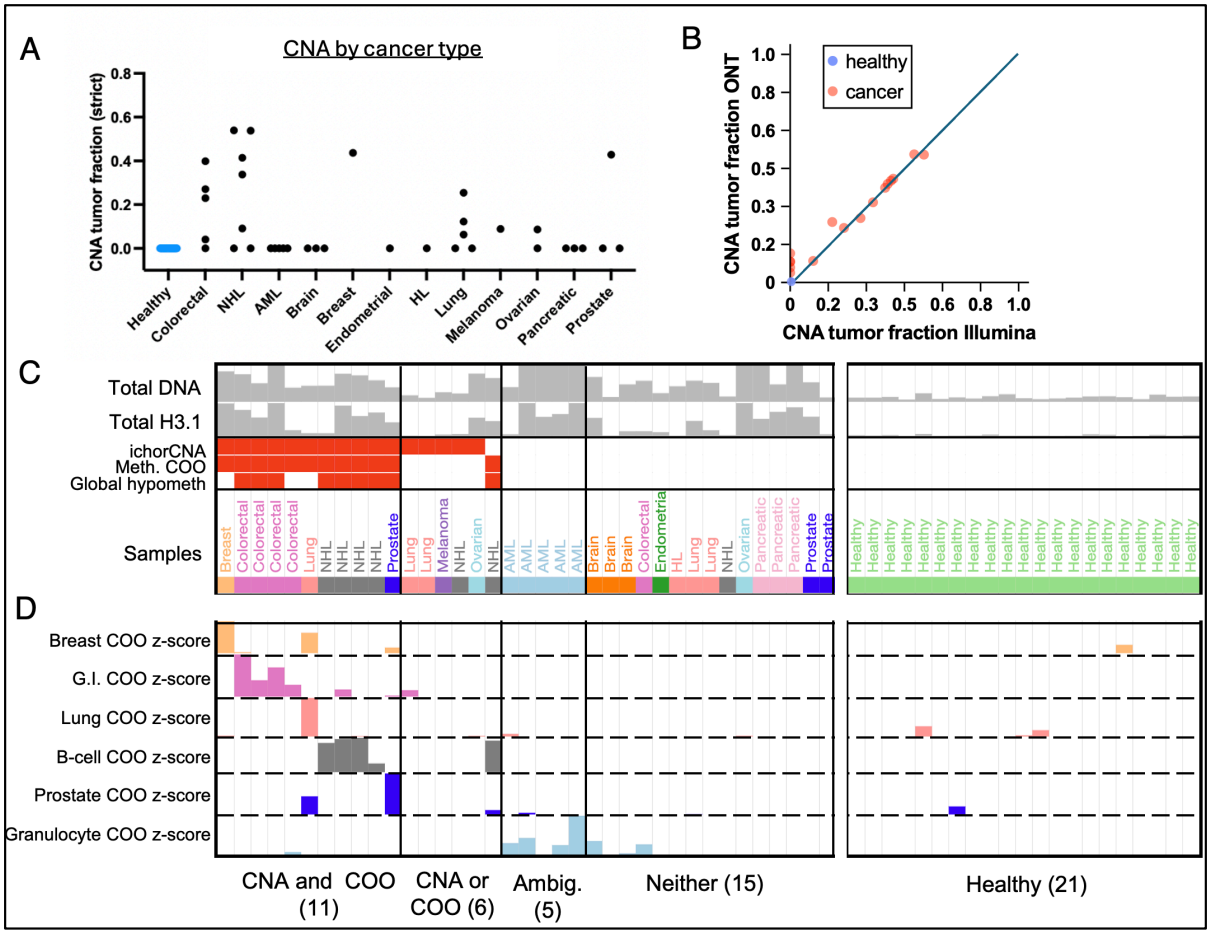


Figure 2: Nanopore sequencing detects cancer DNA by Copy Number Alterations and methylation Cell of Origin (COO) (A) ichorCNA tumor fractions for 37 samples sequenced using Oxford Nanopore Technologies (ONT). (B) Comparison of ichorCNA-estimated tumor fractions derived from Illumina sequencing (x axis) and ONT (y axis). 5/16 cancer samples with detectable CNA in ONT (y axis>0) were not detectable by Illumina (x axis=0). (C) Comparison of ONT ichorCNA cancer detection and DNA methylation-based cancer detection. “ichorCNA” indicates whether CNAs were detected by ichorCNA, and Methylation Cell of Origin (“Meth. COO”) indicates that the correct cancer cell of origin had the highest normalized cell type proportion, as shown below for the six relevant cell types. “Global hypometh” indicates that the sample had significantly lower genome-wide methylation than healthy samples. (D) Details of COO analysis showing z-score normalized contributions for six relevant cell types in the analysis.

Identification of ultra-long sequencing reads originating from genomic contamination

ONT sequencing can analyze long cfDNA fragments, but some of these may represent contamination from genomic DNA released as part of sample processing, especially from cell lysis during freezing and thawing of samples [32,33]. Our standard processing protocol includes a post-thaw high-speed spin to reduce the number of such fragments (Figure 3A). Despite this, three healthy samples had unusually high fractions of ultra-long fragments > 7.5kb in length (Figure 3B). In order to understand the origin of these fragments, we compared aliquots of the same samples with and without the high-speed spin step (Figure 3A, “unspun” protocol). Following centrifugation, the 3 outlier samples showed only slight differences in total H3.1 nucleosome or DNA levels compared to other healthy samples (Figure 3C-D, black bars), whereas the unspun samples had markedly increased levels of both (Figure 3C-D, red bars). Fragment length distributions from ONT sequencing showed that the unspun samples had significantly more ultra-long reads >7.5kb than the post-spin samples (one example shown in Figure 3E). Illumina sequencing yielded no ultra-long fragments, as expected (Figure 3F).

Fragment length distributions in Figure 3E-F represent the proportion of the total *DNA base pairs* contained in fragments of a given size, which is meant to reflect the proportion of cells from which the cfDNA originated. However, previous studies quantified the fraction of long fragments rather than base pairs in those fragments [27,28]. To directly compare our results to these earlier studies, we quantified the proportion of DNA fragments greater than 500, 1000, or 7,500 bp before and after the high-speed spin. In each case, the high-speed spin decreased the fraction of long fragments significantly (Figure 3G-I).

cfDNA fragments disproportionally start and end with the nucleotide motif CCCA and other motifs beginning with “CC”, which is the result of DNASE1L3 activity in circulation [16,34]. To assess whether the ultra-long fragments in our samples were the product of DNASE1L3-associated fragmentation, we analyzed CC end motif frequencies in the three outlier samples along with 4 typical healthy samples (Figure 3J). Typical healthy volunteers had very few fragments over 1kb to analyze, but those that did had typically high CC end motif levels (Figure 3J, black lines). Both the pre- and post- spin samples for the three outlier healthy volunteers had typically high CC end motif levels up to about 500bp, which decreased in larger fragments (Figure 3J, blue and red lines). We confirmed this for the most specific motif, CCCA (Figure 3K). While we cannot confirm in these samples that these ultra-long fragments were generated post-thawing, this is consistent with what has been observed in earlier studies [32,33].

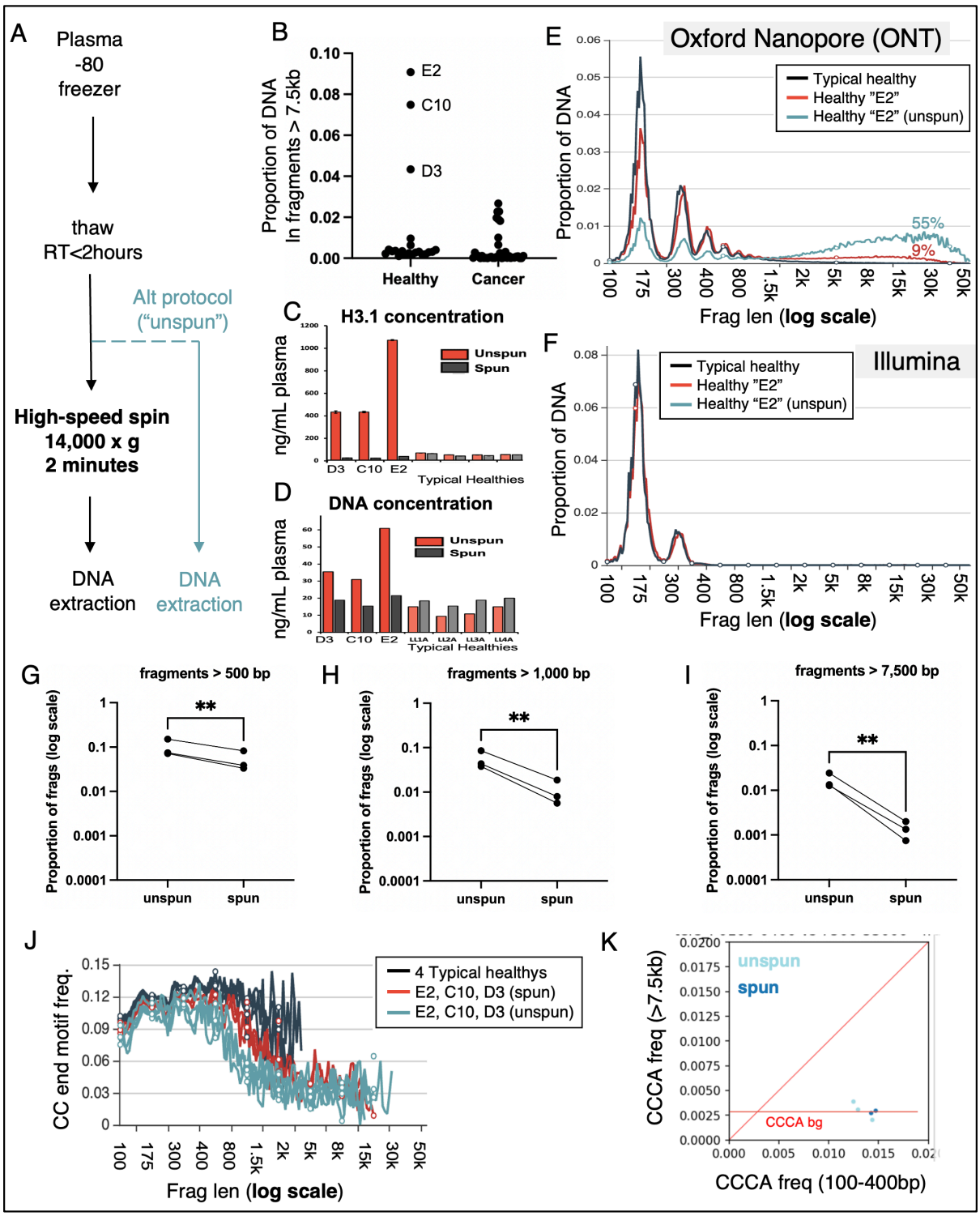

A
Plasma
-80
freezer
thaw
RT<2hours
Alt protocol
("unspun")
High-speed spin
14,000 x g
2 minutes
DNA
extraction
DNA
extraction
B
Proportion of DNA
In fragments > 7.5kb
E2
C10
D3
Healthy
Cancer
C
H3.1 concentration
ng/mL plasma
Unspun
Spun
D3 C10 E2
Typical Healthies
D
DNA concentration
ng/mL plasma
Unspun
Spun
D3 C10 E2 LL1A LL2A LL3A LL4A
Typical Healthies
E
Oxford Nanopore (ONT)
Proportion of DNA
Typical healthy
Healthy "E2"
Healthy "E2" (unspun)
55%
9%
Frag len (log scale)
F
Illumina
Proportion of DNA
Typical healthy
Healthy "E2"
Healthy "E2" (unspun)
Frag len (log scale)
G
fragments > 500 bp
Proportion of frags (log scale)
**
unspun
spun
H
fragments > 1,000 bp
Proportion of frags (log scale)
**
unspun
spun
I
fragments > 7,500 bp
Proportion of frags (log scale)
**
unspun
spun
J
CC end motif freq.
4 Typical healthys
E2, C10, D3 (spun)
E2, C10, D3 (unspun)
Frag len (log scale)
K
unspun
spun
CCCA bg
CCCA freq (>7.5kb)
CCCA freq (100-400bp)

Figure 3 (prev. page): Identification of ultra-long sequencing reads derived from genomic contamination. (A) Typical "spun" plasma processing workflow and alternative "unspun" workflow. (B) All ONT samples, showing proportion of DNA in fragments greater than 7.5kb in length. (C) H3.1 concentration and (D) cfDNA concentration for 3 outlier healthy samples and 4 other healthy samples. (E) Proportion of DNA as a function of fragment length, for ONT sequencing of a typical healthy sample (black) and the "E2" healthy sample processed with the spun (red) and unspun (blue) workflows. (F) Proportion of DNA as a function of fragment length, for Illumina sequencing of the typical healthy sample (black) and the "E2" healthy sample processed with the spun (red) workflow. (G-I) Proportion of fragments greater than 500bp (G), 1,000 bp (H) and 7,500 bp (I). (J) Proportion of fragment ends starting with CC in ONT sequencing, as a function of fragment length, for 4 typical healthy volunteers (black) and the three outlier healthy volunteers processed with the spun workflow (red) and the unspun workflow (blue). (K) The proportion of fragment ends starting with CCCA in ONT sequencing, in typical length fragments (100-400bp, x axis) vs. fragments greater than 7.5kb (y axis). In (G-I), ** indicates $p<0.002$ by ratio paired t-test, two-tailed.

Unsupervised clustering by fragment length identifies both hypofragmented and hyperfragmented cancer classes

In order to assess fragment length differences systematically, we performed Principal Component Analysis (PCA) unsupervised clustering on all samples based on fragment length profiling, and used the values of the first 3 principal components (PC1-3, Supplemental Figure 3A-C) to define 3 groups of cancers and 3 groups of healthy samples (Figure 4A). The first group (PC1-high) contained seven cancers with severely hypofragmented cfDNA, which was highly elevated in the range 900-4300bp (Figure 5A, unnormalized heatmap left, and column normalized heatmap right). The cancers in this group had between 20-66% of all cfDNA in fragments 900-4300bp (Figure 5A, "Frac DNA in PC1"). Interestingly, the PC1-high group included all five Acute Myeloid Leukemias in our study, as well as one Lung cancer and one Non-Hodgkins Lymphoma (NHL) case. An additional Hodgkin's Lymphoma (starred) was weakly positive for the PC1 signature, indicating a tendency for hematopoietic cancers to display this fragmentation profile. Furthermore, these fragments show nucleosomal phasing in cut preference, as indicated by the laddered appearance of the fragment length distribution, as well as CC end motif levels (Figure 4B), consistent with fragmentation by DNASE1L3 in circulation rather than genomic contamination.

The hypofragmented cancer group (PC1-high) was strongly distinguished from the 3 healthy samples with likely genomic contamination, which are high in PC2 (Figure 4A bottom). PC2 is defined by a size range of 7,500-53,000 bp. Consistent with our earlier analysis, the post-spin samples had much lower PC2 scores than the unspun samples (Figure 4A, PC scores), and much less total cfDNA in the 7,500-53,000 bp bin (Figure 4A "Frac DNA", 4-9% for post-spin samples and 38-48% for pre-spin samples). Aside from

these three sample groups, all other healthy volunteers and cancers had low values for PC1 and PC2 (Figure 4A, “PC scores”, and Supplementary Figure 3A-B).

The third principal component (PC3) defined a hyperfragmented set of 17 cancers (Figure 4C, “PC3-high cancers”). These had elevated cfDNA in the range 75-145 bp and 245-295 bp (corresponding to short mono-nucleosomes and short di-nucleosomes, respectively). Longer fragments, including those greater than 1kb, were under-represented. Similar patterns have been described previously both with Illumina sequencing [14,15] and ONT sequencing [25,26]. Interestingly, PC3-high cancers tended to have lower frequencies of the DNASE1L3-associated CC end motifs (Figure 4C).

The PCA analysis above included the 3 unspun samples and to ensure that these were not influencing the clustering, we performed PCA with these samples excluded. This produced nearly identical versions of PC1 and PC3 (as the first two components), and identical grouping of all cancer samples (Supplementary Figure 3D-G).

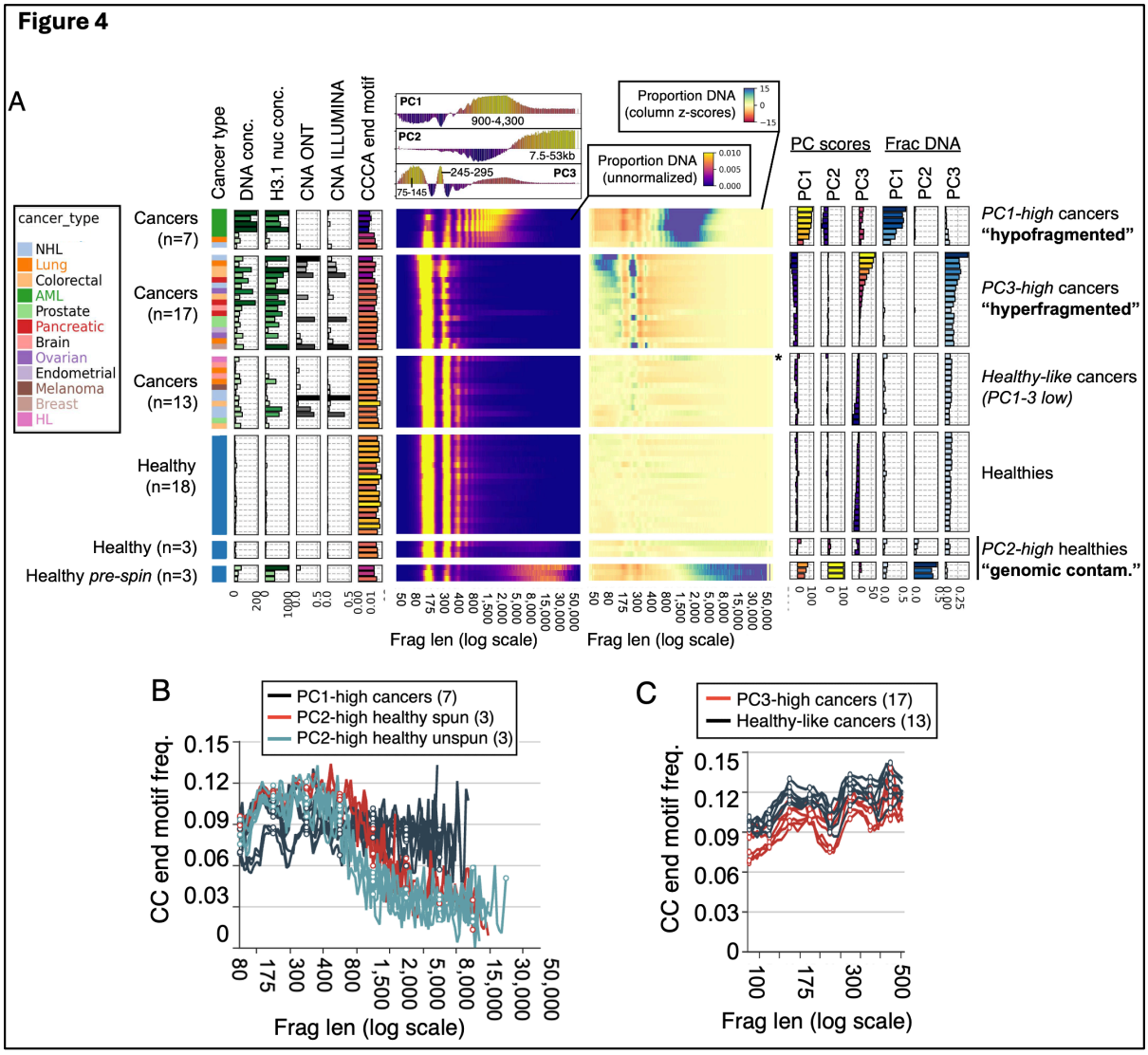


Figure 4. Unsupervised clustering by fragment length. (A) All samples ordered by first three Principal Components (PCs). The first 3 sample groups are cancer samples with the cancer type indicated by color, and the second 3 groups are healthy samples. Bar plots on the left show total cfDNA concentation (ng/mL), total plasma H3.1 nucleosome concentration (ng/mL), tumor DNA fraction from ichorCNA from ONT sequencing ("CNA ONT") and Illumina sequencing ("CNA ILLUMINA"), and the fraction of ONT fragments ending with the motif CCCA. In all bar plots, higher bars are also indicated with darker colors. The heatmap to the right of these bar plots shows the proportion of DNA sequenced in different fragment length bins, on a log scale from 50 bp to 50,000 bp. Above this heatmap are the PC loadings for each of the first 3 PCs. The heatmap to the right shows the same data, but each fragment length bin/column is z-score normalized according to the mean and standard deviation across all healthy samples (not including "pre-spin" samples). The bar plots on the right show the eigenvalues for the 3 principal components ("PC scores"), and the percentage of sequenced DNA in the ranges defined the by those PCs ("Frac DNA"). One sample with a sub-threshold but positive PC1 value is marked with an asterisk. (B) CC end motif frequencies as a function of fragment length, for the PC1 high and PC2 high samples. (C) CC end motif frequencies as a function of fragment length, for the 6 cancer samples with the highest PC3 scores, and the 6 cancer samples with the lowest PC3-scores.

Links between hyperfragmentation and elevated circulating chromatin levels in cancer

Hyperfragmentation occurred in 9 of the 12 cancer types in our ONT cohort. While this phenomenon has been documented in cancer [13–15,26], its significance remains poorly understood. We tested whether a correlation analysis based on our PCA-based ONT signature could provide additional insight. We first excluded the 7 PC1-high hypofragmented cancers, leaving 17 PC3-high cancers and 13 PC3-low cancers. We calculated pairwise correlations between PC3 values and 6 other tumor features and used these to perform hierarchical clustering of the features which resulted in three major clusters (Figure 5A-B).

The PC3 value clustered tightly with the fraction of fragments less than 150 bp, a simple but commonly used definition for short fragments [15]. These two formed a single cluster with the DNASE1L3-associated CCCA end motif (Figure 5A-B, blue cluster and example scatterplots in Figure 5C-D). The clustering of hyperfragmentation and the CCCA end motif is significant given that both features were previously linked to DNASE1L3 activity [16,34,35].

The two metrics that directly detect cancer DNA were also tightly clustered (Figure 5A-B, red cluster). However, these were not well correlated to the PC3/DNASE1L3 (blue) cluster – the average pairwise correlation between features in the red and blue was only 0.17 (Figure 5B, and example scatterplots in Figure 5E-F and Supplementary Figure 4). To specifically measure whether shortened fragments were enriched for cancer-derived DNA, we calculated CNA tumor fraction and methylation cell of origin directly in the shorter fragments associated with PC3. Of the 10 PC3-high cancers with sufficient read coverage, only 2 had measurable enrichment for cancer DNA, and neither was more than 1.5-fold enriched (Supplementary Figure 5).

H3.1 nucleosome and cfDNA concentration also formed a tight cluster (Figure 5A-B, green cluster). In contrast to the features related to cancer fraction, these were moderately correlated with the PC3/DNASE1L3 (blue) cluster (average pairwise correlation 0.43), and indeed, formed a supercluster (Figure 5A-B). While cancers with low PC3 hyperfragmentation values had variable DNA and nucleosome concentrations, those with high PC3 values almost always had elevated concentrations (Figure 5G-H). This suggests a potential link between DNASE1L3-associated hyperfragmentation and the abnormally high levels of immune-derived cfDNA observed in cancer [4].

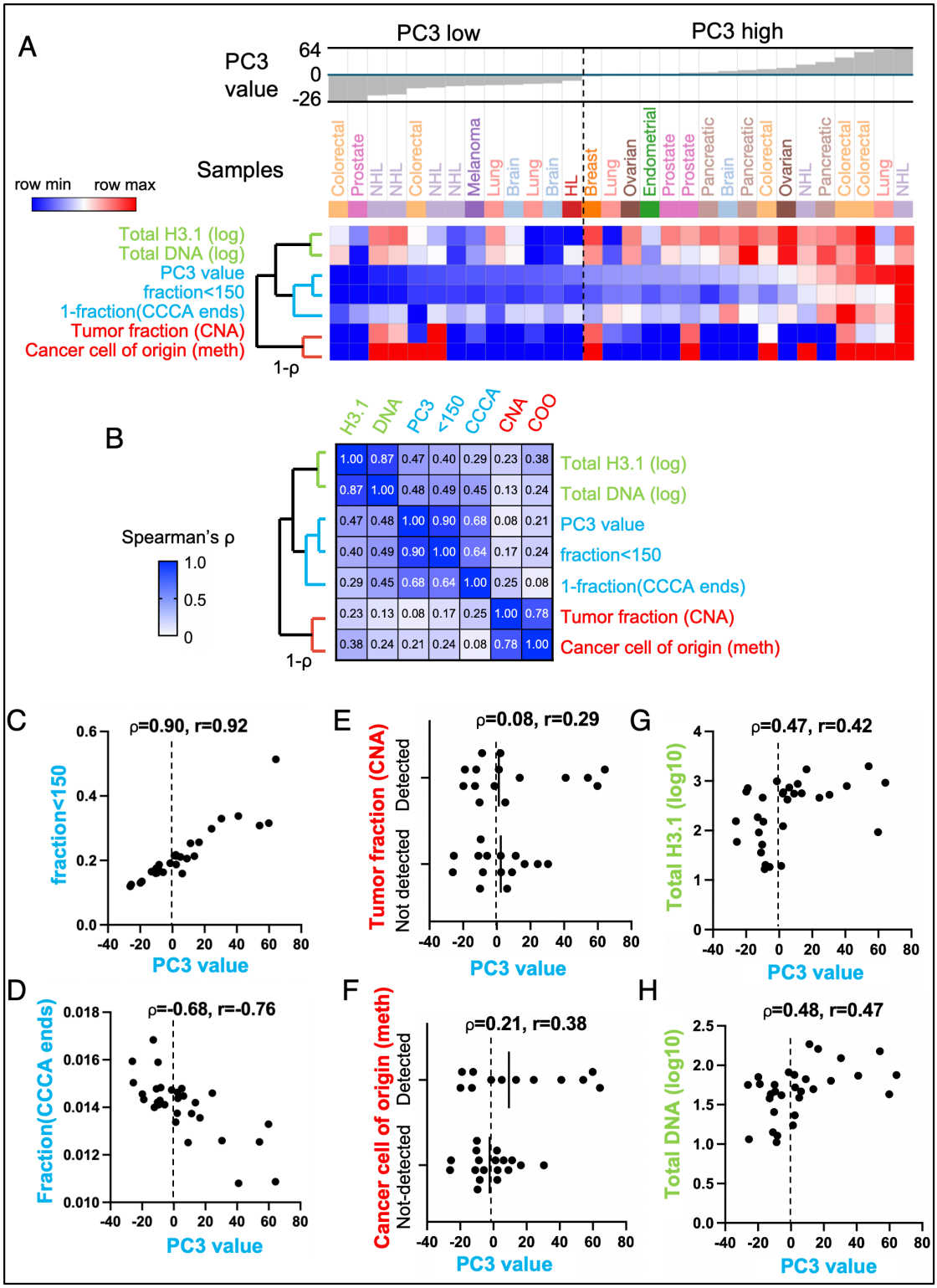


Figure 5: Links between hyperfragmentation and elevated circulating chromatin levels in cancer (A) 30 cancer samples from the PC3-high and PC1-3 low groups, ordered by PC3 value. Heatmap rows show sample features clustered by hierarchical clustering based on 1- ρ (Spearman's rho). Heatmap rows are normalized to row min and max values. (B) Spearman correlation values for all pairwise comparisons. (C-H) PC3 values plotted against the following features: (C) fraction of short fragments, (D) fraction of fragment ends with CCCA motif, (E) detection by ichorCNA, (F) detection by cancer methylation cell of origin, (G) H3.1 nucleosomes by Nu.Q®, and (H) total cfDNA by Qubit.

Discussion

Here, we performed a survey of cancer types using Illumina and Oxford Nanopore (ONT) sequencing, discovering new fragmentation changes that occurred across diverse cancer types. While this study was not powered to investigate heterogeneity within individual cancer types or links to clinical outcomes, we discuss below a number of observations which should be followed up in larger studies.

While several studies have used ONT and Pacific Biosciences sequencing to investigate biological properties of longer fragments >500bp [24,26–29], the issue of genomic contamination has not been addressed. Previous work has shown that genomic contamination occurs as longer fragments and can depend on pre-analytical processing, especially in cryo-preserved samples [32,33]. Here, we showed that the abundance of ultra-long fragments (greater than 7,500bp) was strongly influenced by high-speed centrifugation after thawing, and that such fragments could remain even after centrifugation in some samples. While the majority of these fragments were >7,500bp, all fragments of length 500 bp or longer were significantly influenced by high-speed centrifugation.

The ultra-long fragments we identified had a dramatic loss of DNASE1L3-associated end motifs, which start with the nucleotide sequence CC. Several studies have also shown a loss of end motifs beginning with C in circulating DNA fragments over 500 bp based on ONT or Pacific Biosciences sequencing [24,27,29]. A more recent study suggested that these fragments may be generated by DNase DFFB rather than DNASE1L3, because knockout of the *Dffb* gene in mice led to a reduction in the number of fragments >500bp [28]. However, the degree of reduction was relatively small (~30%) and was not statistically significant based on the small number of animals analyzed. Here, we found an order of magnitude difference in the frequency of ultra-long reads before and after high-speed centrifugation, demonstrating that longer reads are at least sometimes a consequence of pre-analytical variables. This suggests that the abundance of such reads can be used for pre-analytical quality control and assurance and can be bioinformatically subtracted when they do occur.

We were able to clearly distinguish another class of long fragments, which was cancer-specific and included almost 20% of the cancer samples in our study, with the majority being Acute Myeloid Leukemias. These fragments could be distinguished by a shorter length distribution (peaking at 900-4300bp or 5-25 nucleosomes) and had end motif patterns consistent with typical DNASE1L3 fragmentation. The length distribution observed in these cancers is broadly consistent with that of Neutrophil Extracellular Traps (NETs) exposed to circulating nucleases [36], suggesting they might derive

from NETs fragmented in circulation. NETs have been associated with myeloid hyperproliferation [37] and linked to accumulation of total circulating DNA in cancer [38,39]. Alternatively, laddered fragments in this size range can be released by necrotic cells [40]. It's also possible that active DNASE1L3 during blood handling could fragment longer DNA from either necrosing or lysed cells. While our read coverage here is too low to robustly determine the cell of origin of these long fragments, this can be addressed in future studies using higher read coverage and methylation cell-of-origin analysis from ONT sequences.

Our finding that hyperfragmentation was poorly correlated to cancer DNA fraction is somewhat paradoxical in the face of earlier work which documented the preferential shortening of cancer DNA relative to non-cancer DNA using xenograft analysis [13,26,41], copy number analysis [14,15], and mutational analysis [13,42,43]. Despite this statistical association, there is a high degree of variability and many individual cancer cases show no preferential shortening of cancer DNA [15]. Furthermore, a systematic analysis of mutated fragments in cancer shows that fragment shortening is a weak predictor of cancer origin, with an AUROC of 0.51-0.61 across different solid cancer types [42]. Indeed, in our own analysis in hyperfragmented solid tumors, only 20% of cases showed a substantial preference for cancer DNA in short fragments, and the enrichment was modest.

In contrast, we did find a positive association between the degree of hyperfragmentation and total cfDNA and nucleosomes levels. This suggests that hyperfragmentation is linked to a systemic process which results in accumulation of both cancer and blood-derived cfDNA, which is linked to loss of DNASE1L3 activity [16,22,34,35]. The preferential shortening of cancer-derived DNA in some individual cancers is likely related to cancer-associated changes in nucleosome density and/or DNA methylation in long-range chromatin domains [21], both of which can influence DNASE1L3 activity [35]. The events that result in hyperfragmentation and accumulation of total cfDNA in cancer are unclear. In mice, loss of DNASE1L3 activity leads to fragment shortening, increased anti-DNA antibody, and progressive systemic lupus erythematosus (SLE) like disease [16,44]. Likewise, Loss of Function variants in the DNASE1L3 gene in humans are linked to shortening of fragments and SLE [34,45]. Taken together, these suggest that the hyperfragmentation phenotype and elevated cfDNA levels in cancer may both be the result of an imbalance between DNASE1L3 activity and cfDNA clearance that leads to an immune response.

Conflicts

All authors are employees or contractors for VolitionRx. BPB and TKK hold stock in VolitionRx. VolitionRx holds IP in the area of cancer liquid biopsy, including circulating DNA sequencing by Oxford Nanopore.

Author contributions

TK, JC, and BPB conceived the study and co-supervised the research. JC, TK, and SE planned and managed sample acquisition. SE managed and performed sample processing, library construction, and sequencing. BPB, JC, and JVT performed bioinformatic analyses. BPB drafted the manuscript, with TK, JC, and SE contributing. All authors have reviewed the manuscript.

Acknowledgements

The authors thank Christina Wheeler for assistance with lab workflow and sample management, and Daniel Halter and Jacques Avaux for support with cloud computing and Nextflow workflow infrastructure. We also thank Jake Micallef, Catherine Mallalieu, and Gaetan Michel for helpful comments.

Data availability

We have deposited three levels of data. Level 1: BAM files before filtering steps were modified by replacing all single-nucleotide variants with the base in the reference genome hg38, and all insertion variants with Ns (see Methods). These “anonymized” BAM files allow all fragmentomic and copy number analysis to be performed without exposing any personally identifying genetic variants. Level 1 data is deposited in Short Read Archive BioProject ID PRJNA1106648. Level 2: In order to provide combined fragmentation and methylation information, Biscuit [46] was used to create “epiBED” files (see Methods). epiBED files contain each read on a separate line, with each DNA methylation call on the read. epiBED files can be used for combined methylation/fragmentomic analysis, and are compatible with the CelFiE-ISH software [47] that was used for cell of origin deconvolution. Level 3: DNA methylation BED files. BED files created by modkit (see Methods) provide one line for each CpG covered, and can be used with bedtools and a number of other DNA methylation analysis tools. Level 2 and Level 3 data are deposited at Zenodo with DOI 10.5281/zenodo.11092490. These records are under embargo while this manuscript undergoes confidential peer review.

Code availability

Most analysis was performed using publicly available tools, with all software version numbers and command line settings listed in the Methods section. For several in-house scripts that were developed to analyze fragmentation patterns, source code is deposited at Zenodo with DOI 10.5281/zenodo.11092490. This record is under embargo while this manuscript undergoes confidential peer review.

Methods

Plasma sample procurement

Plasma samples were procured through the following Commercial Biobanks: Disovery Life Sciences (DLS), Innovative Research, and Bay Biosciences. All donor and biospecimen information is available in the Supplementrary Data File 1. Selection of 52 self-reported healthy individuals for Illumina sequencing were selected based on donor age criteria: an even distribution of ages from 40 to 90 years old. Selection of 207 cancer cases was based on Stage III/IV status and diversity of cancer types. Selection of healthy samples was based on volume of plasma available and the age of the donor. Samples were evenly distributed across ages with 10 samples in each of the following age groups: 40-49, 50-59, and 60-69 years old. 15 samples were procured in the age group 70-79, and 5 samples in the age group 80-89. Healthy samples were only considered on a lack of cancer diagnosis. Cancer samples were selected based on late stage diagnosis (III or IV), type of cancer, treatment status (untreated) and large sample volume.

All three commercial biobanks collect blood in EDTA tubes, then isolate the plasma by centrifuging at 1300-2000 xg for 10-15 minutes. The plasma is then frozen at -80°C. Exceptions to this protocol are listed in Supplementrary Data File 1.

Sample Storage and Preparation

Plasma samples were received and subsequently stored at -80°C. Aliquots were thawed at room temperature (RT) for up to 2 hours, depending on the aliquot volume. Plasma was then spun at 14000 x *g* for 2 minutes at room temperature and transferred to a new tube avoiding any pelleted fraction.

Nucleosome Quantification

Nucleosome levels in plasma were quantified using the Nu.Q® H3.1 assay developed by Volition for use on the IDS i10 instrument, using 250 uL of plasma according to the manufacturer's recommendations. Samples were completed either in duplicate or triplicate. Raw RLU (Relative Light Units)

were converted to concentrations of nucleosomes (ng/mL) using the provided kit standards, and replicate measurements were averaged.

DNA Extraction and Basic Characterization

All DNA extractions were completed with the QIAamp® Circulating Nucleic Acid Kit (Qiagen, catalogue # 55114) or with the QIAamp® MinElute ccfDNA Kit (Qiagen, catalogue # 55204). When possible, the extraction was completed with 1 mL plasma (if less, 1X PBS was used to bring volume to 1 mL). Additional extractions (if DNA levels were not sufficient for ONT) were carried out with up to 5 mL plasma, according to the manufacturer's protocol.

The Qiagen protocol was followed per the manufacturer's recommendations with the following adjustments: samples were eluted in 30 µL heated (60°C) elution buffer and the elution step was also completed at 60°C to increase DNA yield.

Samples were quantified with Qubit and total DNA levels were calculated based on plasma volume input. DNA profiles were confirmed using BioAnalyzer or Tapestation (Agilent).

Illumina Sequencing and mapping

When possible, DNA from the basic characterization step was used for Illumina Sequencing. If more was needed, additional DNA from 1-5 mL of plasma was extracted using the QIAamp® Circulating Nucleic Acid Kit (Qiagen, catalogue # 55114). Illumina sequencing libraries were constructed using SRSLY® PicoPlus DNA NGS Library Preparation Kit (ClaretBio – Cat: CBS-K250B). Libraries were sequenced by Discovery Life Sciences using the NovaSeq 6000 S4 200 cycle flow cell. Genome mapping to hg38 was performed using the Illumina BaseSpace DRAGEN. Mapping quality and filtering metrics are available as Supplementary Data File 1.

Oxford Nanopore (ONT) Sequencing

Samples were prepped for Nanopore sequencing using the 109 library prep kit and the 104 and 114 barcoding kits (ONT, catalogue #s: SQK-LSK109, EXP-NBD104, and EXP-NBD114, respectively). The Oxford Nanopore Technology (ONT) protocol was used, with the following adjustments: In the DNA repair and end-prep section, DNA input was decreased to 30 ng, and incubation time was increased from 5 min at 20C and 5 minutes at 65C to 30 minutes at each temperature. Throughout the protocol, the AMPure bead: reaction ratio was changed from 1:1 to 3:1, and ethanol wash volumes were increased from 200 uL to 400 uL. When sequenced, additional adaptations were made to the input and amount of library loaded per flow cell as needed

based on the success of the previous sequencing runs and the number of pores available on the flow cell. R9.4.1 flow cells were sequenced on the MinION Mk1c, with sequencing runs lasting for 16-72 hours, depending on the health of the flow cell. Samples were sequenced individually, but were still barcoded to avoid possible contamination. Plasma input and DNA input are included in Supplementrary Data File 1. Barcode demultiplexing and quality filtering were performed by MinKnow using FAST model.

ONT basecalling, mapping, and DNA modification calling

Passing fast5 files were re-processed using Dorado v. 0.2.4+3fc2b0f using the SUP basecalling model dna_r9.4.1_e8_sup@v3.3 and the Remora modification calling model dna_r9.4.1_e8_sup@v3.3_5mCG@v0, which resulted in unaligned BAM files including DNA modification tags. These were mapped to the UCSC hg38.analysisSet assembly (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/) using Minimap v. 2.24-r1122 with “minimap2 -y -t 2 -ax map-ont” and filtered with samtools v. 1.14 with “samtools view -e ([qs] >= 10) && (mapq >= 20)”. Duplicates were marked with samtools v.1.17 “samtools markdup”, and supplementary and secondary mappings were removed using “samtools view -F 0xD04”.

For global methylation analysis, methylation BED files were created using modkit v. 0.1.5 with the command “modkit pileup --cpg --combine-strands --ignore h --filter-threshold 0.9 --bedgraph". For methylation deconvolution, Minimap BAM files were used directly (details below). BED files are available in the Zenodo repository listed in Data Availability.

Copy number alteration analysis

Analysis of CNAs was performed using ichorCNA v. 0.5.0. Default parameters were used with all chromosomes 1-22, with a genomic bin size of 1E6. The following other parameters were used:

> “centromere=’GRCh38.GCA_000001405.2_centromere_acen.txt’, scStates=’’,repTimeWig=",ploidy=2,maxCN=7,includeHOMD=FALSE, estimateNormal=TRUE,estimatePloidy=TRUE,estimateScPrevalence =TRUE,altFracThreshold=0.05,genomeBuild="hg38",genomeStyle="U CSC",flankLength=100000,txnE=0.9999999,chrs="c('chr1',…,'chr22')", chrTrain="c('chr1',…,'chr22')",normal="c(0.5)"

CNA was considered “detected” if tumor fraction was estimated to be greater than 0.

Fragment length analysis

Fragment lengths were extracted from ONT Minimap BAMs using the script "fragmentationReports.py", which uses Pysam and defines the fragment length based on the primary alignment, calculating the difference between the start and end coordinates on the reference genome. Fragment length distributions for PCA analysis were created by defining bins as 10^x where x contains a range 50 to 50,000. Raw fragment lengths are log transformed and assigned to the closest bin by rounding to the nearest increment of 10^0.01. The proportion of fragments is defined as the number of fragments in a given bin divided by the total number of fragments in all bins. For the proportion of total DNA, each fragment in a bin is multiplied by the number of base pairs in the fragment and summed to get the bin base pair total. Each bin total is then divided by the sum of all base pair sums in all bins. The code to perform this analysis is available as the script "fraglens_to_histogram.py".

In order to perform CNA, end motif analysis, and methylation cell of origin analysis in specific fragment length bins, Minimap BAMs were filtered using the script "stratifyBamByFraglen.py", which uses the same Pysam code as above to calculate the fragment length of each read, and write that read to the output BAM file only if it is within the correct length range.

End motif analysis

ONT Minimap BAMs were processed using the script "fragmentationReports.py", which uses Pysam to perform the following analysis. We collect all reads that contain a perfect match to the final 5 base pairs of the ONT adapter sequence ("CACCT") as the final sequence in the soft-clipped portion of the read. We then collect the first 5 base pairs of the aligned portion of the read, which is by definition adjacent to the adapter sequence. This is the "end motif". If any of the adjacent 5 base pairs have a basecalling quality less than 20, the end is not counted. We count each of the two ends of each fragment as independent observations.

Principal Component Analysis (PCA) of fragment length distributions

DNA proportions were calculated for fragment length bins as described above. PCA was performed using the sklearn.decomposition Python package. For main figures 2-5, PCA was performed on column normalized versions of the fragment length distributions. These were defined by taking the proportion of DNA in each bin for of a given sample, and calculating a z-score based on the mean and standard deviation of the bin across all 21 healthy samples (not including the 3 "unspun" healthy samples). PCA in the main figures included 3 "unspun" healthy samples. PCA in Supplementary Figure 3D-G excluded these samples.

Methylation Cell of Origin (COO) analysis

We used a version of CelFiE-ISH package v0.0.2 (https://gitlab.com/methylgrammarlab/deconvolution_models),[47] modified to take Minimap BAM files as input. We used the command:

```
deconvolution --model uxm --min_length 2 --u_threshold 0.25 --modbam_qual 0.9 --percent_u U_atlas.hg38.tsv -m input.bam -i U250.hg38.tsv -b --cpg_coordinates hg38.analysisSet.CGmotifs.modkit.merged.0based.bed.gz --epiformat modbam --outfile output.txt
```

The "--modbam_qual 0.9" setting was used to filter out any modification base with a modification probability score less than 0.9. We used the 31 cell types from the WGBS atlas (Loyfer et al., 2023), as described in the CelFiE-ISH paper [47]. We summed individual cell types to create composite cell type groups such as "Gastrointesintal-Ep" ('Gastric-Ep'+'Colon-Ep'+'Small-Int-Ep'). All composite cell type groups are listed in Supplementary Figure 2. For Figure 2, each cell type was normalized by calculating the Z-score using the mean and standard deviation across all healthy and cancer samples.

Creation of anonymized BAM files.

We used script bamReplaceSeqByRef.py with "-ref hg38.analysisSet.fa", the same UCSC reference genome used for alignment. Files are available in the Zenodo repository listed in Data Availability, and code is available in the Zenodo repository listed in Code Availability.

Creation of Biscuit EpiBED files

Biscuit [46] v.1.4.1-dev was used with the command "biscuit epiread -M -b 0 -m 0 -a 0 -5 0 -3 0 -y 0.9 -L 1000000 hg38.analysisSet.fa", where the reference is the same UCSC reference genome used for alignment. Files are available in the Zenodo repository listed in Data Availability.

Statistical tests

Statistical tests, jitter plots, and scatterplots were generated with GraphPad Prism v. 10.2.1.

References

1. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. The Lancet. 2018;392:777–86.

2. Hasenleithner SO, Speicher MR. A clinician's handbook for using ctDNA throughout the patient journey. Mol Cancer. 2022;21:81.

3. Cohen SA, Liu MC, Aleshin A. Practical recommendations for using ctDNA in clinical decision making. Nature. 2023;619:259–68.

4. Mattox AK, Douville C, Wang Y, Popoli M, Ptak J, Silliman N, et al. The Origin of Highly Elevated Cell-Free DNA in Healthy Individuals and Patients with Pancreatic, Colorectal, Lung, or Ovarian Cancer. Cancer Discov. 2023;13:2166–79.

5. Van Der Pol Y, Mouliere F. Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. Cancer Cell. 2019;36:350–68.

6. Fitzgerald RC, Antoniou AC, Fruk L, Rosenfeld N. The future of early cancer detection. Nat Med. 2022;28:666–77.

7. Baca SC, Seo J-H, Davidsohn MP, Fortunato B, Semaan K, Sotudian S, et al. Liquid biopsy epigenomic profiling for cancer subtyping. Nat Med. 2023;29:2737–41.

8. Bauden M, Pamart D, Ansari D, Herzog M, Eccleston M, Micallef J, et al. Circulating nucleosomes as epigenetic biomarkers in pancreatic cancer. Clin Epigenetics. 2015;7:106.

9. Fedyuk V, Erez N, Furth N, Beresh O, Andreishcheva E, Shinde A, et al. Multiplexed, single-molecule, epigenetic analysis of plasma-isolated nucleosomes for cancer diagnostics. Nat Biotechnol. 2023;41:212–21.

10. Sadeh R, Sharkia I, Fialkoff G, Rahat A, Gutin J, Chappleboim A, et al. ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. Nat Biotechnol. 2021;39:586–98.

11. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. 2021;372:eaaw3616.

12. Thierry AR. Circulating DNA fragmentomics and cancer screening. Cell Genomics. 2023;3:100242.

13. Mouliere F, Robert B, Arnau Peyrotte E, Del Rio M, Ychou M, Molina F, et al. High Fragmentation Characterizes Tumour-Derived Circulating DNA. Lee T, editor. PLoS ONE. 2011;6:e23418.

14. Jiang P, Chan CWM, Chan KCA, Cheng SH, Wong J, Wong VW-S, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci [Internet]. 2015 [cited 2024 Mar 28];112. Available from: https://pnas.org/doi/full/10.1073/pnas.1500076112

15. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018;10:eaat4921.

16. Serpas L, Chan RWY, Jiang P, Ni M, Sun K, Rashidfarrokhi A, et al. *Dnase1l3* deletion causes aberrations in length and end-motif frequencies in plasma DNA. Proc Natl Acad Sci. 2019;116:641–9.

17. Budhraja KK, McDonald BR, Stephens MD, Contente-Cuomo T, Markus H, Farooq M, et al. Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer. Sci Transl Med. 2023;15:eabm6863.

18. Zhou Q, Kang G, Jiang P, Qiao R, Lam WKJ, Yu SCY, et al. Epigenetic analysis of cell-free DNA by fragmentomic profiling. Proc Natl Acad Sci. 2022;119:e2209852119.

19. Ulz P, Thallinger GG, Auer M, Graf R, Kashofer K, Jahn SW, et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. Nat Genet. 2016;48:1273–8.

20. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016;164:57–68.

21. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570:385–9.

22. Han DSC, Lo YMD. The Nexus of cfDNA and Nuclease Biology. Trends Genet. 2021;37:758–70.

23. Martignano F, Munagala U, Crucitta S, Mingrino A, Semeraro R, Del Re M, et al. Nanopore sequencing from liquid biopsy: analysis of copy number variations from cell-free DNA of lung cancer patients. Mol Cancer. 2021;20:32.

24. Choy LYL, Peng W, Jiang P, Cheng SH, Yu SCY, Shang H, et al. Single-Molecule Sequencing Enables Long Cell-Free DNA Detection and Direct Methylation Analysis for Cancer Patients. Clin Chem. 2022;68:1151–63.

25. Katsman E, Orlanski S, Martignano F, Fox-Fisher I, Shemer R, Dor Y, et al. Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. Genome Biol. 2022;23:158.

26. Van Der Pol Y, Tantyo NA, Evander N, Hentschel AE, Wever BM, Ramaker J, et al. Real-time analysis of the cancer genome and fragmentome from plasma and urine cell-free DNA using nanopore sequencing. EMBO Mol Med. 2023;15:e17282.

27. Yu SCY, Deng J, Qiao R, Cheng SH, Peng W, Lau SL, et al. Comparison of Single Molecule, Real-Time Sequencing and Nanopore Sequencing for Analysis of the Size, End-Motif, and Tissue-of-Origin of Long Cell-Free DNA in Plasma. Clin Chem. 2023;69:168–79.

28. Che H, Jiang P, Choy LYL, Cheng SH, Peng W, Chan RWY, et al. Genomic origin, fragmentomics, and transcriptional properties of long cell-free DNA molecules in human plasma. Genome Res. 2024;34:189–200.

29. Yu SCY, Jiang P, Peng W, Cheng SH, Cheung YTT, Tse OYO, et al. Single-molecule sequencing reveals a large population of long cell-free DNA molecules in maternal plasma. Proc Natl Acad Sci. 2021;118:e2114937118.

30. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017;8:1324.

31. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A DNA methylation atlas of normal human cell types. Nature. 2023;613:355–64.

32. Alcaide M, Cheung M, Hillman J, Rassekh SR, Deyell RJ, Batist G, et al. Evaluating the quantity, quality and size distribution of cell-free DNA by multiplex droplet digital PCR. Sci Rep. 2020;10:12564.

33. Markus H, Contente-Cuomo T, Farooq M, Liang WS, Borad MJ, Sivakumar S, et al. Evaluation of pre-analytical factors affecting plasma DNA analysis. Sci Rep. 2018;8:7375.

34. Chan RWY, Serpas L, Ni M, Volpi S, Hiraki LT, Tam L-S, et al. Plasma DNA Profile Associated with DNASE1L3 Gene Mutations: Clinical Observations, Relationships to Nuclease Substrate Preference, and In Vivo Correction. Am J Hum Genet. 2020;107:882–94.

35. Han DSC, Ni M, Chan RWY, Chan VWH, Lui KO, Chiu RWK, et al. The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB. Am J Hum Genet. 2020;106:202–14.

36. Pisareva E, Mihalovičová L, Pastor B, Kudriavtsev A, Mirandola A, Mazard T, et al. Neutrophil extracellular traps have auto-catabolic activity and produce mononucleosome-associated circulating DNA. Genome Med. 2022;14:135.

37. Wolach O, Sellar RS, Martinod K, Cherpokova D, McConkey M, Chappell RJ, et al. Increased neutrophil extracellular trap formation promotes thrombosis in myeloproliferative neoplasms. Sci Transl Med. 2018;10:eaan8292.

38. Pastor B, Abraham J-D, Pisareva E, Sanchez C, Kudriavstev A, Tanos R, et al. Association of neutrophil extracellular traps with the production of circulating DNA in patients with colorectal cancer. iScience. 2022;25:103826.

39. Thierry AR, Pisareva E. A New Paradigm of the Origins of Circulating DNA in Patients with Cancer. Cancer Discov. 2023;13:2122–4.

40. Ungerer V, Bronkhorst AJ, Van Den Ackerveken P, Herzog M, Holdenrieder S. Serial profiling of cell-free DNA and nucleosome histone modifications in cell cultures. Sci Rep. 2021;11:9460.

41. Moldovan N, Van Der Pol Y, Van Den Ende T, Boers D, Verkuijlen S, Creemers A, et al. Multi-modal cell-free DNA genomic and fragmentomic patterns enhance cancer survival and recurrence analysis. Cell Rep Med. 2024;5:101349.

42. Widman AJ, Shah M, Øgaard N, Khamnei CC, Frydendahl A, Deshpande A, et al. Machine learning guided signal enrichment for ultrasensitive plasma tumor burden monitoring [Internet]. 2022 [cited 2024 Mar 28]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.01.17.476508

43. Wan JCM, Heider K, Gale D, Murphy S, Fisher E, Mouliere F, et al. ctDNA monitoring using patient-specific sequencing and integration of variant reads. Sci Transl Med. 2020;12:eaaz8084.

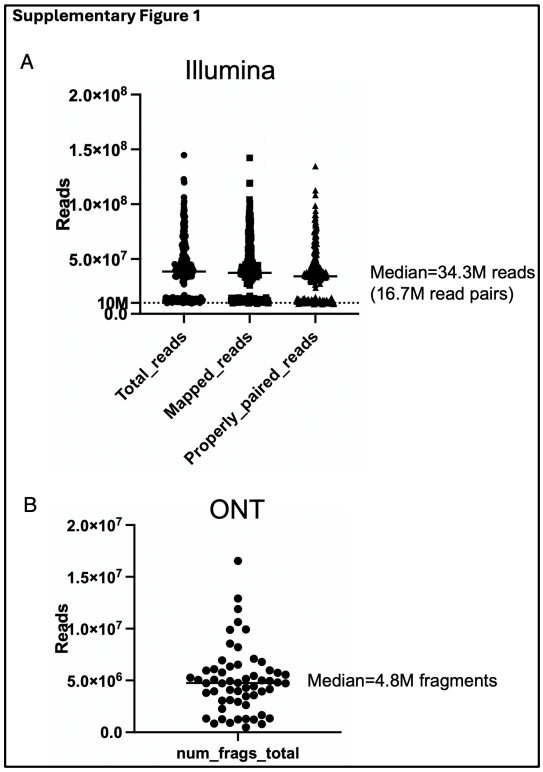
44. Sisirak V, Sally B, D'Agati V, Martinez-Ortiz W, Özçakar ZB, David J, et al. Digestion of Chromatin in Apoptotic Cell Microparticles Prevents Autoimmunity. Cell. 2016;166:88–101.

45. Chan RWY, Jiang P, Peng X, Tam L-S, Liao GJW, Li EKM, et al. Plasma DNA aberrations in systemic lupus erythematosus revealed by genomic and methylomic sequencing. Proc Natl Acad Sci [Internet]. 2014 [cited 2024 Mar 28];111. Available from: https://pnas.org/doi/full/10.1073/pnas.1421126111

46. Zhou W, Johnson BK, Morrison J, Beddows I, Eapen J, Katsman E, et al. BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies. Nucleic Acids Res. 2024;gkae097.

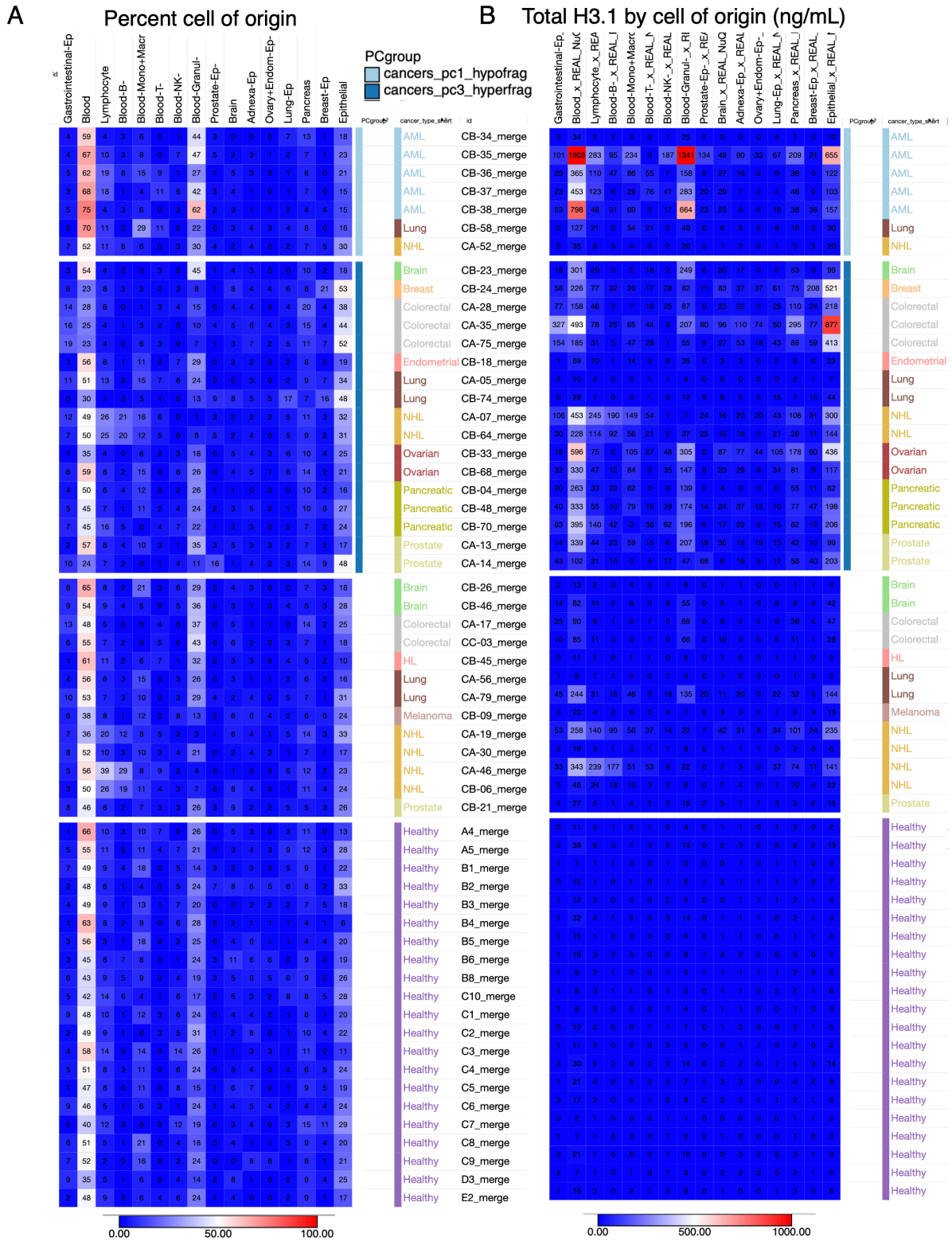
47. Unterman I, Avrahami D, Katsman E, Triche TJ, Glaser B, Berman BP. Multi-cell type deconvolution using a probabilistic model for single-molecule DNA methylation haplotypes [Internet]. 2023 [cited 2024 Mar 28]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.08.20.554012

Supplemental Figures



Supplementary Figure 1: Read counts in Illumina and ONT datasets. (A) Read counts for 259 cancer and healthy samples sequenced using Illumina NovaSeq. Each fragment consists of two paired reads. (B) Read counts for 58 cancer and healthy samples sequenced using ONT MinION Mk1c. Each fragment consists of a single read.

Supplementary Figure 2

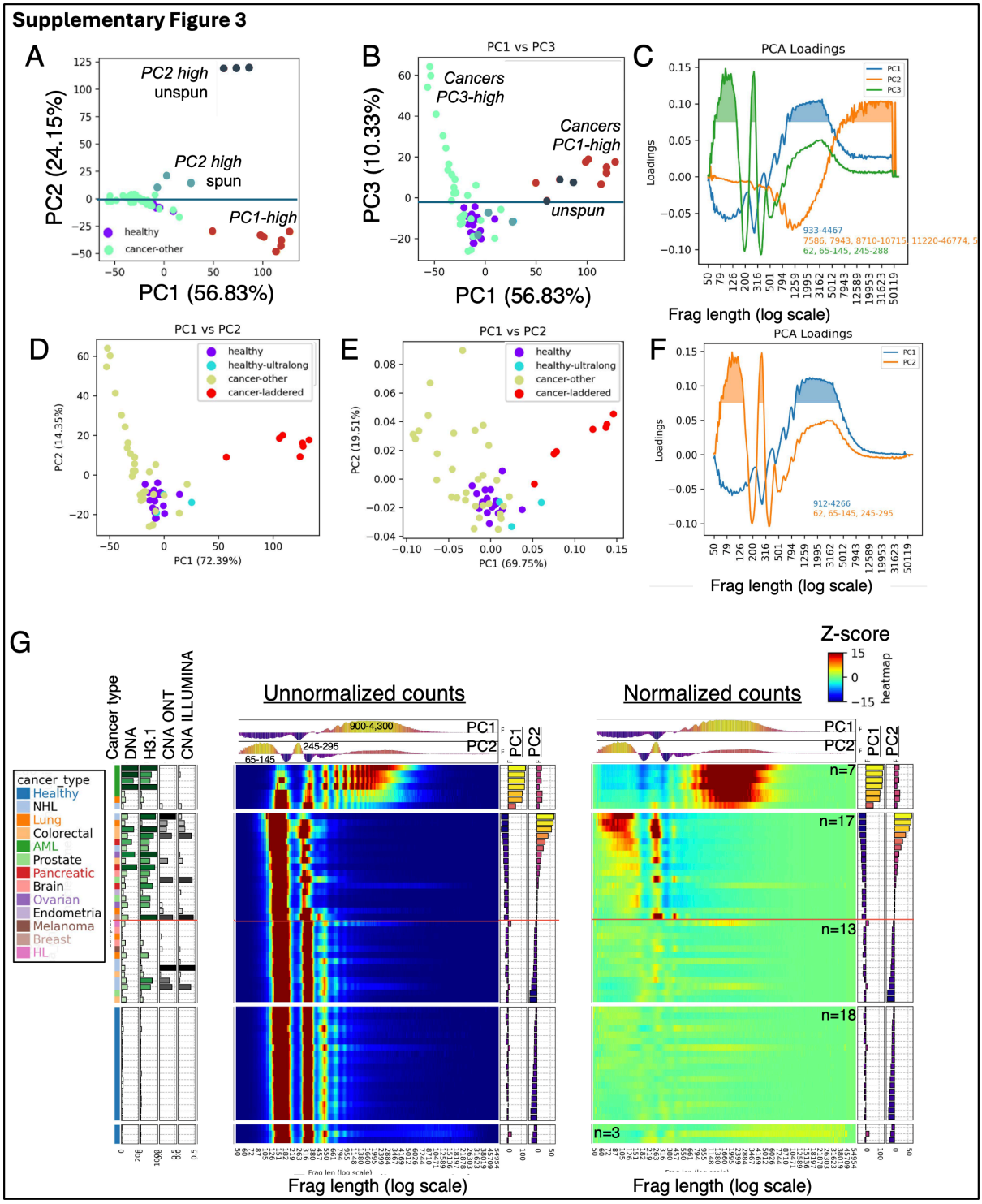


Supplementary Figure 2: Methylation-based cell of origin (COO). (A) For each sample, the non-negative least squares (NNLS) method of CelFiE-ISH was used to estimate the percentage of the mixture derived from each of 31 individual cell types. We then combine several cell type groups by summing percentages, with cell type groups as follows:

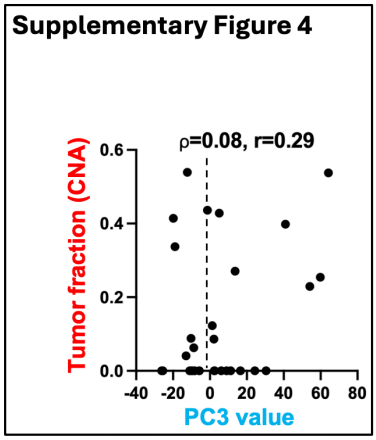
Gastrointestinal-Ep: 'Gastric-Ep','Colon-Ep','Small-Int-Ep'
Blood: 'Blood-T','Blood-NK','Blood-Mono+Macro','Blood-Granul','Blood-B','Eryth-prog'
Lymphocyte: 'Blood-T','Blood-NK','Blood-B'
Adnexa-Ep: 'Ovary+Endom-Ep','Fallopian-Ep'
Lung-Ep: 'Lung-Ep-Bron','Lung-Ep-Alveo'
Pancreas: 'Pancreas-Duct','Pancreas-Acinar','Pancreas-Delta','Pancreas-Beta','Pancreas-Alpha'
Breast-Ep: 'Breast-Basal-Ep','Breast-Luminal-Ep'
Epithelial: 'Bladder-Ep', 'Breast-Basal-Ep', 'Breast-Luminal-Ep', 'Colon-Ep', 'Fallopian-Ep', 'Gastric-Ep', 'Head-Neck-Ep', 'Kidney-Ep', 'Lung-Ep-Alveo', 'Lung-Ep-Bron', 'Ovary+Endom-Ep', 'Prostate-Ep', 'Small-Int-Ep', 'Thyroid-Ep'

Since these groups are not mutually exclusive, the percentages do not sum to 100%, although the 31 primary cell type percentages do. For instance, Blood-B is shown as an individual cell type, but also contained in both “Lymphocyte” and “Blood”. (B) Each cell type or cell type group percentage is multiplied by the total concentration of H3.1 nucleosomes (ng/mL) for the sample. The color scale is trimmed at a maximum value of 1,000 ng/mL.

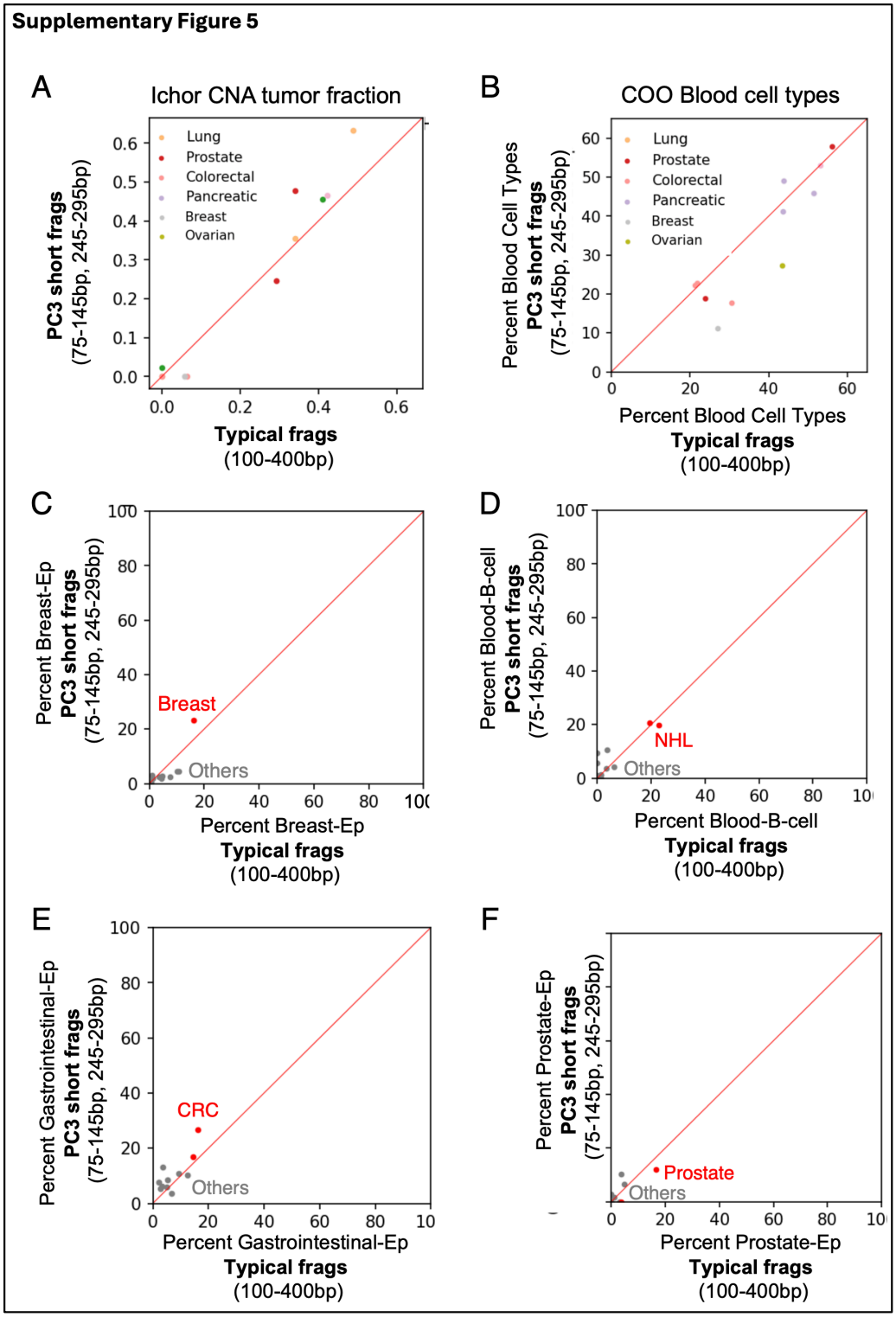
Supplementary Figure 3



Supplementary Figure 3 (prev. page): Principal Component Analysis of fragment length profiles. (A-C) correspond to the PCA presented in the main Figure 4. For this analysis, 58 healthy and cancer samples were included, plus the "unspun" samples of the 3 healthy volunteers with ultra-long fragments (E2, C10, D3). The input vector for each sample included one value for each fragment length bin, which represented the fraction of DNA in that bin, z-score normalized by the mean and standard deviation of the fractions of all healthy samples for the same bin (not including the "unspun" samples). (A-B) PC1, PC2, and PC3 values for each sample. In (B) a line is drawn between the PC3-high hyperfragmented cancer group and the normal-like cancer group in the main Figure 4. (C) shows the PCA loadings for each bin for each of the top 3 PCs. An arbitrary cutoff of 0.075 was used to define the bin ranges strongly associated with each PC, which are listed inside the figure. (D-G) An alternative PCA where the 3 unspun samples were omitted. The PC1 vs. PC2 plot in (D) is nearly identical to the PC1 vs. PC3 plot in the (B) panel. (E) An alternative PCA where the 3 unspun samples were omitted, and raw percent of DNA values were used rather than z-score normalized versions. (F) PCA loadings, defining the bin ranges strongly associated with each of the top 2 PCs, using the same cutoff as the (C) panel. (G) Sample ordering based on the alternative PCA.



Supplementary Figure 4: Quantitative version of the PC3 value vs. CNA tumor fraction. This is an alternative version of the binarized data shown in Figure 5E. Each point is one of 30 cancer samples (the 7 PC1-high hypofragmented cancer samples are omitted). The y axis is the tumor fraction estimated from ichorCNA based on the ONT sequence data.



Supplementary Figure 5: Cancer cell fraction in shortened fragments in hyperfragmented samples. Of the 17 PC3-high samples (PC3>-2), 7 had less than 1 million fragments in either of the PC3-associated length ranges 75-145 or 245-295. We used the remaining 10 to perform ichorCNA and methylation-based cell of origin (COO) analysis. (A) ichorCNA was run using only typical fragments (100-400bp) and fragments in the PC3-associated length ranges. (B-F) Methylation cell of origin was computed for combined Blood cell types (B) as well as cell types which represented the correct cell of origin: Breast-Ep for Breast cancer (C), B-cell for Non-Hodgkins Lymphoma (D), Gastrointestinal-Ep for Colorectal cancer (E), and Prostate-Ep for Prostate Cancer (F).