

METHOD

Open Access



Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing

Efrat Katsman^{1†}, Shari Orlanski^{1†}, Filippo Martignano^{2†}, Ilana Fox-Fisher¹, Ruth Shemer¹, Yuval Dor¹, Aviad Zick³, Amir Eden⁴, Iacopo Petrini⁵, Silvestro G. Conticello^{2,6*†} and Benjamin P. Berman^{1*†}

[†]Silvestro G. Conticello and Benjamin P. Berman jointly supervised the project.

*Correspondence: silvestro.conticello@cnr.it; ben.berman@mail.huji.ac.il

¹ Department of Developmental Biology and Cancer Research, Institute for Medical Research Israel-Canada, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

² Core Research Laboratory, ISPRO, Florence, Italy

³ Department of Oncology, Hadassah Medical Center, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

⁴ Department of Cell and Developmental Biology, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

⁵ Unit of Respiratory Medicine, Department of Critical Area and Surgical, Medical and Molecular Pathology, University Hospital of Pisa, Pisa, Italy

⁶ Institute of Clinical Physiology, National Research Council, Pisa, Italy

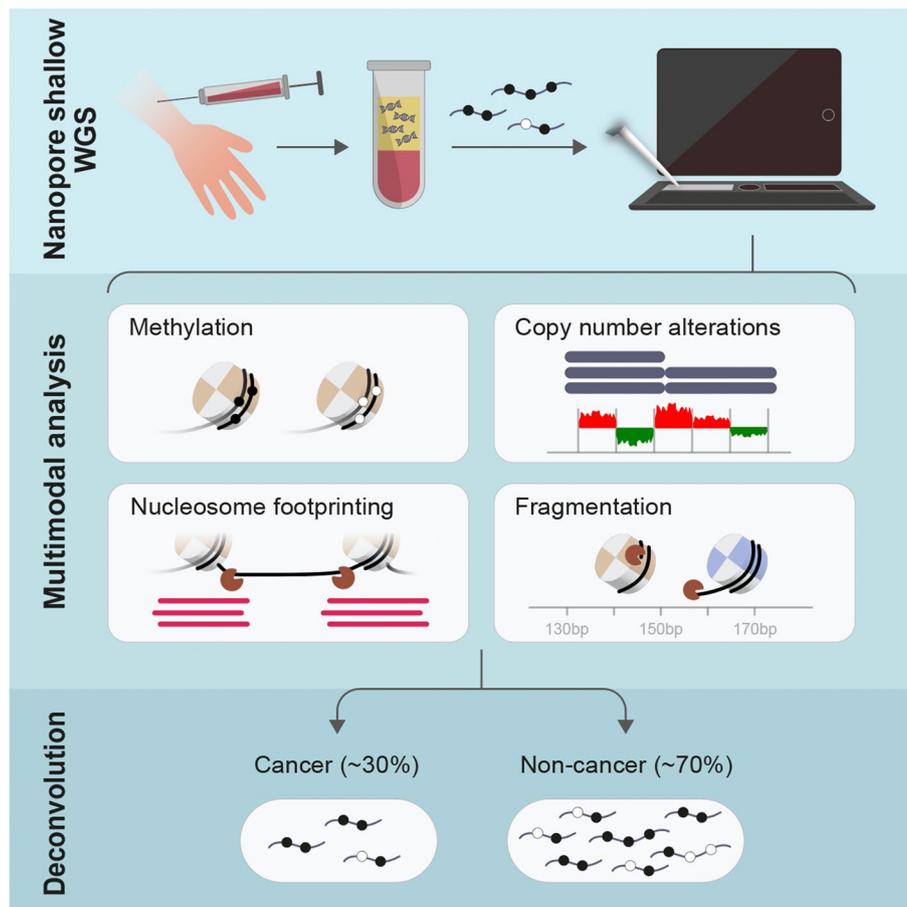
Abstract

The Oxford Nanopore (ONT) platform provides portable and rapid genome sequencing, and its ability to natively profile DNA methylation without complex sample processing is attractive for point-of-care real-time sequencing. We recently demonstrated ONT shallow whole-genome sequencing to detect copy number alterations (CNAs) from the circulating tumor DNA (ctDNA) of cancer patients. Here, we show that cell type and cancer-specific methylation changes can also be detected, as well as cancer-associated fragmentation signatures. This feasibility study suggests that ONT shallow WGS could be a powerful tool for liquid biopsy.



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Graphical Abstract



Background

Circulating cell-free DNA (cfDNA) can reveal informative features of its tissue of origin, including somatic genome alterations, DNA modifications, and cell type-specific fragmentation patterns [1]. DNA methylation is a promising cfDNA biomarker and is in widespread testing as a cancer screening tool [2]. DNA methylation can also be used to detect the turnover of damaged cells in time-sensitive conditions, such as myocardial infarction, sepsis, and COVID-19 [3–6]. These studies used bisulfite-based approaches to profile methylation, but alternative approaches include immunoprecipitation [7] and enzymatic conversion [8] techniques.

Accurate calling of DNA methylation from native DNA Oxford Nanopore (ONT) sequencing has matured and now produces single base-pair resolution results highly similar to bisulfite sequencing [9, 10]. The ONT platform is portable and has a low cost of setup and a rapid sequencing workflow that can enable real-time medicine [11, 12]. Native ONT sequencing requires no complex sample processing steps and no PCR amplification, making it attractive for clinical tests. Bisulfite approaches, in particular, involve significant degradation and loss of input material. For these reasons,

whole-genome sequencing (WGS) using the ONT platform is appealing relative to other whole-genome approaches.

Because of its low cost, targeted bisulfite PCR (including multiplexed NGS versions [13]) is also popular for clinical methylation sequencing, and these would not require the native modification calling capability of Nanopore. However, shallow whole-genome approaches that capture multiple genomic features could potentially be more informative, especially with regard to the course of the disease [8, 14, 15].

ONT has primarily been used for long-read sequencing, but recent work by our group and others has shown that it can be adapted for short fragments without additional processing steps [16–18]. As an added benefit, the ability to capture much longer cfDNA fragments than short-read sequencing may lead to new discoveries or biomarkers, as was demonstrated recently in the case of longer fragments during pregnancy [19].

Here, we perform a feasibility study of ONT sequencing for circulating tumor DNA (ctDNA) detection by comparing methylation and several fragmentation features to matched Illumina samples and comparable Illumina-based datasets.

Results

Estimating cell type fractions from cfNano

We first performed cell type deconvolution of healthy plasma cfDNA using DNA methylation data from either published WGBS datasets or our cfNano samples. For the external WGBS datasets, we used the methylation fractions (beta values) that were provided in the published data files. For our cfNano, we performed direct modification calling using the Megalodon software provided by ONT (<https://github.com/nanoporetech/megalodon>). To perform deconvolution, we used 1000–2000 marker CpGs per cell type based on a previously published atlas of purified cell types (“MethAtlas” [5, 13]) and estimated cell type fractions using non-negative least squares (NNLS) regression as described in [5]. In order to better understand the impact of the relatively low sequencing depth of our cfNano samples ($\sim 0.2\times$ genome coverage), we first performed deconvolution of all samples using downsampling experiments starting with full sequence depth down to $0.0001\times$ genome coverage (Fig. 1A and Additional file 1: Figs. S1–S3). Healthy plasma WGBS samples were taken from a recent study of $50\text{--}100\times$ genomic coverage [13] (Fig. 1A, left “Fox-Fisher” samples) and another WGBS study with $0.5\text{--}1\times$ coverage [20] (Fig. 1A, middle “Nguyen” samples). Finally, healthy cfNano samples were analyzed (Fig. 1A, right “this study”). From full depth down to $0.2\times$ (about 2.5M aligned fragments), all samples were dominated by the expected cell types: monocytes, lymphocytes, megakaryocytes, neutrophils/granulocytes, and sometimes hepatocytes [5]. Cell type proportions became significantly degraded at $0.05\times$ coverage and below (corresponding to less than 700,000 aligned fragments). The common cell types were consistently found across the 23 healthy individuals in the Fox-Fisher dataset, the 3 healthy individuals in the Nguyen dataset, and the 7 healthy individuals in the cfNano dataset, both at full depth (Fig. 1B) and when downsampled to $0.2\times$ depth (Fig. 1C). The same was found when cell type groups, such as lymphocytes, were broken down into individual cell types (Additional file 1: Fig. S4A–C). Notably, a slight epithelial fraction was identified in some of the Fox-Fisher samples at $0.2\times$, which did not appear at full $80\times$ depth, suggesting a small but measurable amount of noise at the $0.2\times$ coverage level.

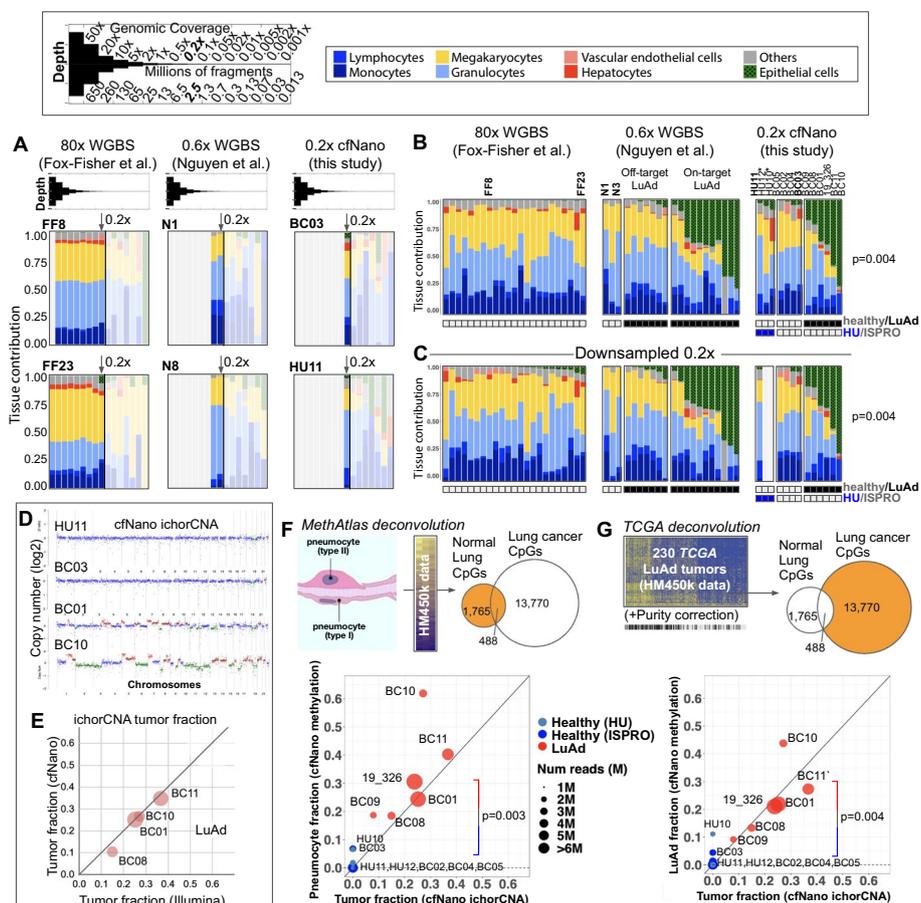


Fig. 1 Estimating cell type fractions from cfNano. **A** Non-negative least squares regression based on [5] was used to deconvolute cell types in healthy plasma cfDNA samples from three whole-genome DNA methylation studies. Two representative samples are shown for each study (FF8 and FF23 for the Fox-Fisher et al. study, N1 and N8 for the Nguyen et al. study, and BC03 and HU11 for our cfNano samples). Each sample is downsampled from full read depth down to an average genome coverage of 0.001 (corresponding to approximately 13,000 fragments). All samples are shown in Additional file 1: Figs. S1-S3. **B** Deconvolution of all samples at full depth, with samples ordered within each group by epithelial cell fraction. Healthy vs. lung adenocarcinoma (LuAd) is shown as an annotation bar, as is the source site (HU Israel vs. BC for ISPRO Italy) for the cfNano samples. Asterisks mark the two HU samples with coverage less than 0.2x sequence depth. Statistical significance (p -value = 0.004) is shown for percent epithelial in healthy cfNano samples vs. LuAd cfNano samples. **C** The same samples downsampled to 0.2x sequence depth. **D** ichorCNA CNA plots for 4 representative cfNano samples, two healthy and two LuAds. Plots for all samples are included in Additional file 5. **E** Tumor fraction (TF) estimates from four LuAd samples based on ichorCNA from cfNano and matched Illumina WGS. **F** Two-component DNA methylation deconvolution of lung fraction using CpGs from MethAtlas-purified lung epithelia samples, showing scatter plot of ichorCNA estimates vs. deconvolution estimates for all cfNano samples. Statistical significance is shown for the DNA methylation estimate of healthy cfNano vs. LuAd cfNano samples (p -value = 0.003). **G** Two-component DNA methylation deconvolution of lung cancer fraction using CpGs from TCGA LuAd tumor samples, showing scatter plot of ichorCNA estimates vs. deconvolution estimates for all cfNano samples (healthy vs. LuAd p -value = 0.004). Statistical significance for **B**, **C**, **F**, and **G** was determined by a one-tailed t -test

The healthy cfNano individuals were divided into two groups based on the source site, with one being collected and sequenced at ISPRO Italy (“BC” samples) and one in Israel (“HU” samples). Despite the HU samples being lower coverage (two were between 0.10–0.15x depth), they displayed relatively similar cell type proportions (Fig. 1B, C and Additional file 1: Fig. S3).

In addition to healthy individuals, the Nguyen WGBS dataset and our cfNano dataset also contained individuals being treated for lung adenocarcinoma, marked as “LuAd” in Fig. 1B, C. In the Nguyen WGBS study, samples were collected at the time of acquired resistance to EGFR inhibitors and were divided into those that acquired resistance mutations in EGFR itself (labeled “on-target”) vs. those that acquired amplifications in alternative oncogenes MET/ERBB2 (labeled “off-target”). The epithelial cell fraction was much higher in the on-target patients, while the off-target patients had very low or no epithelial fraction (Fig. 1B), consistent with the absence of CNAs in the off-target samples in the original study [20]. The 6 LuAd samples in our cfNano study had a similarly high epithelial fraction (Fig. 1B), which was significantly higher than in the healthy patients ($p = 0.004$). In all WGBS and cfNano samples, full-depth results were highly similar to $0.2\times$ downsampled results (Fig. 1C and Additional file 1: Figs. S1-S3). Interestingly, while the Nguyen et al. study interpreted the normal-like methylation levels of the “off-target” tumors as a difference in cancer methylation patterns, our deconvolution results strongly suggest that it is due to the relatively low amount of cancer DNA circulating in the blood.

The fraction of cancer cells in cfDNA (“tumor fraction”) can be estimated from somatic copy number alterations (CNAs) using the ichorCNA tool [21], for cancer cells that contain a sufficient degree of aneuploidy. We estimated tumor fraction for our cfNano samples and four matched Illumina WGS samples from LuAd patients (Fig. 1D, Additional file 2: Table S1, and Additional file 5). While the Illumina samples were sequenced at significantly higher depth (median $1.3\times$), the tumor fraction estimates were highly similar between cfNano and Illumina sequencing (Fig. 1E). Interestingly, the ichorCNA tumor fractions were more similar to the high-depth Illumina samples than the Illumina samples were to themselves when downsampled to the same depth as the cfNano samples (Additional file 1: Fig. S5A).

To compare ichorCNA tumor fraction estimates to methylation-based estimates, we designed a “two-component” deconvolution method based on NNLS regression that used 2253 CpGs with differential methylation between sorted lung epithelia and healthy plasma. This was based on the same array-based MethAtlas samples from [5] as the full deconvolution (Fig. 1F). Three hundred thirty to 1526 of these CpGs were covered by each cfNano sample (Additional file 2: Table S1), which were the CpGs used for NNLS deconvolution. These DNA methylation-based estimates of lung fraction and the ichorCNA estimates of cancer cell fraction were largely in agreement (Fig. 1F, bottom) with all of the six LuAd samples having significantly higher lung fraction compared to the seven healthy plasma samples ($p = 0.003$). Two LuAd cases were markedly higher in the methylation-based than the CNA-based estimate (BC09 and BC10). While we have no independent data to determine which was the more accurate estimate, we hypothesize that the discrepancy may be due to either whole-genome doubling (WGD) events that are not detected by ichorCNA (WGD occurs in 297/503 or 59% of LuAd tumors from the TCGA project [22]) or damage to normal lung cells surrounding the tumor which die and shed their DNA into circulation [23].

To verify the robustness of methylation-based deconvolution, we used a mutually exclusive set of 13,770 CpGs that could distinguish TCGA LuAd tumors from healthy plasma but were not found in the normal lung epithelia set (Fig. 1G). Before applying

the NNLS regression, since most TCGA LuAd samples contain a significant fraction of leukocytes, we corrected the methylation levels of the TCGA LuAd samples based on their non-cancer cell contamination (“purity correction”). After this correction, the tumor fraction estimates of our cfNano samples were highly similar to those based on normal-lung specific CpGs (Fig. 1G, bottom), despite the fact that the two CpG sets were completely non-overlapping. One HU healthy sample (HU005.10) had a higher lung fraction estimate than one of the cancer samples, possibly because this was the cfNano sample with the lowest sequencing coverage ($0.11\times$). However, the methylation-based tumor fraction was still significantly higher in the LuAd samples than in healthy controls ($p = 0.004$).

We performed all deconvolution analyses using a second, and older, base modification caller (DeepSignal [24]). While Megalodon called 10–20% more CpGs, the majority of CpGs called were in common between the two methods and had identical methylation states (Additional file 1: Fig. S6A). Both the full cell type deconvolution (Additional file 1: Fig. S6B) and the two-component deconvolution (Additional file 1: Fig. S6C) were highly similar between the two callers.

Genomic context of DNA methylation changes detected using cfNano

The deconvolution analysis above was based on unannotated differentially methylated regions. In order to investigate the genomic context of lung cancer-specific DNA methylation, we analyzed one hypomethylation feature associated with cell of origin (lineage-specific transcription factor binding sites) and one associated specifically with transformation (global hypomethylation). For the TFBS analysis, we identified 5974 predicted TFBS that were specific to lung epithelia based on a single-cell ATAC-seq atlas of open chromatin within lung and other primary human tissues [25]. In that study, adult lung tissues from multiple donors contained a strong cluster of lung pneumocyte-specific open chromatin regions (“Pal” cluster). This cluster was most strongly enriched for the binding motif for the transcription factor NKX2-1, which is a master regulatory transcription factor in this cell type [26]. NKX2-1 activity is also known to be highly restricted to this cell type [27], and NKX2-1 binding sites were also the most enriched within lung adenocarcinoma ATAC-seq sites in an independent study [28]. Because open chromatin regions are almost universally hypomethylated, we hypothesized that the 5974 predicted NKX2-1 TFBS in lung pneumocytes would be specifically hypomethylated in healthy lung tissues and in lung tumors. We confirmed this using WGBS data from TCGA [29] (Additional file 1: Fig. S7A).

We next plotted plasma cfDNA methylation levels at these same predicted NKX2-1 sites from the published Illumina WGBS studies and our cfNano study (Fig. 2A). In healthy samples from three WGBS studies and our own cfNano samples, NKX2-1 sites were fully methylated. In contrast, the LuAd samples from both the Nguyen et al. WGBS study (Fig. 2A, middle) and our cfNano study (Fig. 2A, right) had substantial demethylation. In both studies, demethylation could only be observed in the higher tumor fraction samples (“on-target” samples in the WGBS study, and samples with ichorCNA TF > 0.15 in the cfNano study). As a negative control, we selected predicted TFBS from a cell type not expected to be found either in healthy plasma or LuAd. We used the adrenal cortical cluster (“Adc” cluster) from [25], which was highly enriched for the KLF5 binding motif.

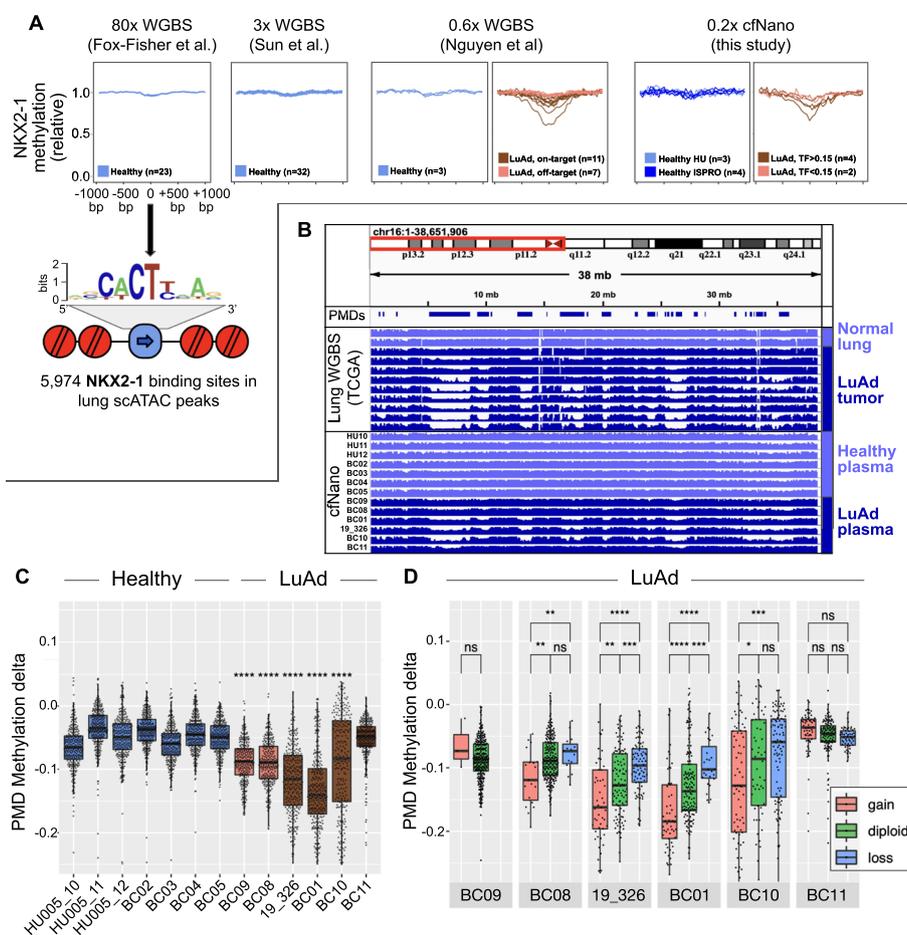


Fig. 2 Genomic context of DNA methylation changes detected using cfNano. **A** Plasma cfDNA methylation levels were averaged from -1 to $+1$ kb at 5974 pneumocyte-specific NKX2-1 transcription factor binding sites (TFBS) taken from [25]. All methylation values are fold change relative to the flanking region (region from 0.8 to 1 kb from the TFBS). From left to right, plots show 23 healthy plasma samples from [13], 32 healthy plasma samples from [30], 3 healthy and 18 LuAd WGBS samples from [20], and 7 healthy and 6 LuAd cfNano samples from this study. **B** Average DNA methylation across chr16p, comparing lung tissue WGBS (top) to plasma cfNano samples from this study (bottom). Reference partially methylated domains (PMDs) are taken from [29]. **C** Methylation delta is shown for all 10-Mbp bins overlapping a reference PMD (methylation delta defined as the average methylation of the bin minus the average methylation genome-wide). Each cancer sample was compared to the group of healthy samples using a one-tailed *t*-test, and statistical significance is shown using asterisks. **D** 10-Mbp PMD bins were stratified by copy number status for each cancer sample, and statistically significant differences were calculated by performing one-tailed Wilcoxon tests within each sample. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$

These sites were fully methylated in plasma samples from both healthy and LuAd individuals (Additional file 1: Fig. S7B). cfNano profiles were nearly identical using DeepSignal methylation calling (Additional file 1: Fig. S7C).

Global DNA hypomethylation is one of the hallmarks of the cancer epigenome. It has long been proposed as a general marker for circulating tumor DNA [31], and this was recently verified for lung cancer using shallow plasma cfDNA WGBS [20]. We have shown that this “global” hypomethylation is not completely global, and occurs preferentially within large domains called partially methylated domains (PMDs) and specifically at CpGs with a local sequence context termed “solo-WCGW” [29]. We replotted WGBS

methylation data from TCGA normal lung and lung tumor tissues, showing a typical chromosome arm where strong hypomethylation occurs within the PMD regions identified in [29] in the cancer samples (Fig. 2B, top). In our cfNano samples, strong hypomethylation was also found exclusively in the cancer samples in the same PMD regions (Fig. 2B, bottom). To quantify this genome-wide, we plotted the methylation change (relative to the sample-specific whole-genome average) of PMD solo-WCGW CpGs within each 10-Mbp genomic bin that overlapped a common PMD region from [29] (Fig. 2C). As expected, five of the six cancer samples were significantly hypomethylated relative to the healthy controls ($p < 0.0001$). Overall, there was significantly more hypomethylation across all cancer sample bins (mean = -0.10 , SD = 0.07) than across all healthy sample bins (mean = -0.05 , SD = 0.04), corresponding to a p -value $< 2.2E-16$ by the one-sided Wilcoxon test. In the final LuAd case (BC11), no PMD hypomethylation could be detected (Fig. 2C). This is not surprising given the high variability associated with global hypomethylation in cancer [29], a process that is not entirely understood but is affected both passively, through mitotic divisions, and actively, by dysregulation of several chromatin modifiers [32, 33].

Reasoning that copy number altered regions would have skewed proportions of tumor-derived DNA and thus different levels of PMD hypomethylation, we divided the PMD bins based on the copy number status of each sample. In four of the five cases with significant hypomethylation overall, the amplified bins had significantly more hypomethylation than diploid regions (Fig. 2D). In the one remaining case (BC09), there were not enough PMDs with CNAs for an accurate measurement. Conversely, deleted regions showed significantly less hypomethylation than diploid regions, although this trend only reached statistical significance in two cases. In the future, the combined analysis of CNAs and global hypomethylation may provide a stronger cancer-specific signal than each feature alone. PMD hypomethylation profiles were nearly identical using DeepSignal methylation calling (Additional file 1: Fig. S7D-F).

cfNano preserves nucleosome positioning signal

Cell-free DNA circulates primarily as mononucleosomal fragments, and mapping the positions of these mononucleosomes can be used to identify cell type-specific positioning (reviewed in [1]). CTCF binding sites provide a good test of whether these signals are detectable, since they eject a central nucleosome and position 10 phased nucleosomes on either side of their binding site [34] (Fig. 3A). Around a set of 9780 CTCF binding sites, cfNano mononucleosome locations recapitulated this expected pattern (Fig. 3B, top), which was identical to the pattern based on matched Illumina WGS of greater sequence depth (Fig. 3B, bottom). These were also identical when both cfNano and Illumina libraries were downsampled to an equal number of 2M fragments (Fig. 3C). In addition to nucleosome positioning, CTCF binding sites also demethylate all CpGs located approximately 200 bp on either side of the binding site [34], and this DNA pattern was also recapitulated in the methylation data from our cfNano samples (Additional file 1: Fig. S8).

We tried the same mononucleosome mapping approach for the 5974 lung-specific NKX2-1 TFBS from Fig. 2A. We could not detect any mononucleosome positioning signal (data not shown). Lung-specific nucleosome positions would only be present on a

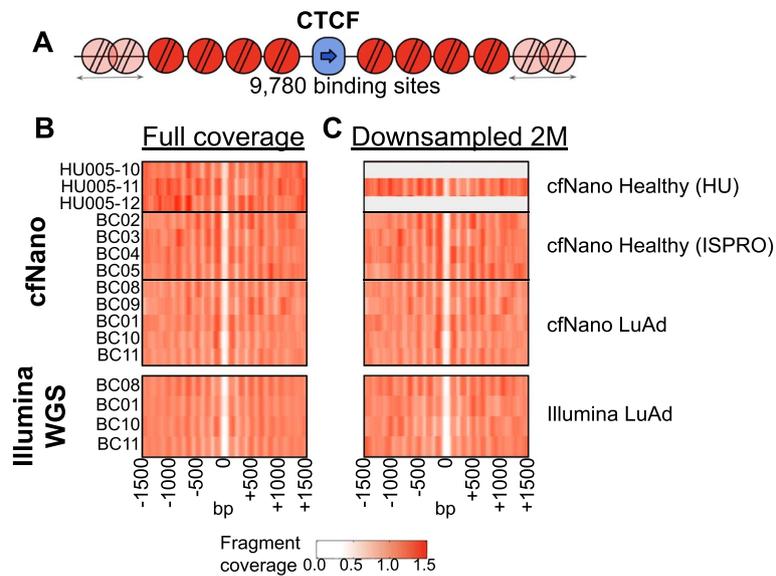


Fig. 3 cfNano preserves nucleosome positioning signal. **A** Alignments to 9780 CTCF motifs within non-promoter ChIP-seq peaks were taken from [34]. **B** Sequence coverage of mononucleosomes (130–155 bp) from cfNano samples is shown as fold change vs. average coverage across the genome (top). Mononucleosome coverage for matched Illumina samples (bottom). **C** The same analysis, using a randomly selected downsampling of 2 million reads from each sample. Two cfNano samples with less than 2M reads total are omitted

fraction of the fragments, so the signal from these fragments may be masked by those from non-lung cell types. But given that the inherent nucleosome positioning information is present (as shown by the CTCF example), more advanced normalization and quantification techniques may reveal these cell type-specific fragments more sensitively in the future.

Cancer-associated fragmentation length features of cfNano vs. Illumina WGS

Specific fragment lengths have been associated with cancer-derived cfDNA fragments (reviewed in [1]), and these have been used as accurate cancer classifiers [15]. Specifically, shorter mononucleosome fragments (< 150 bp) tend to be enriched for cancer-derived fragments [35]. Density plots of fragment length showed that our cfNano cancer samples were enriched in these short mononucleosome fragments relative to healthy controls (Fig. 4A). We used the definition from [35] and [15] to calculate the ratio of short mononucleosomes (100–150 bp) to all mononucleosomes (100–220 bp). The short mononucleosome ratio was significantly higher in the high tumor fraction cancer cases (mean = 0.24, SD = 0.03) than in the healthy cases (mean = 0.16, SD = 0.03), which corresponded to a *t*-test *p*-value of $p = 0.038$ (Fig. 4B). We compared these ratios calculated from our cfNano libraries with those calculated from the matched Illumina WGS libraries which were available for four of the six cancer samples, and the two library types were strongly correlated (Fig. 4C). We hypothesized that improvements to Nanopore base calling could improve alignment and adapter trimming, so we also compared base calling done with the real-time Guppy basecaller at the time of sequencing (“2019” version) to the new “high accuracy calling” base

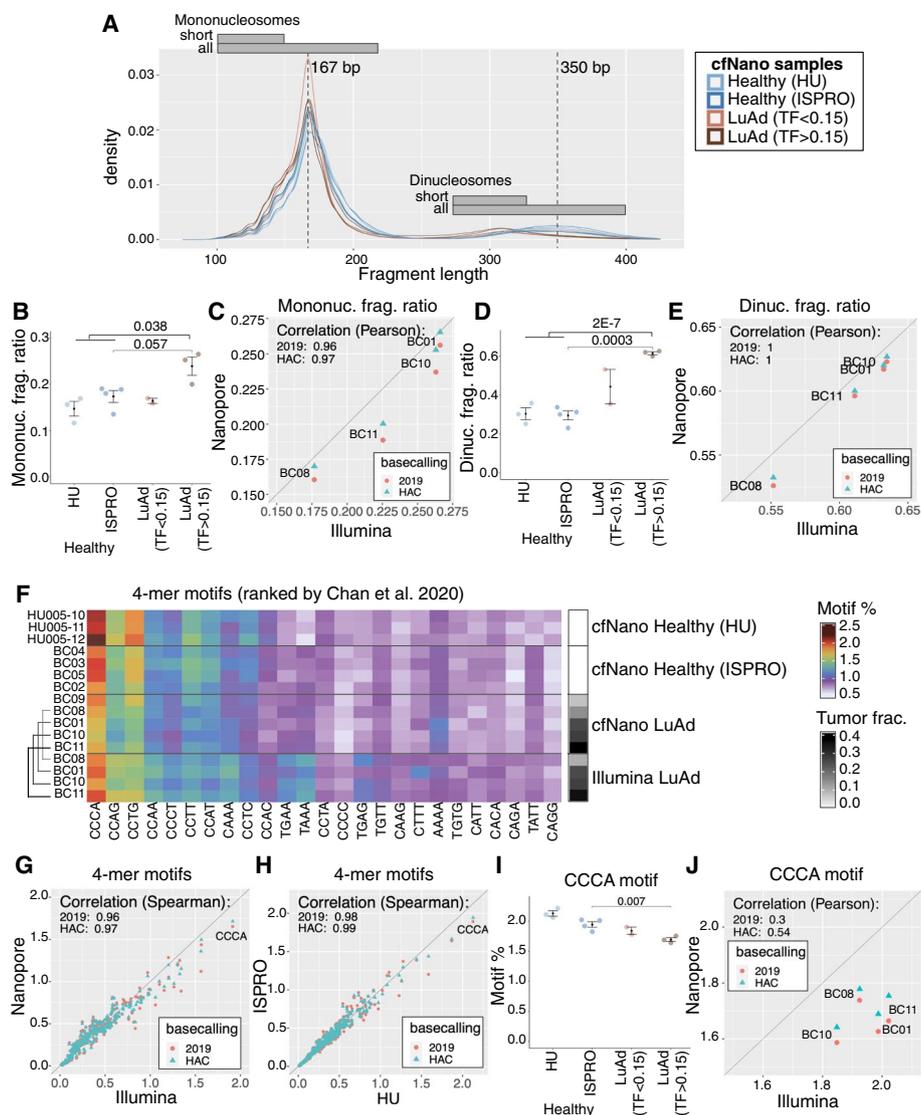


Fig. 4 Cancer-associated fragmentation features of cfNano vs. Illumina WGS. **A** Fragment length density plot for each cfNano sample, with cancer samples divided into low tumor fraction (TF < 0.15) and high tumor fraction (TF > 0.15) based on ichorCNA. Short mononucleosomes are defined as 100–150 bp [15, 35], and short dinucleosomes are defined at 275–325 bp. **B** The ratio (fraction) of short mononucleosome fragments (100–150 bp) to all mononucleosome fragments (100–220 bp). **C** Short mononucleosome ratios based on cfNano are compared to short mononucleosome ratios based on matched Illumina WGS libraries for four LuAd cases. cfNano samples were processed with either the 2019 Oxford Nanopore Real-time base calling model (2019) or the 2022 Oxford Nanopore High Accuracy model (HAC), as indicated by color. **D** The ratio (fraction) of short dinucleosome fragments (275–325 bp) to all dinucleosome fragments (275–400 bp). **E** Short dinucleosome ratios based on cfNano vs. Illumina WGS ratios for matched LuAd samples. **F** Frequency of 4-mer sequences occurring at fragment ends, for cfNano vs. matched Illumina samples. The 25 most frequent 4-mers are shown in ranked order based on frequencies in healthy plasma from [36]. **G** End-motif frequencies for all 256 possible 4-mers, comparing the average frequency in four cfNano samples vs. four matched Illumina WGS samples. **H** End-motif frequencies, comparing the average frequency in four healthy HU Italy cfNano samples vs. three healthy HU Israel cfNano samples. **I** Frequency of CCCA 4-mer in all cfNano samples. **J** CCCA 4-mer frequencies from cfNano samples vs. frequencies calculated from Illumina WGS for four matched LuAd samples. Statistical significance levels for **B**, **D**, and **I** were determined by a two-tailed *t*-test

calling (“HAC”) performed on all samples in 2022. The new ratios with the new base calling were slightly more similar to the matched Illumina libraries (Fig. 4C).

While they have not been studied as extensively as mononucleosomes, ref. [35] also showed that dinucleosomes were significantly shorter in cancer fragments than non-cancer fragments. This is clear from the density plots of our cfNano samples (Fig. 4A), so we used the size range suggested by ref. [35] to calculate the ratio between short dinucleosomes (275–325 bp) and all dinucleosomes (275–400 bp). The short dinucleosome ratio showed even more separation between cancer vs. healthy cfNano samples than the mononucleosome ratio. The high tumor fraction cancer cases had significantly higher dinucleosome ratios (mean = 0.62, SD = 0.01) than the healthy cases (mean = 0.30, SD = 0.04), corresponding to a *t*-test *p*-value of $p = 2E-7$ (Fig. 4D). We compared dinucleosome ratios calculated from our cfNano libraries with those calculated from the matched Illumina WGS libraries, and they were nearly perfectly correlated (Fig. 4E). When we looked across all samples, the short mononucleosome ratio was highly correlated with the short dinucleosome ratio (Additional file 1: Fig. S9D). Interestingly, this correlation held across the healthy samples as well as the cancer samples, indicating that the same underlying mechanism affects circulating cfDNA from both cancer and non-cancer cell types.

One of our cfNano samples used a different (non-barcoded) adapter design method from all other libraries, and this sample was a clear outlier in fragment length (Additional file 1: Fig. S9A-D). This reinforces the caution that should be taken when comparing fragmentomic features across different library designs. We also investigated the effect of sequencing depth on cancer-associated features, by comparing full-depth datasets with datasets created by randomly choosing 2M fragments for each library (Additional file 1: Fig. S9E-H). Sequence depth had almost no effect on either cfNano or Illumina samples down to 2M fragments.

Cancer-associated fragment end features of cfNano vs. Illumina WGS

The four bases immediately flanking cfDNA fragmentation sites have a biased sequence composition that differs between cancer-derived and non-cancer-derived fragments [36, 37]. To study this in our cfNano samples, we first plotted the 25 most abundant 4-mer end motifs that were identified in an Illumina-based study of healthy plasma cfDNA [36], using a heatmap to indicate motif frequencies in each of our cfNano and matched Illumina samples (Fig. 4F). There was broad agreement across all samples, although some differences between cfNano and Illumina libraries were clearly noticeable. When we plotted the average Nanopore vs. Illumina frequencies in matched samples for all 256 possible 4-mers, it appeared that the less abundant motifs had slightly higher frequencies in Nanopore, while the more abundant motifs had slightly lower frequencies in Nanopore (Fig. 4G). Nevertheless, the relative frequencies were highly concordant overall (PCC = 0.97). These were slightly more concordant when we used the 2022 “high accuracy” (HAC) base calling compared to the original 2019 base calling (PCC = 0.97 vs. PCC = 0.96). The degree of difference between the two batches of cfNano healthy samples (“ISPRO” sequenced in 2019 and “HU” sequenced in 2022) was less than the difference between cfNano and Illumina (PCC = 0.99, (Fig. 4H)).

Of particular interest is the CCCA end motif, which is typically the most abundant 4-mer in healthy plasma, and its reduction was shown to be a cancer marker in several cancer types, including lung cancer [36, 37]. CCCA indeed has the highest frequency across all our cfNano and Illumina WGS samples (Fig. 4F–H), and was significantly lower in our three high tumor fraction cancer samples than the healthy samples (Fig. 4I). However, there was a clear difference between the healthy samples generated in the “HU” and “ISPRO” batches, which we presume to be technical since these two batches behaved similarly with respect to fragment length and methylation features. We therefore only did a direct statistical comparison within the ISPRO batch, and indeed CCCA frequency in high TF tumors (mean = 1.7, SD = 0.06) was significantly lower than in ISPRO healthy samples (mean = 1.9, SD = 0.13), leading to a *t*-test *p*-value = 0.007 (Fig. 4I).

Additionally, the relative frequencies of four cfNano cancer samples were not concordant with their matched Illumina WGS libraries (Fig. 4J). We conclude from this that end motifs are particularly sensitive to changes in library strategy and sequencing platform, and caution must be taken when comparing across multiple batches. This is not surprising, given that fragment representation can be skewed by a number of variables during library construction and amplification, as well as sequencing errors and downstream bioinformatic steps such as adapter trimming (in our processing, we also exclude fragment ends that are soft-clipped). End motifs are highly susceptible to these biases, because even a single base pair difference results in a completely different motif. Recent benchmarking has highlighted how error frequencies can differ by sequence context between the Nanopore and Illumina platforms [38]. We found 16 additional motifs that were as significant as CCCA, although none survived FDR adjustment and so will have to be validated in larger studies (Additional file 4: Table S3). Like fragment lengths, end-motif frequencies were not sensitive to down sampling to 2M fragments (Additional file 1: Fig. S10A–D).

Discussion

While the sample size is small, our results suggest that cancer-specific features of DNA methylation, fragmentation, and CNA are broadly concordant between cfNano and Illumina-based WGS and WGBS methods. Downsampling analysis showed that the genomic coverage we targeted with cfNano (minimum of 2.5M aligned reads or 0.2× genome coverage) was sufficient to detect cancer-derived DNA in all samples based on DNA methylation. Cancer-associated fragmentomic features were not detected in all samples, but this is likely due to biological variability rather than sequence depth, based on similar studies using Illumina-based approaches [35], as well as our own results running the fragmentomic analyses on downsampled libraries. Notably, our results suggest that short dinucleosomes could be a more robust cancer marker than short mononucleosomes, although this will need to be validated in larger studies.

While most features agreed between cfNano and Illumina-based datasets, we identified fragment end motifs as one that was especially sensitive to sequencing platform differences. With the small sample size here, it is not possible to determine which of the two platforms provides the truer results. It is tempting to hypothesize that Nanopore provides a more accurate representation of actual fragment frequencies, since it does

not include any PCR amplification. On the other hand, Nanopore error rates are higher and may lead to less accurate adapter trimming or a different spectrum of sequencing errors within the end motifs (although we did try to control for this by using the reference genome sequence rather than the read sequence and by filtering out reads with soft-clipped ends). Despite these differences, we were able to detect the best-known cancer-specific end motif (CCCA) when we constrained the analysis to samples from our main cfNano batch (“ISPRO”). We also detected slight differences in end-motif frequencies between our own Illumina WGS and earlier published WGS data, and between our two cfNano batches which were sequenced two years apart. This indicates that end-motif frequencies are susceptible to batch effects in general and should be analyzed cautiously between batches. In the future, proper controls or normalizations should be developed to compare this feature across datasets sequenced at different times. In general, it is important to note that both the library construction kits and the underlying chemistry for Oxford Nanopore sequencing are still evolving and that future versions may continue to cause differences in the representation of fragmentation features such as length and end motifs.

With the small number of metastatic cancer samples used in our cfNano study, it was not possible to define the lower limit of methylation sensitivity. However, the fact that some epithelial/lung content was found when downsampling higher coverage WGBS samples, as well as in our cfNano healthy samples that had the lowest coverage, suggests that specificity needs to be improved at this very low coverage. We believe this could be improved significantly if whole-genome DNA methylation atlases (WGBS or similar) can be generated for individual cell types purified from human tissues. The purified cell type atlas we used here was based on the Illumina HM450k platform, which covers only about 10–15% of cell type-specific methylation markers. Alternatively, cancer markers could come directly from discovery studies that use WGBS to profile cancer plasma cfDNA. Such datasets do exist in the private domain but remain proprietary (e.g., [2]).

Even if deconvolution can be improved bioinformatically, we will almost certainly need higher read coverage for applications that require more sensitive detection. The cfNano samples analyzed here were all from metastatic disease, and all had relatively high tumor fractions ($\geq 10\%$), and this does not represent the situation for cancer screening or detection of minimal residual disease. Thus, the Nanopore platform will need to be able to consistently produce tens of millions of reads to enable those applications. Using cfNano directly to discover new cancer markers would not necessarily require greater coverage but would require large numbers of patient samples.

Several other current limitations of Nanopore should be considered. The cost per base of Nanopore sequencing is currently several-fold higher than Illumina, although the new generation of ONT PromethION sequencers is meant to address this and increase throughput. Single-nucleotide and indel error rates are higher for the ONT platform, which could pose an issue for whole-genome analysis of mutational signatures [39, 40], something we do not investigate here but is theoretically possible from cfNano. While Nanopore error rates have improved significantly over the past several years, this is a weak point that should be considered if mutations are a priority.

There are several areas where Nanopore methylation sequencing may provide unique strengths over other methods. Bisulfite-based sequencing leads to DNA fragmentation and degradation and can obscure fragmentation patterns [41]. Additionally, bisulfite sequencing can not distinguish between 5mC and other modifications such as 5hmC, 5fC, and 5CaC, whereas these are all in principle detectable by Nanopore. Longer cfDNA fragments have not been extensively studied due to the limitations of short-read technology, and these could be valuable both for biomarker discovery and the basic biology of cfDNA (as shown recently in [19]).

Conclusions

Despite the current limitations discussed above, the simplicity of native ONT sequencing and the number of features that can be extracted from a single run, combined with the low cost and portability of sequencer, make it an interesting proposition for clinical settings. Fast sample prep and sequencing times can allow a complete methylation analysis from sample preparation to computational classification in as little as 1–3 h, enabling real-time medical applications in cancer [11, 12]. Because DNA methylation can differentiate non-cancer cell types as well, Nanopore liquid biopsy could be used to monitor collateral damage to adjacent tissue in cancer [23] or urgent conditions in other areas of medicine such as myocardial infarction, sepsis, and COVID-19 [4–6, 42].

While the sample size of the current study is too small to determine the limits of sensitivity and specificity, we find that both cell type-specific and cancer-specific methylation features can be reliably detected in most of our samples, as well as copy number alterations and cancer-specific fragmentation features. The results provide confidence for pursuing this approach in larger studies.

Methods

ISPRO plasma cfDNA samples, library construction, and sequencing

ISPRO samples, library construction, and sequencing were described in our initial publication of these sequences [17]. The original sample names from that study are listed in Additional file 2: Table S1. Notably, one sample (S1/19_326) was produced using a different library kit (SQK-LSK109 vs. NBD-EXP104+SQK-LSK109 for all other samples). This is the singleplex library kit, which results in shorter adapter-ligated templates overall (due to the lack of barcodes) and thus responds differently to the equivalent clean-up bead concentration. Also, adapter trimming is performed differently in 19_326 due to the library kit differences. For these reasons, fragmentomic properties are not directly comparable between 19_326 and other samples. We thus analyzed 19_326 separately for all fragmentomic analyses (included in Additional file 1: Figs. S9-S10) but included it in all figures when analyzing the methylation and copy number alterations, where small differences in fragment length are not expected to make a difference. Standard MinKNOW runtime control was used without modification (S1 using distribution version 18, and all others using version 19).

HU plasma cfDNA samples, library construction, and sequencing

HU healthy samples are cfDNA extracted from 4-mL plasma, as originally described for these samples in [13]. The original sample names from that study are listed in Additional file 2: Table S1. Barcoded libraries were created using the NBD-EXP104 and SQK-LSK109 kits as described for the ISPRO samples. They were sequenced on a single flow cell, using standard MinKNOW runtime control (distribution v.21.11.7) without modification.

2019 real-time base calling and alignment of Nanopore fast5 files

Base calling was done using “high-accuracy real-time” mode during the run using MinKNOW distribution v.18 for run S1, and v.19.06.9 Guppy version 3.0.6+9999d81 for the others. For multiplex runs, demultiplexing was performed with guppy (version 5.0.16+b9fcd7b5b) using “--trim_barcodes --barcode_kits EXP-NBD104.” For the one singleplex sample (S1/19_326), adapters were trimmed using Porechop with parameters: “--discard_middle --extra_end_trim 0.” Minimap2 alignments were performed to GCF_000001405.39_GRCh38.p13 with minimap2 (version 2.13-r850), using the parameters “-ax map-ont --MD.” The resulting BAM files were used for fragment length and fragment end-motif analysis, below.

2022 high accuracy calling (HAC) base calling and alignment of Nanopore fast5 files

HU and ISPRO Fast5 files were base called and demultiplexed with Guppy (version 5.0.16+b9fcd7b5b) using “--flowcell FLO-MIN106 --kit SQK-LSK109 --trim_barcodes --barcode_kits EXP-NBD104,” model r9.4.1_450bps_hac. The resulting demultiplexed Fast5 files were used as input for Megalodon methylation analysis. Adapter trimming for the one singleplex sample, and alignment, was performed as for 2019 real-time Fast5 processing above. The resulting BAM files were used for ichorCNA, nucleosome (CTCF), fragment length, and fragment end-motif analysis, below.

Megalodon modification mapping to produce mod_mappings.bam files

Demultiplexed Fast5 files from the “2022 high accuracy calling (HAC) base calling and alignment of Nanopore fast5 files” section above were processed using Megalodon v. 2.4.2 with the following command-line parameters “--edge-buffer 0 --mod-min-prob 0 --guppy-params ‘-d /usr/local/hurcs/guppy/6.0.1/data --barcode_kits EXP-NBD104 --trim_barcodes’ --remora-modified-bases dna_r9.4.1_e8 hac 0.0.0 5mc CG 0 --guppy-config dna_r9.4.1_450bps_hac.cfg.” Internally, Megalodon used Guppy server version 6.0.1+652ffd1 and base calling model r9.4.1_450bps_hac. By default, Megalodon filters out multi-mapping (supplementary) reads and uses the minimap2 “map-ont” mode to filter low-quality mappings. Each individual Fast5 tile was run individually, and the resulting mod_mapping.bam files were merged into a single mod_mappings.bam file using samtools merge (v1.14). Samtools/HTSlib versions before v.1.14 were not able to handle the Mm/MI modification stages. Because Megalodon reports only the *reference* sequence in the BAM records, and does not report any base substitutions, these are anonymous BAM files which do not contain any genetic information, and thus contain no personally identifiable information and can be shared publicly. These are the primary

files used for all methylation analysis, described in more detail below, and are available from GEO GSE185307 and at Zenodo DOI: 10.5281/zenodo.6642503.

DeepSignal methylation calling and processing

We used DeepSignal Version 0.1.8 (4), with model “model.CpG.R9.4_1D.human_hx1.bn17.sn360.v0.1.7+/bn_17.sn_360.epoch_9.ckpt,” which was downloaded from the DeepSignal Google Drive (<https://drive.google.com/open?id=1zkK8Q1gyfviWWnXUBMcIwEDw3SocJg7P>). For ISPRO samples, Fast5 were annotated with fastq from 2019 real-time base calling; for HU samples, Fast5 were annotated with fastq from 2022 HAC base calling. We used the DeepSignal call_mods (modification_call) output tsv file, extracting the (strand-specific) methylation calls for each CpG from column 9 (called_label field) and calculated a methylation beta value by taking the number of methylated reads (value 1) divided by the total number of reads (value 0 or value 1). These were collapsed into a bedgraph file with a value between 0 and 1 for every CpG covered. These are available as file “grouped-beta-value_bedgraph.zip” in GEO accession GSE185307 and at Zenodo DOI: 10.5281/zenodo.6642503. All genomic coordinates are in GRCh38 and are zero-based.

Extracting methylation beta values from Megalodon mod_mapping.bam files

Modification mapping by Megalodon to produce mod_mapping.bam files is described above. To extract (stranded) methylation information from the mod_mapping.bam files, we used modbam2bed (<https://github.com/epi2me-labs/modbam2bed>) v.0.4.5, specifying a minimum probability threshold of 0.667, and filtering out positions with 0 confident reads using awk. The full command line was “modbam2bed --cpg -t 4 -a 0.333 -b 0.667 | awk ‘{>0}{print} > out.bed.” All coordinates are in GRCh38 and are 0-based. These files are named “*.5mC.cut0.667.hg38.bed.gz.” Column 11 corresponds to the percent of reads methylated. Modbam2bed does not provide a column for the actual number of reads that this percentage is based on, but it can be calculated from the other columns. $readCount=(col5*col10)/1000$. We also provide a simple bedgraph with just the methylation fraction (beta) values in files named “*cut0.667.hg38.sorted.bedgraph.gz.” These can be loaded into any genome browser. Both file types are available in GEO accession GSE185307 and at Zenodo DOI: 10.5281/zenodo.6642503.

Mapping of cfNano methylation data to HM450k probes

Using the zero-based stranded bed files from modbam2bed (“5mC.cut0.667.hg38.bed.gz” files), we mapped each CpG covering either the forward or reverse strand of each CpG on the Infinium 450k array. For each modbam2bed stranded column, we first got the readCount as $(col5*col10)/1000$. We then multiplied the methylation percentage by the read count to get the number of methylated reads. Then, we divided the sum of the methylated read counts for the two strands, by the sum of the total read counts for the two strands, to get the unstranded percent methylation (beta value).

Methylation calling from external WGBS datasets

For Fisher-Fox et al., methylation “beta.gz” files were obtained from GSE186888 and processed as recommended using `wgbs_tools` (https://github.com/nloyfer/wgbs_tools) `beta2bed` function to obtain fraction methylated and read count for each CpG. For Nguyen et al., bed files with methylation fractions and read counts were obtained from Figshare 10.6084/m9.figshare.16817941.v1.

For Sun et al., we obtained fastq files from EGAD00001001602 and aligned using `Biscuit` (<https://github.com/huishenlab/biscuit>) v.0.3.15.20200318 using the command line “`biscuit align -t 16 hg38.fa CTR153.fq.gz -b 1`” piped into `samblaster` [43] v.0.1.24 to mark and remove duplicates with the command line “`samblaster -i stdin -o stdout -M --excludeDups --addMateTags --ignoreUnmated -d CTR153.hg38_discordant.sam -s CTR153.hg38_split.sam --maxSplitCount 2 --maxUnmappedBases 50 --minIndelSize 50 --minNonOverlap 20 -u CTR153.hg38_fastq --minClipSize 20`”

Methylation coverage downsampling

To downsample methylation coverage from bed files with read count and fraction methylated columns, we used a custom Perl script in the <https://github.com/methylgrammarlab/cfdna-ont> repository called `downsampleMethylBed.pl`. This script treats each read at each CpG as an independent observation, and then randomly samples from these until it has enough observations to reach the average genomic coverage requested. To obtain the coverage levels shown in Fig. 1, it was run with the command line “`downsampleMethylBed.pl --coverageLevels 1E-3,2E-3,5E-3,1E-2,2E-2,5E-2,1E-1,2E-1,5E-1,1E0,2E0,5E0,1E1,2E1,5E1,8E1 --fracTotalFieldsFrom0 -3,4 --ncpgsGenome 28217005`”

Full cell type methylation deconvolution

For the full cell type deconvolution in Fig. 1A–C, we used the non-negative least squares regression (NNLS) method from [5]. Specifically, we used the code from https://github.com/nloyfer/meth_atlas/blob/master/deconvolve.py. We used an input set of methylation markers that included the top 1000 hypermethylated and 1000 hypomethylated CpGs for each of the 25 cell types provided. To generate the reference atlas, we used the script https://github.com/methylgrammarlab/cfdna-ont/blob/main/deconvolution_code/cell_type_probes/creating_reference_atlas/feature_selection_function.m, with the input of “1000” as number of CpGs. Full results were plotted using a modified version of the original `deconvolve.py` which we have deposited in https://github.com/methylgrammarlab/cfdna-ont/blob/main/deconvolution_code/deconvolution_moss/plot_deconv.py. These are shown in Additional file 1: Fig. S4A–B.

For Fig. 1A–C, we collapsed cell types into 8 groups using the file https://github.com/methylgrammarlab/cfdna-ont/blob/main/deconvolution_code/deconvolution_moss/group_file_for_plot_green_epithelial.csv (shown visually in Additional file 1: Fig. S4C). We plotted results using code in https://github.com/methylgrammarlab/cfdna-ont/blob/main/deconvolution_code/deconvolution_moss/deconvolution_plot.R. For DeepSignal methylation data, the procedure was the same except we used the top 2000 hypermethylated and top 2000 hypomethylated CpGs, to account for the significantly smaller number of CpGs called in the DeepSignal data (shown in Additional file 1: Fig. S6A).

ichorCNA analysis

BAM files from the 2022 HAC base calling and alignment step above were used as input. Samtools (Version 1.9) was used to filter BAM alignments, unmapped reads, secondary and supplementary reads, reads with mapping quality less than 20 as in [44], and reads longer than 700 bp. For Illumina alignments, we trimmed all “N” nucleotides from the 3’ ends of fastq data, alignments were performed to GCF_000001405.39_GRCh38.p13 with BWA mem [44], and duplicates were marked using picard MarkDuplicates and removed with samtools; read pairs without the properly-paired flag were removed. Pipelines used for preprocessing and filtering of both Nanopore and Illumina data are available at https://github.com/Puputnik/Fragmentomics_GenomBiol. Somatic copy number analysis was performed using the ichorCNA package v.0.3.2 [21].

We used ichorCNA to determine copy number alterations and tumor fraction for each cancer sample. If the percentage of genome covered by CN alterations was less than 15%, then the tumor fraction was determined to be unstable and set to 0. ichorCNA tumor fraction estimates are available in Additional file 2: Table S1, and genomic CNA plots for all samples are available as Additional file 5. The ichorCNA parameters are, “--ploidy c(2) --normal c(0.5) --maxCN 7 --includeHOMD False --estimateNormal True --estimatePloidy True --estimateScPrevalence True --altFracThreshold 0.001 --rmCentromereFlankLength 1000000.”

Two-component cell type methylation deconvolution using healthy lung epithelia

To determine the lung fraction from different datasets, we devised a “two-component” version of the NNLS regression model described above. To compose the atlas of differentially methylated probes in 25 human tissues and cell types, we used the data collected and tissue-specific feature selection method from the MethAtlas package (https://github.com/nloyfer/meth_atlas) [5]. The script `feature_selection_function.m` (https://github.com/methylgrammlab/cfdna-ont/blob/main/deconvolution_code/cell_type_probes/creating_reference_atlas/feature_selection_function.m) was used to select Lung_cell epithelial-specific CpGs. For the Megalodon version, the cutoff was set to select the top 1000 hypermethylated and the top 1000 hypomethylated probes, using the three Lung_cell epithelia samples vs. the four healthy plasma cfDNA samples from [5]. For DeepSignal methylation data, the procedure was the same except we used the top 2000 hypermethylated and top 2000 hypomethylated CpGs, to account for the significantly smaller number of CpGs called in the DeepSignal data (shown in Additional file 1: Fig. S6A). We removed any probe that did not have valid (non-NA) values for 2 or more of the Lung_cell samples and 2 or more of the healthy plasma samples.

For each probe, the 450k beta values were averaged across three lung samples to produce a single Lung-specific beta value \bar{X}_1 . The same was done for the four plasma cfDNA samples from to yield a healthy cfDNA beta value \bar{X}_2 . We used the Lawson-Hanson algorithm for non-negative least squares (NNLS) (<https://cran.r-project.org/web/packages/npls/>) to perform non-negative least squares regression as in [5]. Specifically, we identified non-negative coefficients β_1 and β_2 , representing the fraction of lung cells and normal blood cells in the Nanopore cfDNA mixture, respectively, subject to the

constraints $\arg\min_{\beta} \|X\beta - Y\|_2$ and $\beta \geq 0$. Then, a single lung fraction β was determined by having β_1 and β_2 sum to 1, with the equation $= \frac{\beta_1}{(\beta_1 + \beta_2)}$.

Two-component cell type methylation deconvolution using TCGA lung tumors

We downloaded the Infinium 450k beta value files for TCGA lung adenocarcinoma (LUAD) tumors using the ELMER packaged in Bioconductor [45]. We removed any probe that did not have valid (non-NA) values for 2 or more of the LUAD samples and 2 or more of the healthy plasma samples. In order to make this analysis completely independent from the healthy lung epithelia deconvolution analysis, we excluded 488 CpGs that were included in the Megalodon "Two-component cell type methylation deconvolution using healthy lung epithelia" analysis above, and an additional 396 that were included in the DeepSignal analysis. We then performed a *t*-test to compare the methylation beta values of these LUAD-specific probes to the four plasma cfDNA samples from the MethAtlas paper [5], requiring a Benjamini-Hochberg corrected FDR of < 0.001 and an absolute beta value difference of 0.3 or greater.

Correcting TCGA methylation model for cancer cell purity

NNLS was performed for the TCGA lung tumor deconvolution as described above in "Two-component cell type methylation deconvolution using healthy lung epithelia", with the following adaptation. The deconvolution assumes that each of the reference cell types is a representation of the purified cell type, but this is not the case for bulk TCGA tumors which have a median of leukocyte fraction of 30% [22]. For each probe in each TCGA cancer sample, we corrected for this by solving for the equation $M_m = M_c\beta + M_l(1 - \beta)$, where M_m is the methylation of the mixture, M_c is (unknown) methylation of the cancer cells, M_l is the (known) methylation of the leukocytes, and β is the (known) percentage of cancer cells in the mixture. M_l was taken as the average of white blood cell samples from the MethAtlas [5], and β was taken as the "tumor purity" estimate based on somatic copy number alterations from the TCGA PanCan Atlas project using the ABSOLUTE program [22], downloaded from the PanCan Atlas website (TCGA_mastercalls.abs_tables_JSedit.fixed.txt, URL <https://gdc.cancer.gov/about-data/publications/pancanatlas>). We used the pure cancer cell estimates M_c and performed NNLS regression as described above in "Two-component cell type methylation deconvolution using healthy lung epithelia".

DNA methylation in 10-Mbp PMD bins

To generate Fig. 2C, D, GRCh19 segmentation results from our previous CNV analysis [17] were divided into non-overlapping 10-Mb bins. Copy number status of each bin was determined by \log_2 ratio segment mean > 0.10 and < -0.10 for gain and loss, respectively. For the healthy samples, 10-Mb bins were generated from the whole genome. GRCh38 methylation files were converted to GRCh37 using *liftover* R package. We selected only the bins overlapping one or more common Partially Methylated Domains (PMDs) from [29]. Within these PMD bins, we took the average of all "solo-WCGW" CpGs overlapping a PMDs, with "solo-WCGW" annotation also from

[29]. We calculated the bin average from these CpGs as $\text{sum}(\text{frac_methylation_each_CpG})/\text{CpG_count}$. We then subtracted this bin average from the average of all CpGs in the genome, to get the methylation delta shown in Fig. 2C, D. Common PMD and solo-WCGW annotations were taken from file https://zwdzwd.s3.amazonaws.com/pmd/solo_WCGW_inCommonPMDs_hg38.bed.gz. Statistical significance between each cancer sample's bins and all pooled healthy sample bins in Fig. 2C was calculated by the one-sided Wilcoxon test (because we decided a priori to look only for hypomethylation in the cancer samples). For copy number analysis in Fig. 2D, each pair of copy number groups was compared using a one-sided Wilcoxon test, to test the hypotheses that diploid should have higher methylation than amplified regions, and deleted regions should have the highest methylation. The files and pipeline used for this analysis (including segmentation results from Martignano et al. [1]) are available at https://github.com/Puputnik/CNV_Methylation_Genome_Biol_2022.

Transcription factor binding site (TFBS) analysis

First, we used HOMER to identify predicted NKX2-1 binding sites (using the HOMER built-in matrix “nkx2.1.motif”) across the GRCh38 genome, and removed any site within the ENCODE blacklist. For normal lung cell analysis, we intersected this list with 6754 ATAC-seq peaks identified in the pneumocyte (PAL) cluster 13 CREs from [46] (downloaded from supplemental table 6 of that paper “Table_S6_Union_set_of_cCREs.xlsx”). We then selected 5974 peaks that overlapped a predicted NKX2-1 TFBS and centered each on the predicted NKX2-1 TFBS. If multiple TFBS were present in the peak, we took the motif with the highest HOMER log-odds match score. This TFBS set is available as file “nkx2.1.incluster13_distalPeaks_PAL.bed.highestScoreMotifs.hg38.bed” in at Zenodo DOI: 10.5281/zenodo.6642503. To calculate relative methylation levels, raw methylation levels in each bin were divided by the mean methylation within all bins from -1000 to -800 and $+800$ to $+1000$ across all NKX2-1 sites. For TCGA lung and non-lung samples in Additional file 1: Fig. S7A, we downloaded TCGA WGBS bedgraph files from <https://zwdzwd.github.io/pmd> [29]. We used all WGBS cancer types that were represented by normal tissues in the scATAC-seq atlas, as this was the atlas used to define pneumocyte-specific (PAL) peaks. These TCGA types included LUAD and LUSC (lung tissue from atlas), CRC (transverse colon tissue from atlas), BRCA (breast tissue from atlas), STAD (stomach tissue from atlas), and UCEC (uterus tissue from atlas).

KLF5 transcription factor binding site (TFBS) analysis (Additional file 1: Fig. S7B)

As with *NKX.2* above, we used HOMER to identify predicted KLF5 binding sites (using the HOMER built-in matrix “klf5.motif”) across the GRCh38 genome and removed any site within the ENCODE blacklist. As a control, we intersected this list with 9274 ATAC-seq peaks identified in the cluster 43 CREs from [46] (downloaded from supplemental table 6 of that paper “Table_S6_Union_set_of_cCREs.xlsx”). We then selected 1762 peaks that overlapped a predicted KLF5 TFBS and centered each on the predicted KLF5 TFBS. If multiple TFBS were present in the peak, we took the motif with the highest

HOMER log-odds match score. This TFBS set is available as file “klf5.incluster43Distal.txt.highestScoreMotifs.bed” in GEO accession GSE185307.

CTCF nucleosome positioning analysis

We used 9780 evolutionarily conserved CTCF motifs occurring in distal CHIP-seq peaks, which were taken from [34]. Nanopore or Illumina fragments within the size range of 130–155 bp were used for fragment coverage analysis, with reads being extracted from BAMs as described above. These shorter mononucleosomal fragments showed similar nucleosomal patterns but gave higher spatial resolution than 156–180-bp fragments. DeepTools (version 3.5.0) bamCoverage was used with the parameters “--ignoreDuplicates --binSize -bl ENCODE_blacklist -of bedgraph --effectiveGenomeSize 2913022398 --normalizeUsing RPGC.” For Illumina WGS, we used the additional parameter “--extendedReads 145.” The bedgraph was converted to a bigwig file using bigWigToBedGraph downloaded from UCSC Genome Browser. This bigwig file was passed to DeepTools computeMatrix with the command line parameters “reference-point --referencePoint center -out table.out,” and the table.out file was imported into R to create fragment coverage heatmap.

Fragment length analysis

BAM files from either the 2019 real-time basecalling and alignment, or 2022 HAC base calling and alignment, above, were used as input. Samtools (version 1.9) was used to filter BAM alignments, unmapped reads, secondary and supplementary reads, reads with mapping quality less than 20 as in [44], and reads longer than 700 bp. For Illumina alignments, we trimmed all “N” nucleotides from the 3′ ends of fastq data, alignments were performed to GCF_000001405.39_GRCh38.p13 with BWA mem [44], duplicates were marked using Picard MarkDuplicates and removed with samtools. Pipelines used for preprocessing and filtering of both Nanopore and Illumina data, and analyzed data are available at https://github.com/Puputnik/Fragmentomics_GenomBiol. In addition, only reads with barcodes at both ends (obtained using the --require_barcodes_both_ends flag while demultiplexing) were used for fragment length analysis of the multiplexed samples (all except 19_326). Read identifiers of double-barcoded reads are available in the “doubleBarcodeIds” file in Zenodo DOI: 10.5281/zenodo.6642503. Reads with soft clipping at either the 5′ or 3′ ends were removed. Fragment length was calculated from the Minimap2 BAM CIGAR column by summing all counts. Short mononucleosome ratio was calculated as $\frac{\text{numfrags}_{100-150\text{bp}}}{\text{numfrags}_{100-220\text{bp}}}$ (150 bp is the same cutoff for short fragments used in [35]). Short dinucleosome ratio was calculated as $\frac{\text{numfrags}_{275-325\text{bp}}}{\text{numfrags}_{275-400\text{bp}}}$ (this was determined visually from Fig. 2D of publication [35]).

End-motif analysis

BAM files from 2019 real-time base calling and alignment or 2022 HAC base calling and alignment above were used as input. Fragments and reads were processed and filtered as in fragment length analysis. For cfNano, we only used read end1 because end2 could occasionally not represent the actual end of the fragment. To avoid biases that would affect end-motif analysis, we also removed reads with any soft clipping at end 1. The first

4 bases of each fragment were extracted and used for 4-mer analysis. To avoid errors in Nanopore base calling, these 4 bases were extracted from the reference genome. Motif frequency was calculated as $\frac{\text{numfrags}_{4\text{-mer}}}{\text{numfrags}_{\text{total}}}$. For the top 25 motifs and ranking order in Fig. 4 and Additional file 1: Fig. S10, we used [36]. Files and pipelines used for Fragment length and end-motif analyses are available at https://github.com/Puputnik/Fragmentomics_GenomBiol

Statistical tests

Student's *t*-test for all sample comparisons where at least one test group had less than five samples, otherwise the Wilcoxon test was used.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-022-02710-1>.

Additional file 1. Supplementary Figs. S1-S10.

Additional file 2. Table S1. cfNano sample information and statistics.

Additional file 3. Table S2. External WGBS sample information.

Additional file 4. Table S3. Tumor vs. normal differences for 4-mer end motifs.

Additional file 5. ichorCNA plots for all cfNano and matched Illumina WGS samples.

Additional file 6. Review history.

Acknowledgements

We thank Joshua Moss and Yaping Liu for the critical review of the manuscript, Sara Isaac and Myriam Maoz for the assistance with sample processing, and Tiago Silva and Irene Unterman for useful discussions. We thank the Dennis Lo lab for sharing healthy control cfDNA data from EGAD00001001602 and the Le Son Tran lab for sharing healthy and LuAd plasma cfDNA data from 10.6084/m9.figshare.16817941.v1. We thank the two anonymous reviewers for constructive comments that improved the final version of the paper. Computation was performed on the Hebrew University Research Computing Services cluster, with assistance from Yaron Weitz and Ori Adam.

Peer review information

Barbara Cheifet and Kevin Pang were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 6.

Authors' contributions

BPB, FM, and SGC conceived the project. BPB, EK, SO, FM, and SGC designed the computational analyses. BPB, EK, SO, and FM performed the computational and statistical analyses and generated the figures. IP provided cancer plasma samples and associated clinical information. IFF, RS, and YD provided healthy plasma samples and associated clinical information. FM, AZ, and AE performed the ONT sequencing. BPB wrote the manuscript, with contributions from SGC, FM, EK, and SO. BPB and SGC co-supervised the project. The first three authors (EK, SO, FM) are co-lead authors with an equal contribution to the work, and they have the right to list their names first in their CVs. The last two authors (SGC, BPB) are co-senior authors who jointly supervised the work, and they have the right to list their names last in their CVs. The authors read and approved the final manuscript.

Authors' information

Twitter handles: @Filo_Mart (Filippo Martignano); @silvoLab (Silvestro G. Conticello); @benbfly (Benjamin P. Berman)

Funding

Ben Berman received startup support from The Hebrew University, the Kamea B program of the Israel Ministry of Aliyah and Immigrant Integration, and the Beethoven Foundation. This work was funded by a project grant to Berman from the Israel Cancer Research Fund Project Grant (Grant #845755). Filippo Martignano was supported by the Italian Ministry of Health (SG-2019-12370279).

Availability of data and materials

Source code availability:

Source code for fragmentomic analysis is available at https://github.com/Puputnik/Fragmentomics_GenomBiol. Source code for combined CNV and PMD methylation analysis is available at https://github.com/Puputnik/CNV_Methylation_Genome_Bio_2022. Source code for methylation deconvolution is available at <https://github.com/methylgrammarlab/cfdna-ont> [47]. An archive of all source code is deposited in Zenodo DOI:10.5281/zenodo.6641763 [48]. All source code is open-source under the MIT License.

Data availability:

All processed data files for the analyses described here are available at Zenodo DOI: 10.5281/zenodo.6642503 [49]. These include “anonymized BAM files” generated by Megalodon for all cfNano samples. Anonymized BAM files contain read-specific DNA methylation calls, but subject genetic variants have been removed. We also include methylation BED files for both Megalodon and DeepSignal. The same processed files are available from GEO, accession GSE185307 [50], but the DNA modification data is automatically removed from BAM files in GEO (we therefore recommend the Zenodo version). Raw data files for cfNano samples are deposited in EGAD00001006888 [51].

Declarations

Ethics approval and consent to participate

The study protocol for the Italian samples was reviewed and approved by the Pisa University Hospital Ethics Committee (approval number 363/2014), and written informed consent was provided by each patient. The study for the Israeli samples was conducted according to the protocols approved by the Institutional Review Board at Hadassah Medical Center (HMO-14-0198), with procedures performed in accordance with the Declaration of Helsinki. Blood and tissue samples were obtained from donors who have provided written informed consent.

Consent for publication

Not applicable

Competing interests

BPB, EK, SO, FM, and SGC are inventors on IP filings of this work by the Yissum Research Development Company of The Hebrew University of Jerusalem Ltd. BPB and SO receive research funding from Volition Belgium Rx for an unrelated study.

Received: 11 October 2021 Accepted: 15 June 2022

Published online: 15 July 2022

References

- Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science*. 2021;372. <https://doi.org/10.1126/science.aaw3616>.
- Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol*. 2021;32:1167–77.
- Dor PY, Cedar PH. Principles of DNA methylation and their implications for biology and medicine. *Lancet*. 2018;392:777–86.
- Zemmour H, Planer D, Magenheimer J, Moss J, Neiman D, Gilon D, et al. Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA. *Nat Commun*. 2018;9:1443.
- Moss J, Magenheimer J, Neiman D, Zemmour H, Loyfer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun*. 2018;9:5068.
- Andargie TE, Tsuji N, Seifuddin F, Jang MK, Yuen PS, Kong H, et al. Cell-free DNA maps COVID-19 tissue injury and risk of death and can cause tissue injury. *JCI Insight*. 2021;6. <https://doi.org/10.1172/jci.insight.147610>.
- Shen SY, Singhanian R, Fehring G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*. 2018;563:579–83.
- Siejka-Zielińska P, Cheng J, Jackson F, Liu Y, Soonawalla Z, Reddy S, et al. Cell-free DNA TAPS provides multimodal information for early cancer detection. *Sci Adv*. 2021;7:eabh0534.
- Yuen ZW-S, Srivastava A, Daniel R, McNevin D, Jack C, Eyras E. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat Commun*. 2021;12:3438.
- Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol*. 2021;22:295.
- Dogan H, Patel A, Herold-Mende C, Pfister S, Wick W, Loose M, et al. P07.04 Rapid-CNS2: rapid comprehensive adaptive nanopore-sequencing of CNS tumors, a proof of concept study. *Neuro Oncol*. 2021;23:ii25–6.
- Djirackor L, Halldorsson S, Niehusmann P, Leske H, Capper D, Kuschel LP, et al. Intraoperative DNA methylation classification of brain tumors impacts neurosurgical strategy. *Neurooncol Adv*. 2021;3:vdab149.
- Fox-Fisher I, Piyanzin S, Ochana BL, Klochendler A, Magenheimer J, Peretz A, et al. Remote immune processes revealed by immune-derived circulating cell-free DNA. *Elife*. 2021;10. <https://doi.org/10.7554/eLife.70520>.
- Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun*. 2021;12:1–14.
- 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.
- Cheng SH, Jiang P, Sun K, Cheng YKY, Chan KCA, Leung TY, et al. Noninvasive prenatal testing by nanopore sequencing of maternal plasma DNA: feasibility assessment. *Clin Chem*. 2015;61:1305–6.
- 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.
- Baslan T, Kovaka S, Sedlazeck FJ, Zhang Y, Wappel R, Tian S, et al. High resolution copy number inference in cancer using short-molecule nanopore sequencing. *Nucleic Acids Res*. 2021;49:e124.
- 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 U S A*. 2021;118. <https://doi.org/10.1073/pnas.2114937118>.

20. Nguyen H-N, Cao N-PT, Van Nguyen T-C, Le KND, Nguyen DT, Nguyen Q-TT, et al. Liquid biopsy uncovers distinct patterns of DNA methylation and copy number changes in NSCLC patients with different EGFR-TKI resistant mutations. *Sci Rep*. 2021;11:1–12.
21. 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.
22. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell*. 2018;33:676–89.e3.
23. Lubotzky A, Zemmour H, Neiman D, Gotkine M, Loyfer N, Piyanzin S, et al. Liquid biopsy reveals collateral tissue damage in cancer. *JCI Insight*. 2022;7. <https://doi.org/10.1172/jci.insight.153559>.
24. Ni P, Huang N, Zhang Z, Wang D-P, Liang F, Miao Y, et al. DeepSignal: detecting DNA methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics*. 2019;35:4586–95.
25. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell*. 2021;184:5985–6001.e19.
26. Maeda Y, Davé V, Whitsett JA. Transcriptional control of lung morphogenesis. *Physiol Rev*. 2007;87:219–44.
27. Eraslan G, Drokhllyansky E, Anand S, Subramanian A, Fiskin E, Slyper M, et al. Single-nucleus cross-tissue molecular reference maps to decipher disease gene function. *bioRxiv*. 2021:2021.07.19.452954 Available from: <https://www.biorxiv.org/content/10.1101/2021.07.19.452954v1.abstract>. Cited 2021 Sep 26.
28. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. *Science*. 2018;362. <https://doi.org/10.1126/science.aav1898>.
29. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet*. 2018;50:591–602.
30. Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A*. 2015;112:E5503–12.
31. Chan KCA, Jiang P, Chan CWM, Sun K, Wong J, Hui EP, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci U S A*. 2013;110:18761–8.
32. López-Moyado IF, Tsagaratou A, Yuita H, Seo H, Delatte B, Heinz S, et al. Paradoxical association of TET loss of function with genome-wide DNA hypomethylation. *Proc Natl Acad Sci U S A*. 2019;116:16933–42.
33. Farhangdoost N, Horth C, Hu B, Bareke E, Chen X, Li Y, et al. Chromatin dysregulation associated with NSD1 mutation in head and neck squamous cell carcinoma. *Cell Rep*. 2021;34:108769.
34. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res*. 2012;22:2497–506.
35. Moulriere 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. <https://doi.org/10.1126/scitranslmed.aat4921>.
36. 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.
37. Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov*. 2020;10:664–73.
38. Foox J, Tighe SW, Nicolet CM, Zook JM, Byraska-Bishop M, Clarke WE, et al. Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. *Nat Biotechnol*. 2021;39:1129–40.
39. Zviran A, Schulman RC, Shah M, Hill STK, Deochand S, Khamnei CC, et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med*. 2020;26:1114–24.
40. 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. *bioRxiv*. 2022:2022.01.17.476508 Available from: <https://www.biorxiv.org/content/10.1101/2022.01.17.476508v1.abstract>. Cited 2022 Apr 11.
41. Erger F, Nörling D, Borchert D, Leenen E, Habbig S, Wiesener MS, et al. cfNOME—a single assay for comprehensive epigenetic analyses of cell-free DNA. *Genome Med*. 2020;12:1–14.
42. Cuadrat RRC, Kratzer A, Arnal HG, Wreczycka K, Blume A, Ebenal V, et al. Cardiovascular disease biomarkers derived from circulating cell-free DNA methylation. *medRxiv*. 2021;2021(11):05.21265870.
43. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*. 2014;30:2503–5.
44. Baslan T, Kovaka S, Sedlazeck FJ, Zhang Y, Wappel R, Lowe SW, et al. High resolution copy number inference in cancer using short-molecule nanopore sequencing. *bioRxiv*. 2020:2020.12.28.424602 Available from: <https://www.biorxiv.org/content/10.1101/2020.12.28.424602v1.abstract>. Cited 2021 Sep 19.
45. Silva TC, Coetzee SG, Gull N, Yao L, Hazelett DJ, Noushmehr H, et al. ELMER v2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics*. 2019;35:1974–7.
46. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A cell atlas of chromatin accessibility across 25 adult human tissues. *bioRxiv*. 2021; Available from: <https://www.biorxiv.org/content/early/2021/02/17/2021.02.17.431699>.
47. Katsman E, Orlanski S, Martignano F, Conticello SG, Berman, BP. Github source code from Katsman et al. Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. 2021. Available from: https://github.com/Puputnik/Fragmentomics_GenomBiol, https://github.com/Puputnik/CNV_Methylation_Genome_Biol_2022, <https://github.com/methylgrammarlab/cfdna-ont>
48. Katsman E, Orlanski S, Martignano F, Conticello SG, Berman, BP. Static Zenodo version of all source code from Katsman et al. Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. 2022. <https://doi.org/10.5281/zenodo.6641763>.
49. Katsman E, Orlanski S, Martignano F, Eden E, Conticello SG, Berman, BP. Processed data files in Zenodo from Katsman et al. Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. 2022. <https://doi.org/10.5281/zenodo.6642503>.

50. Katsman E, Orlanski S, Martignano F, Eden E, Conticello SG, Berman, BP. Processed data files in the GEO database from Katsman et al. Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. 2021. Available from: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE185307>.
51. Martignano F, Munagala U, Crucitta S, Mingrino S, Semeraro R, Del Re M, Petrini I, Magi A, Katsman E, Orlanski S, Eden E, Berman, BP, Conticello SG. Raw plasma cfDNA data files in EGA from Martignano et al. 2021 and Katsman et al. 2022. 2021. Available from: <https://ega-archive.org/datasets/EGAD00001006888>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

