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

Efrat Katsman[1,*], Shari Orlanski[1,*], Filippo Martignano[2,*], Amir Eden[3], Iacopo Petrini[4], Silvestro G. Conticello[2,5,#], Benjamin P. Berman[1,#]

1. Department of Developmental Biology and Cancer Research, Hebrew University of Jerusalem, Faculty of Medicine. Jerusalem, Israel

2. Core Research Laboratory, ISPRO, Florence, Italy.

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

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

5. Institute of Clinical Physiology, National Research Council, Pisa, Italy.

* Equal contribution

Jointly supervised project

Correspondence to: s.conticello@ispro.toscana.it, ben.berman@mail.huji.ac.il

Abstract

DNA methylation (5mC) is a promising biomarker for detecting circulating tumor DNA (ctDNA), providing information on a cell's genomic regulation, developmental lineage, and molecular age. Sequencing assays for detecting ctDNA methylation involve pre-processing steps such as immunoprecipitation, enzymatic treatment, or the most common method, sodium bisulfite treatment. These steps add complexity and time that pose a challenge for clinical labs, and bisulfite treatment in particular degrades input DNA and can result in loss of informative ctDNA fragmentation patterns. In this feasibility study, we demonstrate that whole genome sequencing of circulating cell-free DNA using conventional Oxford Nanopore Technologies (ONT) sequencing can accurately detect cell-of-origin and cancer-specific 5mC changes while preserving important fragmentomic information. The simplicity of this approach makes it attractive as a liquid biopsy assay for cancer as well as non-cancer applications in emergency medicine.

Introduction

Cell-free DNA captures informative features of its originating cell, which include genomic alterations, DNA modifications such as 5mC, fragmentation patterns due to differential DNase activities, and nucleosomal organization (*1*). One of the most promising cfDNA biomarkers for cancer is 5mC, which has been validated in a large clinical study and is now in widespread use for cancer detection (*2*). Unlike other cancer-specific cfDNA biomarkers, 5mC can detect the presence of other unusual cell types in cfDNA to detect non-cancer conditions including myocardial infarction and sepsis (*3*). Most of these studies have used bisulfite-based approaches, but immunoprecipitation-based (*4*) and enzymatic (*5*) techniques have also shown promising results.

Native sequencing with the ONT platform is attractive for a number of reasons. First, single base pair resolution DNA methylation calling on the Nanopore platform has improved significantly in the past several years, and now achieves high concordance with the gold standard whole-genome bisulfite sequencing (WGBS) in several benchmarking studies (*6*, *7*). ONT sequencing is also rapid, with recent clinical demonstrations of end to end turnaround time from sample collection to DNA methylation-based classification in as little as 1-3 hours (*8*, *9*). Other benefits of ONT for clinical settings include the low buy-in cost and portable nature of the device. ONT native WGS is unique among DNA methylation sequencing approaches in that it does not require a PCR amplification step, which can bias both fragmentation patterns and uniformity of genomic coverage.

ONT sequencing has primarily been used for long-read sequencing, but recent work has shown that it can be adapted for short fragments to detect copy number alterations, where long read sequencing is not cost effective (*10*–*12*). In our recent publication (*11*), we showed that optimizations in library construction could generate 4-20 million sequencing reads from 4mL of plasma of healthy and cancer patients. Here, we perform additional analysis on that same dataset to extract 5mC and fragmentomic information which we did not investigate previously.

Results

All analyses described below are from sequence data generated in our previous publication, which used cfDNA extracted from 4mL of plasma for four healthy control individuals and six metastatic lung adenocarcinoma cases (*11*). Six of the samples had between 3.8M and 5.3M raw reads (2.2M-2.6M uniquely mapping reads), and the remaining four samples had between 8.4M and 20.2M raw reads (4.8M-11.2M uniquely mapping reads). All sequencing statistics are available in Supplemental Table 1. We used ichorCNA (*13*) to estimate the tumor fraction of each sample using somatic copy number alterations (SCNAs) (Supplemental Table 1). Four of the six cancer cases had tumor fraction estimates greater than 0.1 (high tumor fraction), one case had 0.086 (BC09), and another (BC08) was under the detection limit for TF estimation via IchorCNA so tumor fraction was set to 0 (Figure 1A-1B, top).

Nanopore DNA methylation detects cancer-specific and cell-of-origin of ctDNA

Global DNA hypomethylation is one of the hallmarks of the cancer epigenome and has been proposed as a general ctDNA biomarker (*14*), and was recently verified in WGBS of cfDNA from NSCLC cases (*15*). In order to investigate this, we processed the original fast5 sequencing files with DeepSignal (*16*) to call methylation at individual CpGs. The six lower coverage samples covered between 4.3M and 5.5M CpGs (usually by a single read per CpG), while the remaining four samples covered 8.1M-18.9M CpGs (Supplemental Table 1). Next, we calculated global methylation within 10 Mbp genomic windows genome-wide. This analysis showed high methylation levels for the four healthy control plasmas and the two low tumor fraction cases, and significantly reduced methylation for three of the four high tumor fraction samples (Figure 1A, bottom). Reasoning that regions of copy number alteration would have skewed proportions of tumor-derived DNA and thus skewed methylation levels, we split out methylation by SCNA status for all cancer samples. In the three cases with globally reduced methylation (BC01, 19_326, BC10), amplified regions were significantly more hypomethylated than diploid regions, as expected (Figure 1B). While hypomethylation could not be detected genome-wide in the low tumor fraction sample BC08, amplified regions were significantly hypomethylated. Conversely, deleted regions showed reduced hypomethylation relative to diploid regions, but this trend only reached statistical significance in two of the three cases with global hypomethylation (19_326 and BC10). In the final case (BC11), DNA methylation overall was higher than in healthy plasma, and SCNAs levels suggested this was due specifically to the high methylation of cancer-derived DNA (Figure 1B). While this is an interesting case, it is not surprising given the high degree of variability associated with global hypomethylation (*17*), a process that is not entirely understood but is known to be affected by various chromatin modifiers that are dysregulated in cancer(*18*, *19*).

“Global” cancer hypomethylation is not truly global and occurs primarily within long regions of lamina-associated heterochromatin called Partially Methylated Domains (PMDs)(*17*). In all of our hypomethylated samples, hypomethylation was concentrated within previously identified PMDs from (*17*) (Figure 1C). When considering only bins within PMDs, significant hypomethylation was identified not only within the three cases where it was significant genome-wide, but also in the two low tumor fraction cases (Supplemental Figure 1A). The same

association between SCNAs and hypomethylation that occurred genome-wide analysis (Figure 1B) was also significant in the PMD-only analysis (Supplemental Figure 1B).

Since global hypomethylation is a relatively generic cancer change, we next sought to investigate regions marking the cell-of-origin of lung adenocarcinoma cells. The lack of a suitable whole-genome DNA methylation dataset for lung epithelia prompted us to use regulatory regions defined by ATAC-seq, since ATAC-seq open chromatin regions are almost universally demethylated in cancer (*20*). A recent single-cell ATAC-seq atlas identified open chromatin regions in 25 distinct human tissue types from multiple donors, and identified a strong cluster of lung pneumocytes (the “Pal” cluster) in primary lung samples (*21*). NKX2-1 is a known master regulator transcription factor in lung pneumocytes (*22*), and the binding site for NKX2-1 was the most enriched motif within this pneumocyte-specific cluster of scATAC-seq peaks (*21*). NKX2-1 expression also has highly restricted expression across all known organs (*23*), making it an ideal marker for lung pneumocyte cell-of-origin analysis. Predicted NKX2-1 binding sites are the most enriched motifs in open chromatin of TCGA lung adenocarcinoma tumors (*20*), suggesting they are not only a good cell type marker but also a good marker of this cancer type.

To analyze NKX2-1 binding site DNA methylation, we first identified the 5,974 predicted binding NKX2-1 sites within pneumocyte-specific (“Pal”) ATAC-seq peaks from (*21*) (Figure 1D). We confirmed lung cancer specificity using the TCGA WGBS dataset from (*17*), which contained 9 NSCLC samples and 18 other samples from four other non-lung epithelial cancer types (Breast, Colorectal, Stomach, and Endometrial). NKX2-1 sites showed almost no demethylation in non-lung tumors (Figure 1E, left), but substantial demethylation in both lung lung tumors and adjacent normal lung tissue, with lung adenocarcinomas having the strongest demethylation (Figure 1E, right). We next looked at methylation in plasma cfDNA from published studies using Illumina WGBS. Neither Healthy plasmas, liver cancer, nor colorectal cancer plasmas showed demethylation, confirming the lung specificity of these NKX2-1 sites (Figure 1F). In our Nanopore WGS samples, plasma from healthy individuals showed no demethylation (Figure 1G, left), but at least three of the four cancer samples with high tumor fraction cancer samples were demethylated (Figure 1G, right, and Supplemental Figure 2). This analysis shows that shallow Nanopore WGS can detect highly cell type specific features of the cancer cell-of-origin.

The quantitative nature of DNA methylation allows accurate estimates of cell type mixtures from reference datasets of pure cell types (*24*), including applications to cfDNA (*25*). While there is currently no whole-genome methylation dataset that includes pure lung epithelial cells, such a dataset was recently generated on the Illumina HumanMethylation450k (HM450k) platform (*25*). We adapted the non-negative least squares (NNLS) regression method used in (*25*) to deconvolute our Nanopore plasma samples into lung cell and healthy plasma cell type components (Figure 1H). We had to significantly expand the number of cell-type specific marker CpGs used in the MethAtlas paper (*25*), due to the relatively low degree of overlap between HM450k probes and CpGs called in our Nanopore samples (the majority of Nanopore samples overlapped less than 20% of HM450k probes, see Supplemental Table 1). We identified a total of 4,355 lung-specific marker CpGs (Supplemental Figure 3), which covered a median of 818 CpGs per sample (Supplemental Table 1). For example, healthy sample BC05 overlapped 1,760 lung-specific CpGs, while cancer sample BC11 overlapped 1,251 (Figure 1H, Supplemental Figure 4, Supplemental Table 1). These were used as input to NNLS regression

to estimate the lung cell fraction (β) and healthy plasma cell fraction (1-β) for all Nanopore samples (Figure 1H).

We used the 4,355 lung-specific marker CpGs and NNLS regression to estimate lung cell fraction (β) for all Nanopore samples, which yielded 100% separation between healthy plasma and lung cancer samples (Figure 1I-J). We evaluated the quantitative accuracy of these estimates by comparing to the ichorCNA tumor fraction estimates (Figure 1K). While these two estimates are based on completely independent features, they showed overall strong agreement (PCC 0.884). One case, BC08, had a lower read count (2.6M reads) and ichorCNA failed to detect SCNAs in this case, leading to a tumor fraction estimate of 0. We sequenced BC08 with higher coverage using Illumina WGS (17M uniquely alignable read pairs), which allowed ichorCNA to detect sufficient SCNAs for a valid tumor fraction estimate of 0.11 (Supplemental Figure 4, Supplemental Table 1). This was extremely close to the methylation-based estimate of BC08 from Nanopore data (Figure 1K, yellow point), suggesting that Nanopore DNA methylation can be a more sensitive ctDNA detector then SCNAs for cases with low tumor fraction or few SCNAs.

To verify the robustness of the NNLS deconvolution results, we performed the same analysis using a mutually exclusive set of 14,654 HM450k marker probes differentially methylated between TCGA LUAD tumors and healthy plasma (Figure 1L). This analysis yielded very similar results to the normal lung-based analysis (Figure 1M-N), reinforcing the idea that circulating tumor DNA can be detected using reference data from either the appropriate normal cell type or from tumors (*3*). This analysis revealed an interesting outlier, BC10, which had a methylation-based lung cell estimate that was almost 2-fold higher than the ichorCNA estimate in both the normal lung based (Figure 1K) and tumor based (Figure 1N) estimates. Interestingly, the NKX2-1 methylation analysis agreed with these deconvolution results, with BC10 showing the highest degree of NKX2-1 demethylation of any cancer sample (Supplemental Figure 2). While it would require study in a larger cohort, it is possible that this represents a case where low read coverage (2.6M reads) leads SCNA analysis to fail at detecting a whole-genome doubling event, and that the model could be improved by incorporating DNA methylation data.

Nanopore preserves fragmentomic features of ctDNA

Tumor-derived cfDNA can be distinguished by several DNA fragmentation features, including shorter fragment lengths and altered fragment end motifs. These features likely reflect the specific DNase enzymes present in the cancer cells as well as the chromatin organization in those cells (reviewed in (*1*)). We were able to investigate these fragmentation features for nine of the ten Nanopore samples that we previously sequenced. The tenth sample, 19_326, was generated with a different library construction kit that affects fragment size representation as well as adapter trimming. It was thus not included in our primary fragmentomic analyses (we analyzed it separately in Supplemental Figure 5A-G).

Cancer-derived circulating cfDNA fragments tend to be shorter than those from healthy individuals, with an overabundance of fragments of length <150 bp (*26*), and these fragment length differences can classify different cancer types (*27*). We compared fragment lengths in our Nanopore samples, and indeed found that high tumor fraction samples had shorter fragments

than the healthy samples, using the criteria developed by (*26*) (Figure 2A-B). The two low tumor fraction samples did not have detectably shorter fragments.

We also investigated the four bases immediately flanking cfDNA fragmentation sites, as these have been shown to have biased sequence composition which are predictive of cancer (*28*, *29*). To compare these biases between Illumina and Nanopore end motif frequencies, we first sorted the 25 most frequent 4-mers from a prior Illumina-based study of healthy plasma (*29*). Visualizing 4-mer frequencies using this ordering showed that seven of the top eight 4-mers from the previous study were also top ranked in our Nanopore samples as well as our Illumina WGS samples (Figure 2C). CCCA was the most frequent 4-mer motif in both our Nanopore and Illumina samples, consistent with earlier studies of healthy plasma (*28*, *29*). In a previous study, CCCA had a significantly lower frequency in lung and four other common cancers than in healthy plasma (*28*). Consistent with this, we found that CCCA was significantly lower in our cancer samples, most notably our high tumor fraction samples (Figure 2C-D, Supplemental Figure 6). That same previous study (*28*) highlighted two other cancer-increased and two cancer-decreased motifs, and all of these followed the same trend in our Nanopore samples, with two of the four (CCTG and AAAA) rising to statistical significance in our small sample set (Supplemental Figure 6). Despite these similarities, the overall frequencies show clear differences between Nanopore and Illumina at several 4-mers such as CCAA (Figure 2C). Future work will be required to determine which sequencing technology gives more accurate representations, but the absence of PCR bias in Nanopore sequencing could be a determining factor.

Cell-free DNA circulates primarily as mono-nucleosomal fragments, and nucleosome positions inferred from fragment cut sites can be used to detect cell-of-origin (reviewed in (*1*)). Bisulfite conversion used with Illumina-based sequencing can degrade these fragmentation patterns (*30*). We have previously showed that mono-nucleosomal fragment lengths are largely preserved in Nanopore cfDNA sWGS (*11*), but we wanted to further investigate the ability of Nanopore to reveal biologically relevant nucleosome structure based on fragmentation patterns. CTCF binding sites present the best model for nucleosome organization - they position 10 phased nucleosomes on either side of a central binding site, which itself lies within a 100bp nucleosome depleted region (*31*). The Nanopore fragmentation pattern around CTCF binding sites recapitulated this structure (Figure 2E, top), and reproduced the pattern based on our deeper Illumina WGS sequencing (Figure 2E, bottom). The CTCF binding site is also known to sit at the center of a 400bp demethylated region, and this was also recapitulated from 5mC levels from our Nanopore sequencing (Figure 2F, top). We verified this pattern in lung adenocarcinoma tumors using bisulfite-seq (WGBS) data from TCGA (*17*) (Figure 2G). CTCF binding is largely not cell-type specific, and thus we did not observe any differences between healthy and cancer samples in this analysis. We tried the same fragment coverage analysis for the lung-specific NKX2-1 binding sites discussed earlier, but we were not able to detect any nucleosome structure (Supplemental Figure 6). It is possible that higher read depth will be necessary to detect nucleosomal fragmentation patterns from cell-type specific components of cfDNA, but new computational methods such as Griffin (*32*) may allow for more sensitive detection in the future.

Discussion

While this is only a small feasibility demonstration, the results are very encouraging. We were able to detect cell-type specific and cancer-specific DNA methylation patterns that recapitulate known patterns from Illumina-based Bisulfite sequencing (WGBS), as well as cancer-specific fragmentation signatures of Illumina-based WGS.

The ability to independently estimate tumor DNA fraction using either (1) ichorCNA or (2) methylation-based cell type deconvolution was somewhat limited by the low sequencing depth (median 6.4M reads in this study). Nevertheless, these two estimates were in very good agreement, and could potentially be combined to increase detection accuracy since they derive from highly independent features. In at least one case (BC08) and possibly a second (BC10), the DNA methylation-based estimate of tumor cell fraction appeared to be more accurate. The accuracy of DNA methylation-based detection could be significantly improved by generating whole-genome methylation atlases of purified human cell types (*33*), or by generating large-scale WGBS sequencing of human cancer (such a dataset was described in (*2*), but remains proprietary). The loci we used here for deconvolution were based on Illumina HM450k reference data, which covers less than 3% of all CpGs in the genome and only about 13% of highly cell type specific methylation markers (Tommy Kaplan, personal communication).

While bisulfite-based approaches have been successful in identifying ctDNA biomarkers, the ability to analyze DNA methylation from native DNA has a number of advantages. First, bisulfite treatment can lead to significant DNA loss (especially relevant within the limitations of clinical samples) and the loss of informative fragmentation features (*30*). Second, the requirement for PCR amplification in bisulfite-based and other approaches provides less uniform representation of the genome, and could skew or overshadow informative fragmentation patterns. Third, bisulfite-based approaches do not differentiate between 5mC and the other informative CpG modification 5hmC. Nanopore can identify and distinguish both of these two marks, and potentially in the future additional modifications such as 5fC and 5CaC.

The Nanopore platform could have other practical advantages for the clinical setting. Current high-throughput sequencing technologies allow for reasonable per assay costs only with a large capital equipment investment and in a high throughput environment where many samples are available for multiplexing. Nanopore offers an alternative with fast turnaround times for individual samples in as little as 2 hours from sample collection to DNA methylation based classification (8, 9). Since all cell types have specific DNA methylation patterns, rapid Nanopore sequencing could have a variety of biomarker applications outside of cancer, including emergency medicine applications such as myocardial infarction or sepsis (3).

Code availability

R code for deconvolution is available on https://github.com/methylgrammarlab/cfdna-ont.

Data availability

Processed data files for the analyses described here are available at GEO accession GSE185307 or Figshare .

Competing interests

BPB, EK, SO, FM, and SGC are inventors on filings for patent protection of Nanopore-based detection of cfDNA by Yissum Research Development Company of the Hebrew University of Jerusalem Ltd. Dr. Berman's lab at Hebrew University receives funding from Volition Belgium Rx, for a collaboration on fragmentomic analysis of cfDNA.

Acknowledgments

We acknowledge Tiago Silva, Irene Unterman, and Yuval Dor for helpful discussions, and Joshua Moss, Yaping Liu, and Aviad Zick for critical review of the manuscript. We thank Jasmine Zhou lab for sharing HCC cfDNA data EGAD00001004317, the Dennis Lo lab for sharing CRC and healthy control cfDNA data EGAD00001004568 and EGAD00001001602. Computation was carried out on the Hebrew University Research Computing Services cluster, and we acknowledge Yaron Weitz for his help and support. Dr. Berman's lab received startup support from the Kamea B program of the Israel Ministry of Aliyah and Immigrant Integration, and from the Beethoven Foundation. Dr. Martignano was supported by the Italian Ministry of Health (SG-2019-12370279). Yitzhak Yadgari contributed artwork.

Author Contributions

BPB, FM, and SGC conceived the project. BPB, EK, SO, FM, and SGC designed the analyses. BPB, EK, SO, and FM performed computational and statistical analyses and generated figures. BPB drafted the initial manuscript, and SGC, EK, SO, and FM contributed. JP provided samples and clinical information and expertise. AE provided critical feedback on ONT sequencing and DNA methylation analysis. BPB and SGC co-supervised the project.

Figure legends

Figure 1: Nanopore DNA methylation detects cancer-specific and cell-of-origin of ctDNA features. (A) ONT cfDNA WGS of 4 healthy controls and 6 lung adenocarcinoma cases. (A, top) Tumor fraction was calculated using ichorCNA (*13*). (A, bottom) Average DNA methylation of all 10 Mbp bins (removing CpG islands.) (B) 10 Mbp bins stratified by copy number for all cancer cases. (C) Average DNA methylation across chr16p, comparing lung tissue WGBS from the TCGA project (top) to plasma cfDNA ONT samples from this study (bottom). Common Partially Methylated Domains (PMDs) from an earlier study from an earlier study (*17*) are shown as a reference. (D) Illustration of NKX2-1 binding sites within pneumocyte-specific ATAC-seq peaks taken from (*21*). (E) Relative methylation levels within -1kb to +1kb of pneumocyte-specific NKX2-1 sites, for 18 TCGA WGBS non-lung tumors (left) and 11 TCGA WGBS lung tumors and adjacent normal tissue (right) from (*17*). Non-lung tumors included 5 Breast (BRCA), 4 Colorectal (CRC), 4 Stomach (STAD), and 5 endometrial (UCEC). Relative methylation was defined as raw methylation divided by the mean methylation from -1,000 to -800 and +800 to +1,000 across all NKX2-1 sites. (F) NKX2-1 relative methylation for previously published cfDNA methylation studies using WGBS. (G) NXK2-1 relative methylation for healthy control plasmas (blue) and lung cancer plasmas (red) from this study, grouped by ichorCNA tumor fraction estimates. (H) Deconvolution of ONT methylation profiles into lung component and healthy cell-of-origin component. Reference datasets using the HM450k platform are sorted lung epithelial cells (X_1) and healthy plasma (X_2) from (*25*). These are used to identify lung-hypermethylatedCpGs (top) and lung-hypomethylated CpGs (bottom). Two ONT cfDNA samples are shown, healthy BC05 (left) and cancer BC11 (right). Each ONT sample overlaps different subsets of hyper- and hypomethylated CpGs, with BC05 overlapping 846+914=1,760 CpGs and BC11 overlapping 604+647=1,251 CpGs. Lung fractions β are estimated using non-negative least squares (NNLS) regression. (I) Full set of 4,355 differentially methylated CpGs used for normal lung NNLS analysis, showing 884 CpG overlap with differentially methylated CpGs from TCGA lung adenocarcinoma (LUAD) tumors. (J) Estimated lung fraction β for all Nanopore plasma samples. (K) Estimated lung fraction β plotted against ichorCNA tumor fractions. SCNAs were undetectable in BC08 leading to an ichorCNA estimate of 0. For this sample, we performed higher depth Illumina sequencing where ichorCNA estimated tumor fraction as 0.11 (shown as a yellow circle “β vs. deep WGS CNA”). (L-N) Same methods as (I-K) except using 13,770 differentially methylated CpGs from TCGA LUAD tumors (14,654 total minus 884 probes overlapping normal lung CpGs). Statistical significance for panels A and B determined by one-tailed wilcoxon test. Statistical significance for J and M determined by two-tailed Student’s t-test. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

Figure 1: Nanopore DNA methylation detects cancer-specific and cell-of-origin ctDNA features

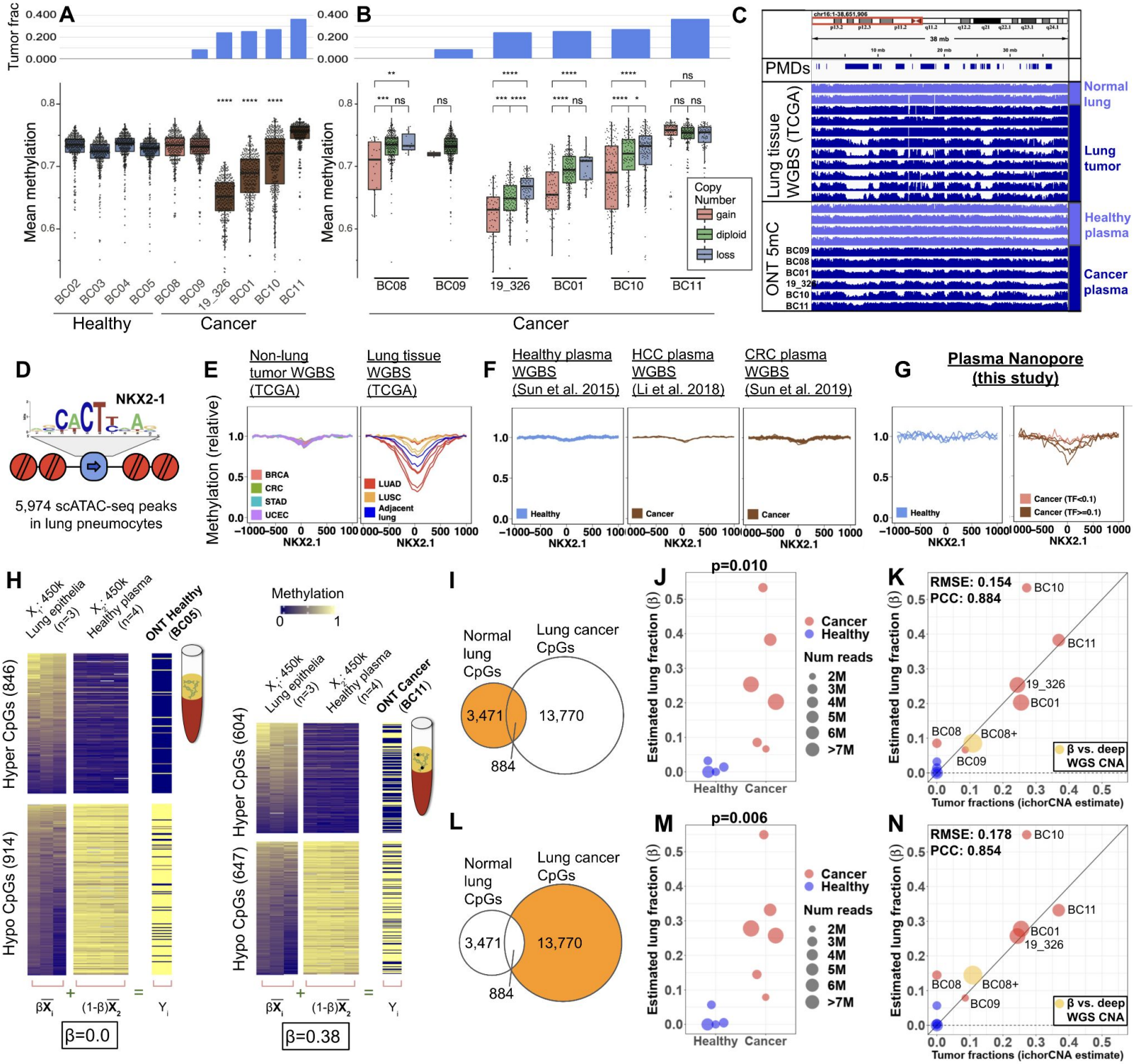
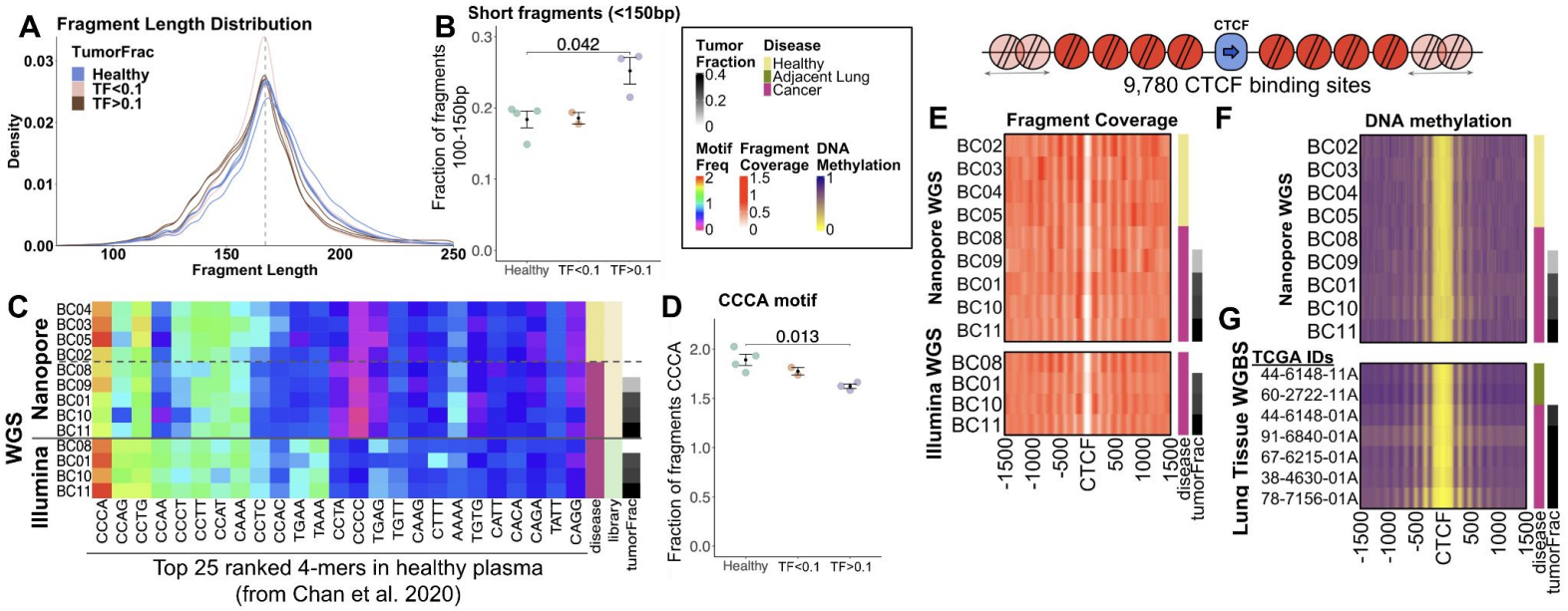


Figure 2: Nanopore preserves fragmentomic features of ctDNA. (A) Fragment length density of all samples, colored by tumor fraction estimated by ichorCNA. (B) Fragment length ratio calculated as the number of short fragments (100-150bp, same cutoff used in (*26*)) divided by the total number of total mononucleosome reads (100-220bp). (C) Frequency of different 4-mer sequences at 5' fragment ends, comparing ONT cfDNA WGS samples and matched Illumina samples. The 25 most frequent 4-mers in healthy plasma from (*29*) are shown in rank order. Samples are ordered by healthy vs. cancer and then by increasing tumor fraction. (D) Frequency of CCCA motif at 5' fragment ends. (E-G) Alignments to CTCF motifs within 9,780 distal ChIP-seq peaks from (*31*). (E, top) cfDNA fragment coverage shown as fold-change vs. average coverage depth across the genome. The plot includes only fragments of length 130-155bp to maximize resolution. (E, bottom) Matched Illumina samples of higher sequencing depth (median 17.0M fragments in Illumina vs. 6.4M in ONT samples). (F) CTCF DNA methylation of Nanopore samples from this study at CTCF sites. (G) DNA methylation from seven lung tissue WGBS samples from TCGA (*17*). Statistical significance for panels A and B was determined by two-tailed t-test.

Figure 2: Nanopore preserves fragmentomic features of ctDNA



References

1. Y. M. D. Lo, D. S. C. Han, P. Jiang, R. W. K. Chiu, Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science*. **372** (2021), doi:10.1126/science.aaw3616.

2. E. A. Klein, D. Richards, A. Cohn, M. Tummala, R. Lapham, D. Cosgrove, G. Chung, J. Clement, J. Gao, N. Hunkapiller, A. Jamshidi, K. N. Kurtzman, M. V. Seiden, C. Swanton, M. C. Liu, Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann. Oncol.* **32**, 1167–1177 (2021).

3. P. Y. Dor, P. H. Cedar, Principles of DNA methylation and their implications for biology and medicine. *Lancet*. **392**, 777–786.

4. S. Y. Shen, R. Singhania, G. Fehringer, A. Chakravarthy, M. H. A. Roehrl, D. Chadwick, P. C. Zuzarte, A. Borgida, T. T. Wang, T. Li, O. Kis, Z. Zhao, A. Spreafico, T. da S. Medina, Y. Wang, D. Roulois, I. Ettayebi, Z. Chen, S. Chow, T. Murphy, A. Arruda, G. M. O'Kane, J. Liu, M. Mansour, J. D. McPherson, C. O'Brien, N. Leighl, P. L. Bedard, N. Fleshner, G. Liu, M. D. Minden, S. Gallinger, A. Goldenberg, T. J. Pugh, M. M. Hoffman, S. V. Bratman, R. J. Hung, D. D. De Carvalho, Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*. **563**, 579–583 (2018).

5. P. Siejka-Zielińska, J. Cheng, F. Jackson, Y. Liu, Z. Soonawalla, S. Reddy, M. Silva, L. Puta, M. V. McCain, E. L. Culver, N. Bekkali, B. Schuster-Böckler, P. F. Palamara, D. Mann, H. Reeves, E. Barnes, S. Sivakumar, C.-X. Song, Cell-free DNA TAPS provides multimodal information for early cancer detection. *Sci Adv*. **7**, eabh0534 (2021).

6. Z. W.-S. Yuen, A. Srivastava, R. Daniel, D. McNevin, C. Jack, E. Eyras, Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat. Commun.* **12**, 3438 (2021).

7. Y. Liu, W. Rosikiewicz, Z. Pan, N. Jillette, P. Wang, A. Taghbalout, J. Foox, C. Mason, M. Carroll, A. Cheng, S. Li, DNA methylation calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation, , doi:10.1101/2021.05.05.442849.

8. L. P. Kuschel, J. Hench, S. Frank, I. B. Hench, E. Girard, M. Blanluet, J. Masliah-Planchon, M. Misch, J. Onken, M. Czabanka, P. Karau, N. Ishaque, E. G. Hain, F. Heppner, A. Idbaih, N. Behr, C. Harms, D. Capper, P. Euskirchen, Robust methylation-based classification of brain tumors using nanopore sequencing. *medRxiv*, 2021.03.06.21252627 (2021).

9. L. Djirackor, S. Halldorsson, P. Niehusmann, H. Leske, L. P. Kuschel, J. Pahnke, B. J. Due-Tønnessen, I. A. Langmoen, C. J. Sandberg, P. Euskirchen, E. O. Vik-Mo, EPCT-15. RAPID EPIGENOMIC CLASSIFICATION OF BRAIN TUMORS ENABLES INTRAOPERATIVE NEUROSURGICAL RISK MODULATION. *Neuro. Oncol.* **23**, i50–i50 (2021).

10. S. H. Cheng, P. Jiang, K. Sun, Y. K. Y. Cheng, K. C. A. Chan, T. Y. Leung, R. W. K. Chiu, Y. M. D. Lo, Noninvasive prenatal testing by nanopore sequencing of maternal plasma DNA: feasibility assessment. *Clin. Chem.* **61**, 1305–1306 (2015).

11. F. Martignano, U. Munagala, S. Crucitta, A. Mingrino, R. Semeraro, M. Del Re, I. Petrini, A. Magi, S. G. Conticello, Nanopore sequencing from liquid biopsy: analysis of copy number

variations from cell-free DNA of lung cancer patients. *Mol. Cancer*. **20**, 32 (2021).

12. T. Baslan, S. Kovaka, F. J. Sedlazeck, Y. Zhang, R. Wappel, S. W. Lowe, S. Goodwin, M. C. Schatz, High resolution copy number inference in cancer using short-molecule nanopore sequencing. *bioRxiv* (2020), p. 2020.12.28.424602.

13. V. A. Adalsteinsson, G. Ha, S. S. Freeman, A. D. Choudhury, D. G. Stover, H. A. Parsons, G. Gydush, S. C. Reed, D. Rotem, J. Rhoades, D. Loginov, D. Livitz, D. Rosebrock, I. Leshchiner, J. Kim, C. Stewart, M. Rosenberg, J. M. Francis, C.-Z. Zhang, O. Cohen, C. Oh, H. Ding, P. Polak, M. Lloyd, S. Mahmud, K. Helvie, M. S. Merrill, R. A. Santiago, E. P. O'Connor, S. H. Jeong, R. Leeson, R. M. Barry, J. F. Kramkowski, Z. Zhang, L. Polacek, J. G. Lohr, M. Schleicher, E. Lipscomb, A. Saltzman, N. M. Oliver, L. Marini, A. G. Waks, L. C. Harshman, S. M. Tolaney, E. M. Van Allen, E. P. Winer, N. U. Lin, M. Nakabayashi, M.-E. Taplin, C. M. Johannessen, L. A. Garraway, T. R. Golub, J. S. Boehm, N. Wagle, G. Getz, J. C. Love, M. Meyerson, Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).

14. K. C. A. Chan, P. Jiang, C. W. M. Chan, K. Sun, J. Wong, E. P. Hui, S. L. Chan, W. C. Chan, D. S. C. Hui, S. S. M. Ng, H. L. Y. Chan, C. S. C. Wong, B. B. Y. Ma, A. T. C. Chan, P. B. S. Lai, H. Sun, R. W. K. Chiu, Y. M. D. Lo, Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18761–18768 (2013).

15. H.-N. Nguyen, N.-P. T. Cao, T.-C. Van Nguyen, K. N. D. Le, D. T. Nguyen, Q.-T. T. Nguyen, T.-H. T. Nguyen, C. Van Nguyen, H. T. Le, M.-L. T. Nguyen, T. V. Nguyen, V. U. Tran, B. A. Luong, L. G. H. Le, Q. C. Ho, H.-A. T. Pham, B. T. Vo, L. T. Nguyen, A.-T. H. Dang, S. D. Nguyen, D. M. Do, T.-T. T. Do, A. V. Hoang, K. T. Dinh, M.-D. Phan, H. Giang, L. S. Tran, Liquid biopsy uncovers distinct patterns of DNA methylation and copy number changes in NSCLC patients with different EGFR-TKI resistant mutations. *Sci. Rep.* **11**, 1–12 (2021).

16. P. Ni, N. Huang, Z. Zhang, D.-P. Wang, F. Liang, Y. Miao, C.-L. Xiao, F. Luo, J. Wang, DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*. **35**, 4586–4595 (2019).

17. W. Zhou, H. Q. Dinh, Z. Ramjan, D. J. Weisenberger, C. M. Nicolet, H. Shen, P. W. Laird, B. P. Berman, DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).

18. I. F. López-Moyado, A. Tsagaratou, H. Yuita, H. Seo, B. Delatte, S. Heinz, C. Benner, A. Rao, Paradoxical association of TET loss of function with genome-wide DNA hypomethylation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 16933–16942 (2019).

19. N. Farhangdoost, C. Horth, B. Hu, E. Bareke, X. Chen, Y. Li, M. Coradin, B. A. Garcia, C. Lu, J. Majewski, Chromatin dysregulation associated with NSD1 mutation in head and neck squamous cell carcinoma. *Cell Rep.* **34**, 108769 (2021).

20. M. R. Corces, J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, C. Groeneveld, C. K. Wong, S. W. Cho, A. T. Satpathy, M. R. Mumbach, K. A. Hoadley, A. G. Robertson, N. C. Sheffield, I. Felau, M. A. A. Castro, B. P. Berman, L. M. Staudt, J. C. Zenklusen, P. W. Laird, C. Curtis, C. G. A. A. Network, W. J. Greenleaf, H. Y. Chang, The chromatin accessibility landscape of primary human cancers. *Science*. **362** (2018),

doi:10.1126/science.aav1898.

21. K. Zhang, J. D. Hocker, M. Miller, X. Hou, J. Chiou, O. B. Poirion, Y. Qiu, Y. E. Li, K. J. Gaulton, A. Wang, S. Preissl, B. Ren, A cell atlas of chromatin accessibility across 25 adult human tissues. *bioRxiv* (2021), doi:10.1101/2021.02.17.431699.

22. Y. Maeda, V. Davé, J. A. Whitsett, Transcriptional control of lung morphogenesis. *Physiol. Rev.* **87**, 219–244 (2007).

23. G. Eraslan, E. Drokhlyansky, S. Anand, A. Subramanian, E. Fiskin, M. Slyper, J. Wang, N. Van Wittenberghe, J. M. Rouhana, J. Waldman, O. Ashenberg, D. Dionne, T. S. Win, M. S. Cuoco, O. Kuksenko, P. A. Branton, J. L. Marshall, A. Greka, G. Getz, A. V. Segrè, F. Aguet, O. Rozenblatt-Rosen, K. G. Ardlie, A. Regev, Single-nucleus cross-tissue molecular reference maps to decipher disease gene function. *bioRxiv* (2021), p. 2021.07.19.452954.

24. D. Aran, M. Sirota, A. J. Butte, Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).

25. J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, K.-Y. Fu, E. Kiss, K. L. Spalding, G. Landesberg, A. Zick, A. Grinshpun, A. M. J. Shapiro, M. Grompe, A. D. Wittenberg, B. Glaser, R. Shemer, T. Kaplan, Y. Dor, Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).

26. F. Mouliere, D. Chandrananda, A. M. Piskorz, E. K. Moore, J. Morris, L. B. Ahlborn, R. Mair, T. Goranova, F. Marass, K. Heider, J. C. M. Wan, A. Supernat, I. Hudecova, I. Gounaris, S. Ros, M. Jimenez-Linan, J. Garcia-Corbacho, K. Patel, O. Østrup, S. Murphy, M. D. Eldridge, D. Gale, G. D. Stewart, J. Burge, W. N. Cooper, M. S. van der Heijden, C. E. Massie, C. Watts, P. Corrie, S. Pacey, K. M. Brindle, R. D. Baird, M. Mau-Sørensen, C. A. Parkinson, C. G. Smith, J. D. Brenton, N. Rosenfeld, Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10** (2018), doi:10.1126/scitranslmed.aat4921.

27. S. Cristiano, A. Leal, J. Phallen, J. Fiksel, V. Adleff, D. C. Bruhm, S. Ø. Jensen, J. E. Medina, C. Hruban, J. R. White, D. N. Palsgrove, N. Niknafs, V. Anagnostou, P. Forde, J. Naidoo, K. Marrone, J. Brahmer, B. D. Woodward, H. Husain, K. L. van Rooijen, M.-B. W. Ørntoft, A. H. Madsen, C. J. H. van de Velde, M. Verheij, A. Cats, C. J. A. Punt, G. R. Vink, N. C. T. van Grieken, M. Koopman, R. J. A. Fijneman, J. S. Johansen, H. J. Nielsen, G. A. Meijer, C. L. Andersen, R. B. Scharpf, V. E. Velculescu, Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*. **570**, 385–389 (2019).

28. P. Jiang, K. Sun, W. Peng, S. H. Cheng, M. Ni, P. C. Yeung, M. M. S. Heung, T. Xie, H. Shang, Z. Zhou, R. W. Y. Chan, J. Wong, V. W. S. Wong, L. C. Poon, T. Y. Leung, W. K. Jacky Lam, J. Y. K. Chan, H. L. Y. Chan, K. C. Allen Chan, R. W. K. Chiu, Y. M. Dennis Lo, Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. *Cancer Discov.* **10**, 664–673 (2020).

29. R. W. Y. Chan, L. Serpas, M. Ni, S. Volpi, L. T. Hiraki, L.-S. Tam, A. Rashidfarrokhi, P. C. H. Wong, L. H. P. Tam, Y. Wang, P. Jiang, A. S. H. Cheng, W. Peng, D. S. C. Han, P. P. P. Tse, P. K. Lau, W.-S. Lee, A. Magnasco, E. Buti, V. Sisirak, N. AlMutairi, K. C. A. Chan, R. W. K. Chiu, B. Reizis, Y. M. D. Lo, Plasma DNA Profile Associated with DNASE1L3 Gene Mutations: Clinical Observations, Relationships to Nuclease Substrate Preference, and In

Vivo Correction. *Am. J. Hum. Genet.* **107**, 882–894 (2020).

30. F. Erger, D. Nörling, D. Borchert, E. Leenen, S. Habbig, M. S. Wiesener, M. P. Bartram, A. Wenzel, C. Becker, M. R. Toliat, P. Nürnberg, B. B. Beck, J. Altmüller, cfNOMe — A single assay for comprehensive epigenetic analyses of cell-free DNA. *Genome Med.* **12**, 1–14 (2020).

31. T. K. Kelly, Y. Liu, F. D. Lay, G. Liang, B. P. Berman, P. A. Jones, Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).

32. A.-L. Doebley, M. Ko, H. Liao, A. Eden Cruikshank, C. Kikawa, K. Santos, J. Hiatt, R. D. Patton, N. De Sarkar, A. C. H. Hoge, K. Chen, Z. T. Weber, M. Adil, J. Reichel, P. Polak, V. A. Adalsteinsson, P. S. Nelson, H. A. Parsons, D. G. Stover, D. MacPherson, G. Ha, Griffin: Framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA, , doi:10.1101/2021.08.31.21262867.

33. C. Caggiano, B. Celona, F. Garton, J. Mefford, B. L. Black, R. Henderson, C. Lomen-Hoerth, A. Dahl, N. Zaitlen, Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. *Nat. Commun.* **12**, 2717 (2021).

34. T. C. Silva, S. G. Coetzee, N. Gull, L. Yao, D. J. Hazelett, H. Noushmehr, D.-C. Lin, B. P. Berman, ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics*. **35**, 1974–1977 (2019).

SUPPLEMENTARY FIGURES AND METHODS

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

Efrat Katsman[1,*], Shari Orlanski[1,*], Filippo Martignano[2,3,*], Amir Eden[4], Iacopo Petrini[5], Silvestro G. Conticello[2,6,#], Benjamin P. Berman[1,#]

1. Department of Developmental Biology and Cancer Research, Hebrew University of Jerusalem, Faculty of Medicine. Jerusalem, Israel

2. Core Research Laboratory, ISPRO, Florence, Italy.

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

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

5. Institute of Clinical Physiology, National Research Council, Pisa, Italy.

* Equal contribution

Jointly supervised project

Correspondence to: s.conticello@ispro.toscana.it, ben.berman@mail.huji.ac.il

Methods

Plasma cfDNA samples, library construction, and sequencing. Samples, library construction and sequencing were described in our initial publication of these sequences (*1*). Notably, one sample (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 (because adapters are shorter) and thus responds differently to the equivalent clean up bead concentration. Furthermore, the multiplex libraries (all except for 19_326) are pooled and an additional bead cleanup step is performed. We also found that adapter clipping performed differently in 19_326 due to the library kit differences. For these reasons, fragmentomic properties are not directly comparable between 19_236 and other samples. We thus analyzed 19_326 separately for all fragmentomic analyses, but included it together with others for methylation and copy number analyses where small differences in fragment length are not expected to make a difference.

Basic processing of nanopore sWGS data. Fastq files were taken from our previous publication (*1*), where they were generated using real-time high-accuracy basecalling during the GridION run. These files were demultiplexed with guppy_barcoder (Version 5.0.11+2b6dbffa5) using "--trim_barcodes --barcode_kits EXP-NBD104" . For singleplex sample (19_326) adapters were trimmed via Porechop (https://github.com/rrwick/Porechop) using: "--discard_middle --extra_end_trim 0". Alignments were performed to GRCh38 with minimap2 (Version 2.13-r850), using the parameters "-ax map-ont --MD", as described in our initial study (*1*).

Filtering of alignments for ichorCNA and fragmentomic analysis. Samtools (Version 1.9) was used to filter BAM alignments, unmapped reads, secondary and supplementary reads, reads with Minimap2 mapping quality less than 20 as in (*2*), and reads longer than 700bp. Bedtobam was used to create bed files, which are available as file "bedsFromBAMsForGEO.zip" in GEO accession GSE185307 and figshare https://doi.org/10.6084/m9.figshare.c.5665966.v1. All genomic coordinates are in GRCh38.

Tumor fraction estimation from somatic copy number analysis (ichorCNA). Somatic copy number analysis was performed using the ichorCNA package v.0.3.2 (*3*). We used default settings to determine copy number alterations and tumor fraction for each cancer sample. If the percentage of genome covered by CN alterations was less than 5%, then the tumor fraction was determined to be unstable and set to 0.

Methylation calling of nanopore sWGS data. 5mC was called using DeepSignal Version: 0.1.8 (*4*), with 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). We used the DeepSignal call_mods (mofidication_call) output tsv file. To aid in combined methylation and fragmentomic analysis, we added additional columns to this file The final 14 fields were extracted from Minimap2 alignment files, matched by read id. They are as follows:

> Column 11: flag, 12: contig, 13: start, 14: end, 15: mapping quality of the chosen alignment, 16: mapping quality of the alignment with the best mapping quality (check),

17: mapping quality of the alignment with the "second" best mapping quality (check), 18: number of alignments for that read, 19: number of alignments for those alignments with mapping quality > 0, 20: left hard clipped bases, 21: left soft clipped bases, 22: read length (from CIGAR, not including soft clipped bases), 23: right soft clipped bases, 24: right hard clipped bases

These files are available as file “raw-modifcation-calls-fixed-format.zip” in GEO accession GSE185307 and figshare https://doi.org/10.6084/m9.figshare.c.5665966.v1. All genomic coordinates are in GRCh38.

From the modified DeepSignal modification_call output described above (in “raw-modifcation-calls-fixed-format.zip”), we then extracted the methylation calls for each (strand-specific) 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 0-1 for every CpG covered. These are available as file “grouped-beta-value_bedgraph.zip” in GEO accession GSE185307 and figshare https://doi.org/10.6084/m9.figshare.c.5665966.v1. All genomic coordinates are in GRCh38.

DNA methylation in 10 Mbp bins. To generate Figure 1A-B and Supplemental Figure 1 plots, segmentation results from our previous CNV analysis (*1*) were converted from GRCh37 to GRCh38 using NCBI remap API and divided into non-overlapping 10Mb bins. These are available as file “SegmentationResultsMartignano2021.zip” in GEO accession GSE185307 and figshare https://doi.org/10.6084/m9.figshare.c.5665966.v1. Copy number status of each bin was determined by log2ratio segment mean > 0.10 and < -0.10 for Gain and Loss respectively. For healthy samples, 10Mb bins were generated from the whole genome. Mean methylation levels for each bin were calculated as sum(frac_methylation_each_cytosine)/cytosine_count. For the analysis of the methylation at Partially Methylated Domains (PMDs), we used only “Common PMDs” from (*5*), splitting regions within PMDs into non-overlapping 10 Mbp bins. For PMDs, we calculated average methylation using only “solo-WCGW” CpGs from the same paper (*5*). Common PMD and solo-WCGW annotations were taken from file https://zwdzwd.s3.amazonaws.com/pmd/solo_WCGW_inCommonPMDs_hg38.bed.gz.

NKX2-1 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 6,754 ATAC-seq peaks identified in the pneumocyte (PAL) cluster 13 CREs from (*6*) (downloaded from supplemental table 6 of that paper “Table_S6_Union_set_of_cCREs.xlsx”). We then selected 5,974 peaks that overlapped a predicted NKX2-1 TFBS, and centered each on the predicted NXK2-1 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 GEO accession GSE185307 and figshare https://doi.org/10.6084/m9.figshare.c.5665966.v1.

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 NKX2-1 methylation levels in TCGA lung and non-lung samples, we downloaded TCGA WGBS bedgraph files from https://zwdzwd.github.io/pmd (*5*). 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 TGCA 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). Plasma WGBS of HCC was downloaded from EGAD00001004317. Plasma WGBS of CRC patients was downloaded from EGAD00001004568. Plasma WGBS of healthy controls was downloaded from EGAD00001001602.

450k healthy tissue reference atlas. 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)(*7*). The script feature_selection.m was used to select Lung_cell specific CpGs for 2,000 hypermethylated and 2,000 hypomethylated probes. 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. We plotted the values using the a custom R script available at (https://github.com/methylgrammarlab/cfdna-ont/deconvolution_code/cell_types_probes/plot_tumor_fractions_vs_score.R).

450k TCGA tumor reference atlas. We downloaded the Infinium 450k beta value files for TCGA Lung Adenocarcinoma (LUAD) tumors using the ELMER packaged in Bioconductor (*8*). 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. We then performed a t-test to compare the methylation beta values of these Lung_specific probes to the four plasma cfDNA samples from the MethAtlas paper (*7*), requiring a Benjamini-Hochberg corrected FDR of <0.01 and an absolute beta value difference of 0.3 or greater. We plotted the values using the a custom R script available at (https://github.com/methylgrammarlab/cfdna-ont/deconvolution_code/TCGA_probes/plot_tumor_fractions_vs_score.R file).

450k cell type deconvolution. First, CpGs covering either the forward or reverse strand of each CpG on the Infinium 450k array were extracted from each Nanopore beta value file (averaged by taking the total number of methylated reads on either strand divided by the total number of methylated+unmethylated reads on both strands) to produce a beta value vector Y. Each of these files was intersected with the normal Lung_cell specific probes as described above (an example of this is shown in Fig 1A). For each probe, the 450k beta values were averaged to produce a single Lung-specific beta value $\underline{X_1}$. The same was done for the four plasma cfDNA samples from (*7*) to yield a healthy cfDNA beta value $\underline{X_2}$. We used the Lawson-Hanson algorithm for non-negative least squares (NNLS) (https://cran.r-project.org/web/packages/nnls) to perform non-negative least squares regression as in (*7*). Specifically, we identified non-negative coefficients $\beta_1$and β_2, representing the fraction of Lung cells and normal blood cells in

the Nanopore cfDNA mixture, respectively, subject to the constraints $argmin_{\beta}\left||X\beta - Y|\right|_2$ and $\beta \geq 0$. Then a single Lung fraction β was determined by having $\beta_1$and β_2 sum to 1. $\beta = \frac{\beta_1}{(\beta_1+\beta_2)}$.

Correcting TCGA methylation model for cancer cell purity. For the deconvolution based on TCGA LUAD tumors, we had to account for the fact that most TCGA LUAD samples are a mixture of cancer cells and leukocytes, with a median cancer fraction of ~50%. 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 (*7*), 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 (*9*), 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.

Fragment length analysis. Minimap2 alignments were filtered as described above. Reads with soft clipping at either the 5' or 3' ends were removed. Fragment length was calculated from Minimap2 BAM CIGAR column by summing all counts. Short fragment ratio was calculated as $\frac{numfrags_{100-150bp}}{numfrags_{100-220bp}}$ (150bp is the same cutoff for short fragments used in (*10*)).

End motif analysis. Minimap2 alignments were filtered as described above. To avoid biases that would affect 5' end motif analysis, we also removed reads with any 5' soft clipping. The first 4 bases of each fragment were extracted and used for 4-mer analysis. Motif frequency was calculated as $\frac{numfrags_{4mer}}{numfrags_{total}}$.

CTCF nucleosome positioning analysis. We used 9,780 evolutionarily conserved CTCF motifs occurring in distal ChIP-seq peaks, which were taken from (*11*). Nanopore or Illumina fragments within the size range of 130-155bp were used for fragment coverage analysis. These shorter mononucleosomal fragments give better nucleosome-level resolution than longer 167 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.

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

683 ## Supplemental tables and data files

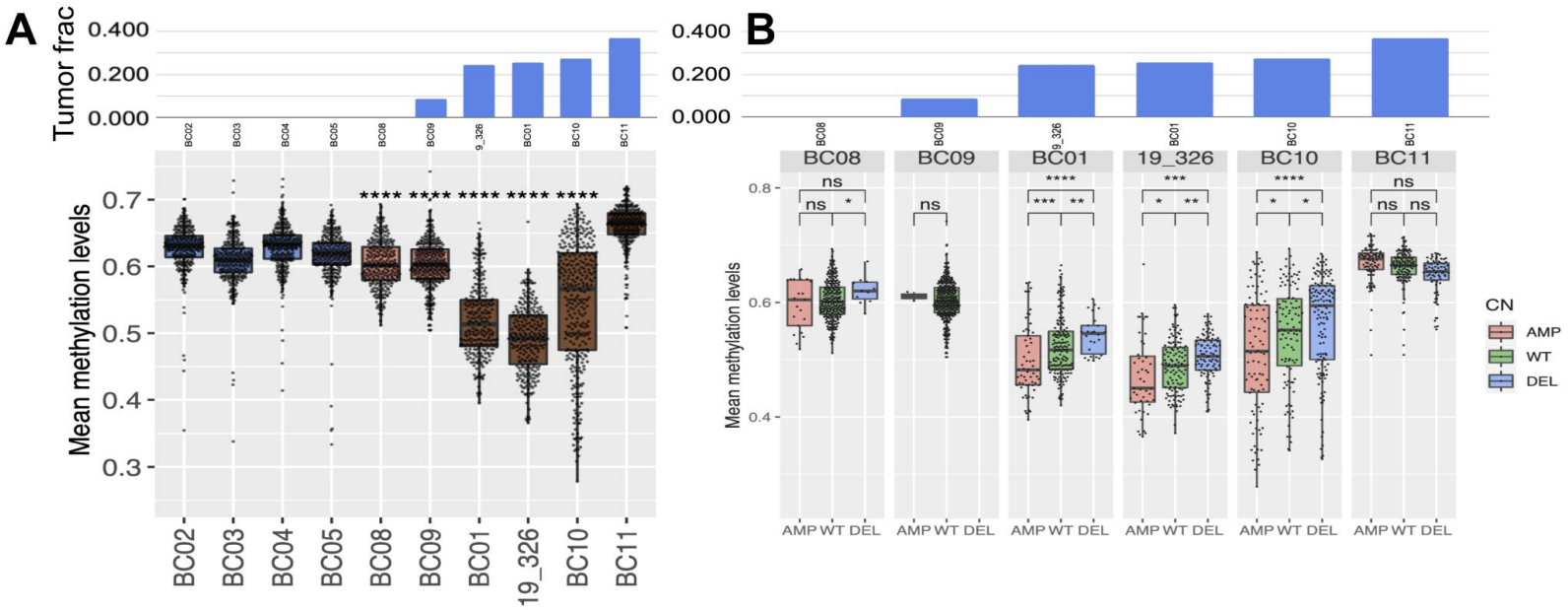
684 **Supplemental Table 1. Samples and sequencing statistics.**

685

Supplemental figures

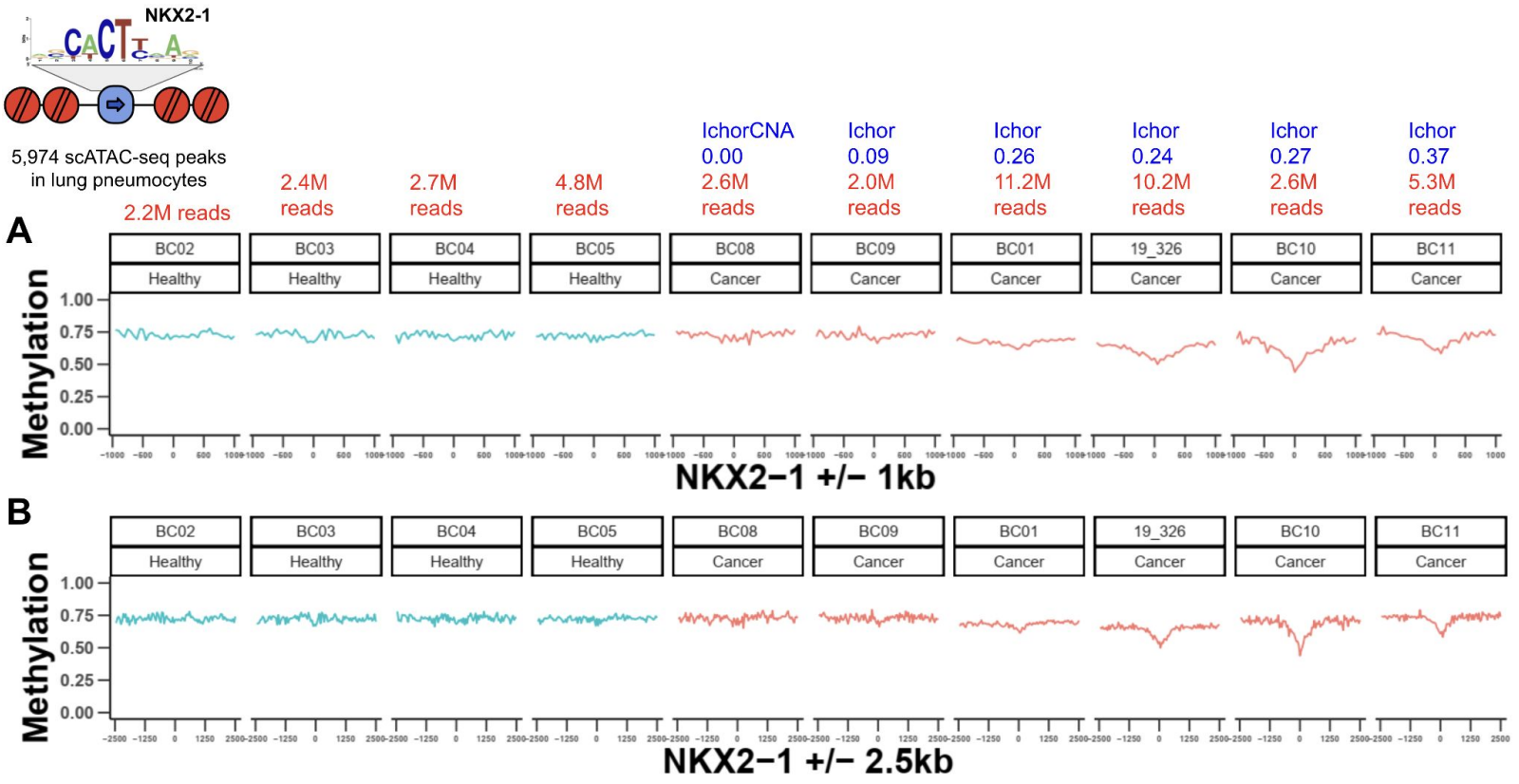
Supplemental Figure 1. Hypomethylation within PMDs. (A) Average DNA methylation from all 10 Mbp bins contained in common Partially Methylated Domains (PMDs) from an earlier study (*5*). Only the 1,669,234 CpGs defined as "solo-WCGW" CpGs from (*5*) were used. All 10 Mbp bins are shown for Oxford Nanopore (ONT) cfDNA WGS of 4 healthy controls and 6 lung adenocarcinoma cases. Tumor fraction (panel A, top) was calculated from somatic copy number alterations using ichorCNA (*3*). (B) 10 Mbp bins stratified by copy number status for all cancer cases. Statistical significance for A and B determined by one-tailed Wilcoxon test. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

Supplemental Figure 1: Hypomethylation within PMDs



Supplemental Figure 2. Methylation at Lung-specific NKX2-1 binding sites. Illustration of predicted NKX2-1 binding sites within single-cell ATAC-seq peaks specific to primary lung pneumocytes. (A) Relative methylation level of healthy controls and lung cancer samples from ONT cfDNA WGS within -1kb to +1kb of pneumocyte-specific NKX2-1 sites. Cancer cases are ordered by tumor fraction estimates from ichorCNA. (B) Same analysis, but within -2.5kb to +2.5kb of pneumocyte-specific NKX2-1 sites.

Supplemental Figure 2: Methylation at Lung-specific NKX2-1 binding sites

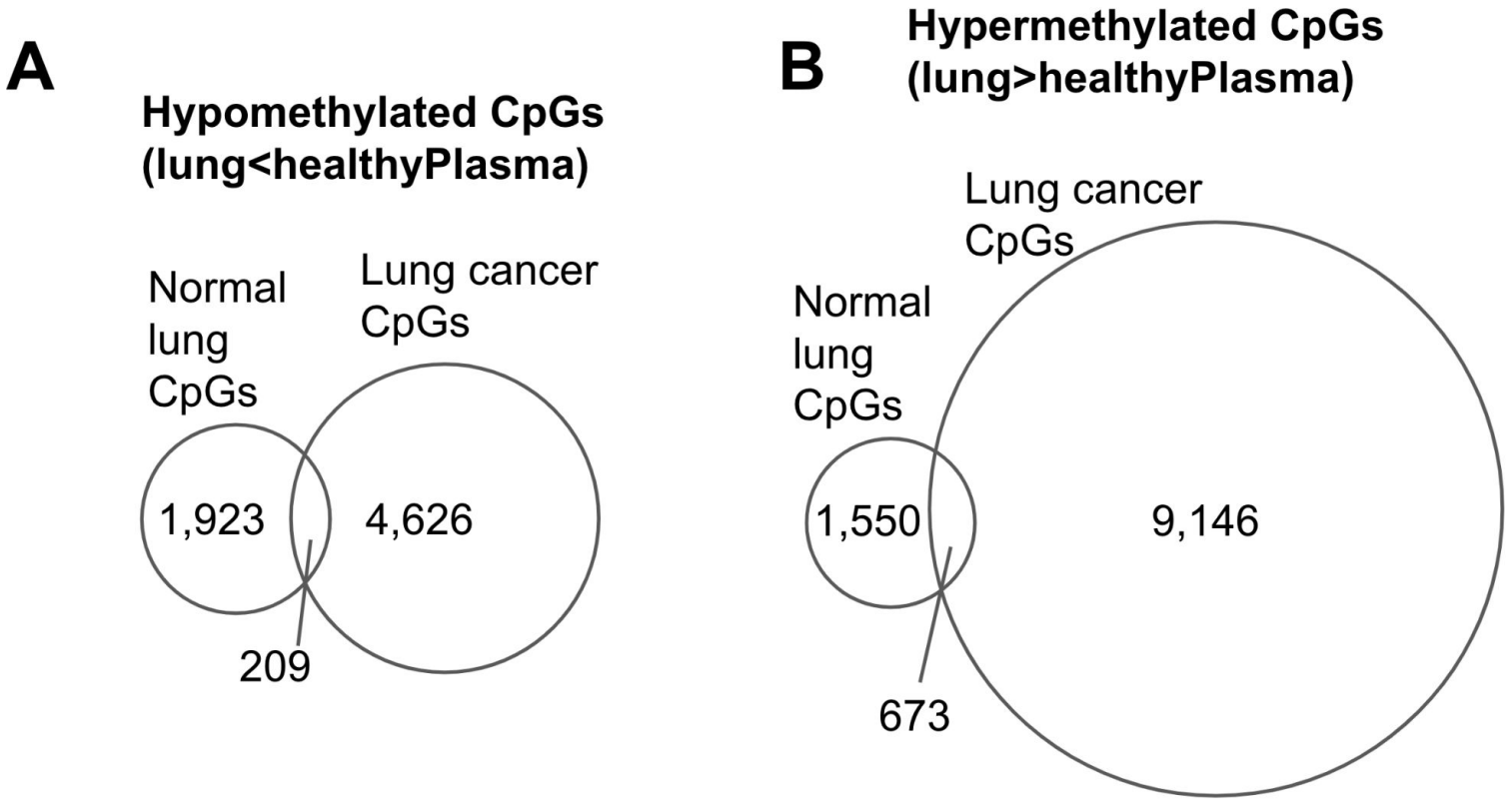


705 **Supplemental Figure 3. Deconvolution of ctDNA using HumanMethylation450k reference**
706 **data.** Estimated lung fraction β plotted for all Nanopore plasma samples. (A) Number of probes
707 hypomethylated in Normal lung cells relative (red) or Lung cancer cells (blue) relative to healthy
708 plasma cfDNA methylation, using the feature_selection.m script of the MethAtlas package (*7*).
709 (B) Number of hypermethylated probes.

710

711

Supplemental Figure 3: Deconvolution of cfDNA using HumanMethylation450k reference data

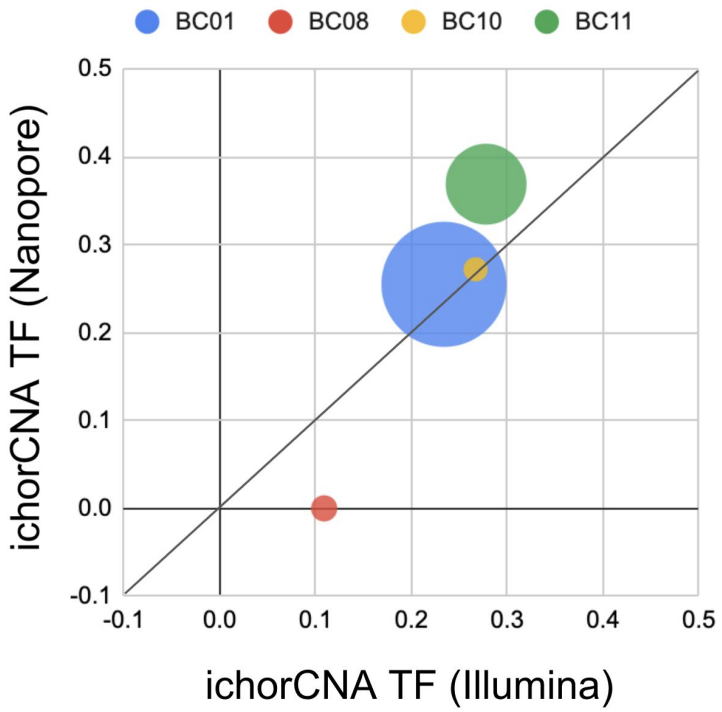


712 **Supplemental Figure 4. ichorCNA tumor fraction estimates from deep Illumina WGS.** (B)
713 ichorCNA (*12*) tumor fraction estimates for samples with matched Nanopore and Illumina WGS
714 data. Size of each bubble is the number of uniquely aligned reads in the Nanopore library. The
715 median number of Nanopore reads per sample is 6.4M and the median number of Illumina
716 reads per sample is 17.0M.

717

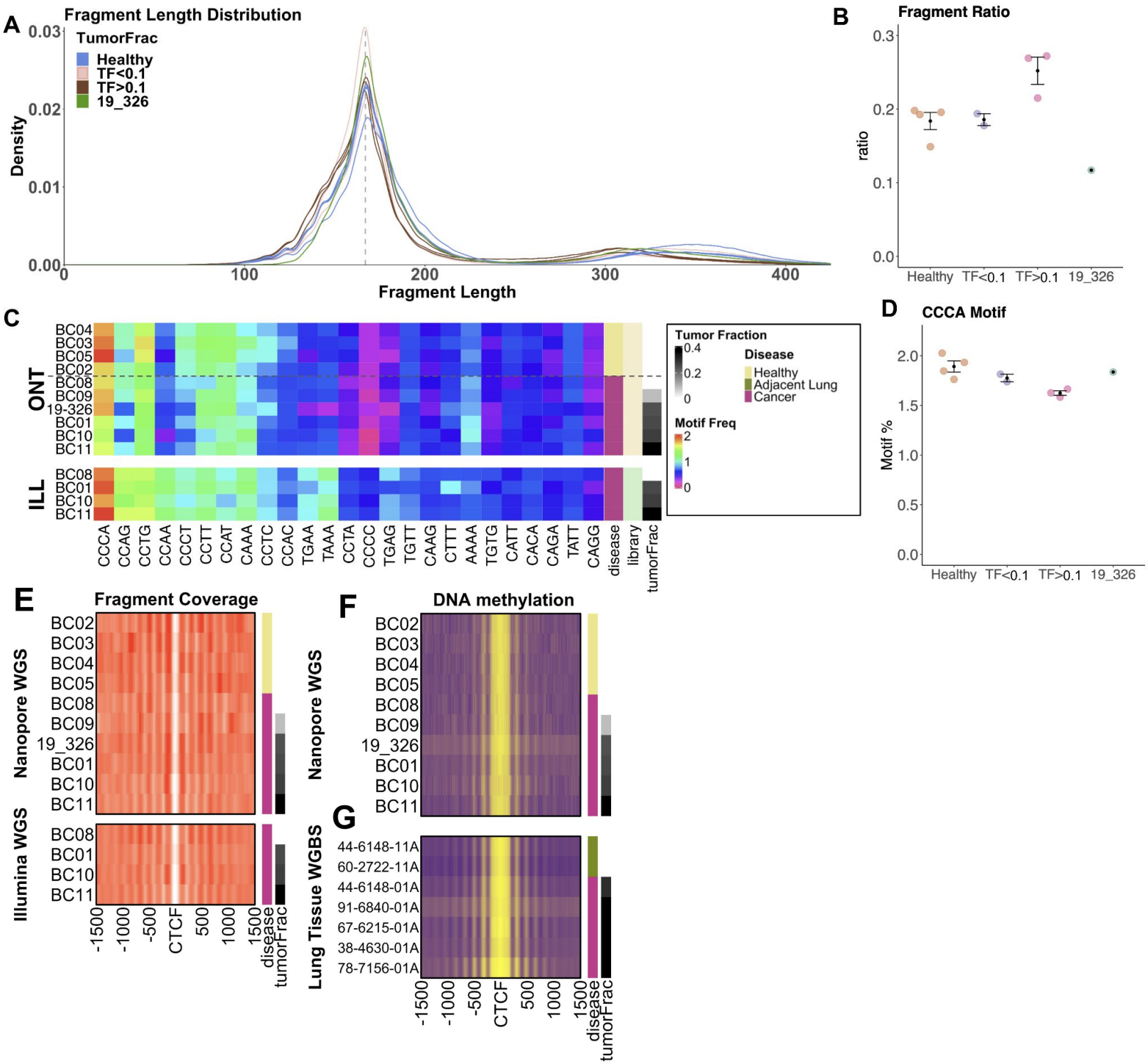
718

Supplemental Figure 4: Normal lung methylation estimates and deep sequencing ichorCNA tumor fraction estimates



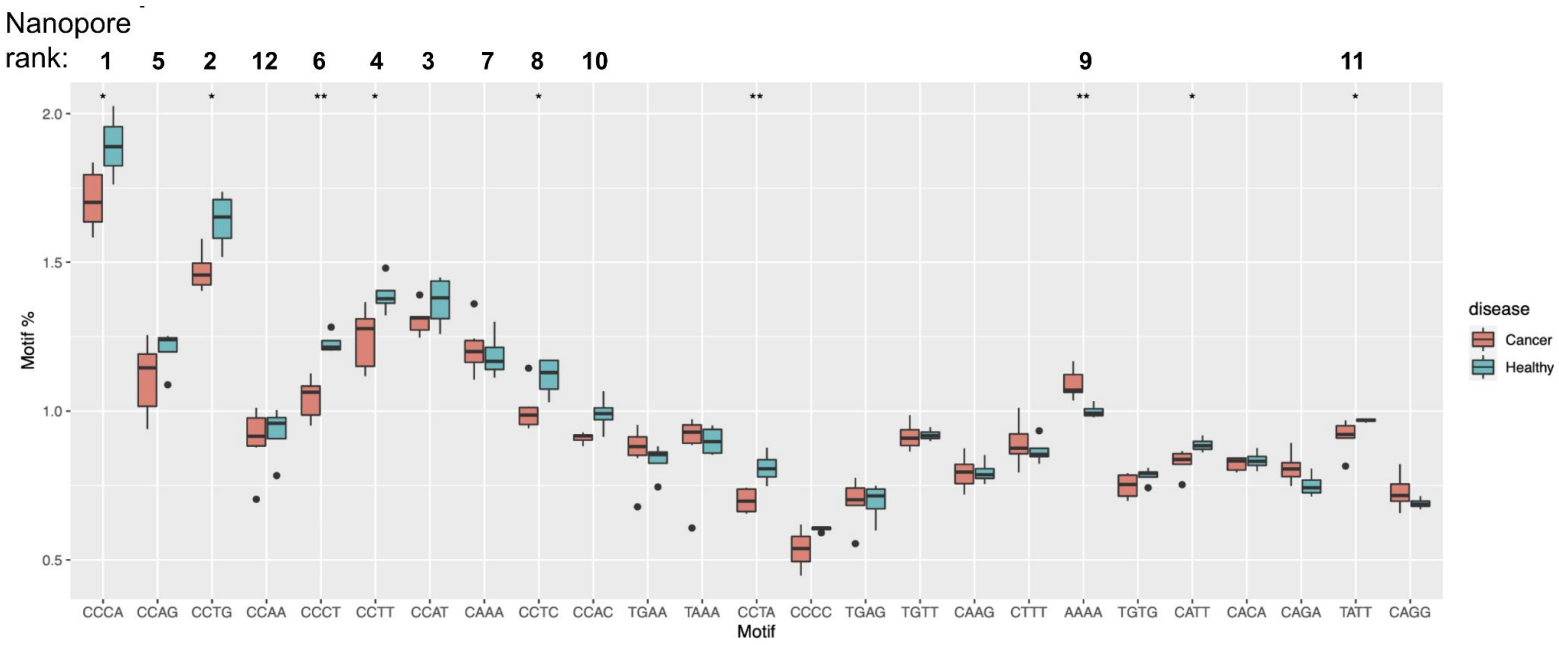
719 **Supplemental Figure 5. Fragment analysis of sample 19_326.** Sample 19_326 used different bead
720 selection parameters and a different library version, and thus is not included in the main analysis. Here,
721 we show 19_326 in context in all fragmentomic analyses. (A-D) General fragment properties. (A)
722 Fragment length density of all samples, colored by tumor fraction estimated by ichorCNA. (B) Fragment
723 length ratio calculated as the number of short fragments (100-150bp, same cutoff used in (*10*)) divided by
724 the total number of total mononucleosome reads (100-220bp). (C) Frequency of different 4-mer
725 sequences at 5' fragment ends, comparing ONT cfDNA WGS samples and matched Illumina samples.
726 The 25 most frequent 4-mers in healthy plasma from (*13*) are shown in rank order. Samples are ordered
727 by healthy vs. cancer and then by increasing tumor fraction. (D) Frequency of CCCA motif at 5' fragment
728 ends. (E-G) Alignments to CTCF motifs within distal ChIP-seq peaks from (*11*). (E, top) cfDNA fragment
729 coverage shown as fold-change vs. average coverage depth across the genome. Includes only fragments
730 of length 130-155bp to maximize resolution. (E, bottom) Matched Illumina samples of higher sequencing
731 depth (median 17.0M fragments in Illumina vs. 6.4M in ONT samples) show similar patterns. (F) CTCF
732 DNA methylation of Nanopore samples from this study. (G) DNA methylation from seven lung tissue
733 WGBS samples from TCGA (*5*).

Supplemental Figure 5: General fragment properties



734 **Supplemental Figure 6. End motif comparison of cancer and healthy samples.** (A) 4-mer
735 frequencies were grouped into Healthy samples (blue) and cancer samples (red), and ordered
736 as above. Rank within Nanopore healthy samples is shown for the top 12 4-mers (top).
737 Statistical significance between healthy and cancer groups in panel E was calculated using two-
738 tailed Student's t-test. *p<0.05, **p<0.01

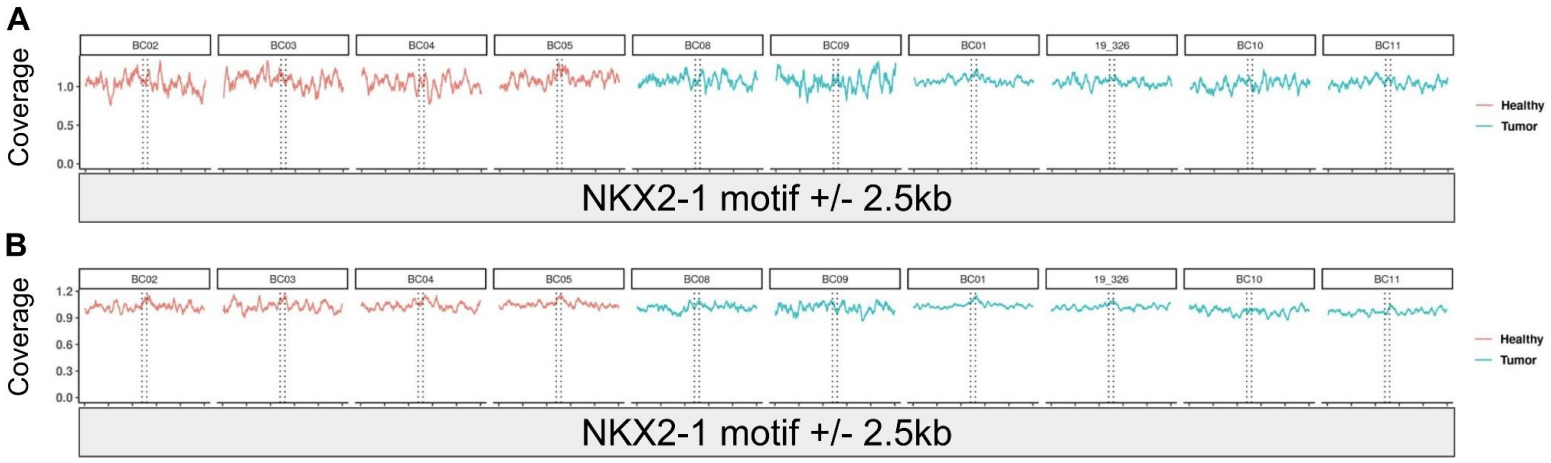
Supplemental Figure 6: End motif comparison of cancer and healthy samples



Top 25 ranked 4-mers in healthy plasma (Chan et al. 2020)

Supplemental Figure 7. Fragment analysis of Lung-specific NKX2-1 binding sites. (A) Analysis of 5,974 lung-specific regions aligned to predicted NKX2-1 binding sites contained within single-cell ATAC-seq peaks specific to primary lung pneumocytes (same sites as Figure 1H-K). Plot shows fragment coverage relative to average coverage across the genome, including only fragments of length 130-155bp. (B) Same plot, but using 28,298 predicted NKX2-1 binding sites from ATAC-seq peaks of TCGA Lung adenocarcinoma samples instead of ATAC-seq peaks from normal lung pneumocytes (same sites from Supplemental Figure 3).

Supplemental Figure 7: Fragment analysis of Lung-specific NKX2-1 binding sites



References

1. F. Martignano, U. Munagala, S. Crucitta, A. Mingrino, R. Semeraro, M. Del Re, I. Petrini, A. Magi, S. G. Conticello, Nanopore sequencing from liquid biopsy: analysis of copy number variations from cell-free DNA of lung cancer patients. *Mol. Cancer*. **20**, 32 (2021).

2. T. Baslan, S. Kovaka, F. J. Sedlazeck, Y. Zhang, R. Wappel, S. W. Lowe, S. Goodwin, M. C. Schatz, High resolution copy number inference in cancer using short-molecule nanopore sequencing. *bioRxiv* (2020), p. 2020.12.28.424602.

3. V. A. Adalsteinsson, G. Ha, S. S. Freeman, A. D. Choudhury, D. G. Stover, H. A. Parsons, G. Gydush, S. C. Reed, D. Rotem, J. Rhoades, D. Loginov, D. Livitz, D. Rosebrock, I. Leshchiner, J. Kim, C. Stewart, M. Rosenberg, J. M. Francis, C.-Z. Zhang, O. Cohen, C. Oh, H. Ding, P. Polak, M. Lloyd, S. Mahmud, K. Helvie, M. S. Merrill, R. A. Santiago, E. P. O'Connor, S. H. Jeong, R. Leeson, R. M. Barry, J. F. Kramkowski, Z. Zhang, L. Polacek, J. G. Lohr, M. Schleicher, E. Lipscomb, A. Saltzman, N. M. Oliver, L. Marini, A. G. Waks, L. C. Harshman, S. M. Tolaney, E. M. Van Allen, E. P. Winer, N. U. Lin, M. Nakabayashi, M.-E. Taplin, C. M. Johannessen, L. A. Garraway, T. R. Golub, J. S. Boehm, N. Wagle, G. Getz, J. C. Love, M. Meyerson, Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).

4. P. Ni, N. Huang, Z. Zhang, D.-P. Wang, F. Liang, Y. Miao, C.-L. Xiao, F. Luo, J. Wang, DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*. **35**, 4586–4595 (2019).

5. W. Zhou, H. Q. Dinh, Z. Ramjan, D. J. Weisenberger, C. M. Nicolet, H. Shen, P. W. Laird, B. P. Berman, DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).

6. K. Zhang, J. D. Hocker, M. Miller, X. Hou, J. Chiou, O. B. Poirion, Y. Qiu, Y. E. Li, K. J. Gaulton, A. Wang, S. Preissl, B. Ren, A cell atlas of chromatin accessibility across 25 adult human tissues. *bioRxiv* (2021), doi:10.1101/2021.02.17.431699.

7. J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, K.-Y. Fu, E. Kiss, K. L. Spalding, G. Landesberg, A. Zick, A. Grinshpun, A. M. J. Shapiro, M. Grompe, A. D. Wittenberg, B. Glaser, R. Shemer, T. Kaplan, Y. Dor, Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).

8. T. C. Silva, S. G. Coetzee, N. Gull, L. Yao, D. J. Hazelett, H. Noushmehr, D.-C. Lin, B. P. Berman, ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics*. **35**, 1974–1977 (2019).

9. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell*. **33**, 676–689.e3 (2018).

10. F. Mouliere, D. Chandrananda, A. M. Piskorz, E. K. Moore, J. Morris, L. B. Ahlborn, R. Mair, T. Goranova, F. Marass, K. Heider, J. C. M. Wan, A. Supernat, I. Hudecova, I. Gounaris, S. Ros, M. Jimenez-Linan, J. Garcia-Corbacho, K. Patel, O. Østrup, S. Murphy, M. D. Eldridge, D. Gale, G. D. Stewart, J. Burge, W. N. Cooper, M. S. van der Heijden, C. E. Massie, C. Watts, P. Corrie, S. Pacey, K. M. Brindle, R. D. Baird, M. Mau-Sørensen, C. A. Parkinson, C. G. Smith, J. D. Brenton, N. Rosenfeld, Enhanced detection of circulating

789 tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10** (2018),
790 doi:10.1126/scitranslmed.aat4921.

791 11. T. K. Kelly, Y. Liu, F. D. Lay, G. Liang, B. P. Berman, P. A. Jones, Genome-wide mapping
792 of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome*
793 *Res.* **22**, 2497–2506 (2012).

794 12. A. D. Choudhury, L. Werner, E. Francini, X. X. Wei, G. Ha, S. S. Freeman, J. Rhoades, S.
795 C. Reed, G. Gydush, D. Rotem, C. Lo, M.-E. Taplin, L. C. Harshman, Z. Zhang, E. P.
796 O'Connor, D. G. Stover, H. A. Parsons, G. Getz, M. Meyerson, J. C. Love, W. C. Hahn, V.
797 A. Adalsteinsson, Tumor fraction in cell-free DNA as a biomarker in prostate cancer. *JCI*
798 *Insight*. **3** (2018), doi:10.1172/jci.insight.122109.

799 13. Plasma DNA Profile Associated with DNASE1L3 Gene Mutations: Clinical Observations,
800 Relationships to Nuclease Substrate Preference, and In Vivo Correction. *Am. J. Hum.*
801 *Genet.* **107**, 882–894 (2020).

802