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.

63 ## Results

64 All analyses described below are from sequence data generated in our previous publication,
65 which used cfDNA extracted from 4mL of plasma for four healthy control individuals and six
66 metastatic lung adenocarcinoma cases (*11*). Six of the samples had between 3.8M and 5.3M
67 raw reads (2.2M-2.6M uniquely mapping reads), and the remaining four samples had between
68 8.4M and 20.2M raw reads (4.8M-11.2M uniquely mapping reads). All sequencing statistics are
69 available in Supplemental Table 1. We used ichorCNA (*13*) to estimate the tumor fraction of
70 each sample using somatic copy number alterations (SCNAs) (Supplemental Table 1). Four of
71 the six cancer cases had tumor fraction estimates greater than 0.1 (high tumor fraction), one
72 case had 0.086 (BC09), and another (BC08) was under the detection limit for TF estimation via
73 IchorCNA so tumor fraction was set to 0 (Figure 1A-1B, top).

74

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

76 Global DNA hypomethylation is one of the hallmarks of the cancer epigenome and has been
77 proposed as a general ctDNA biomarker (*14*), and was recently verified in WGBS of cfDNA from
78 NSCLC cases (*15*). In order to investigate this, we processed the original fast5 sequencing files
79 with DeepSignal (*16*) to call methylation at individual CpGs. The six lower coverage samples
80 covered between 4.3M and 5.5M CpGs (usually by a single read per CpG), while the remaining
81 four samples covered 8.1M-18.9M CpGs (Supplemental Table 1). Next, we calculated global
82 methylation within 10 Mbp genomic windows genome-wide. This analysis showed high
83 methylation levels for the four healthy control plasmas and the two low tumor fraction cases,
84 and significantly reduced methylation for three of the four high tumor fraction samples (Figure
85 1A, bottom). Reasoning that regions of copy number alteration would have skewed proportions
86 of tumor-derived DNA and thus skewed methylation levels, we split out methylation by SCNA
87 status for all cancer samples. In the three cases with globally reduced methylation (BC01,
88 19_326, BC10), amplified regions were significantly more hypomethylated than diploid regions,
89 as expected (Figure 1B). While hypomethylation could not be detected genome-wide in the low
90 tumor fraction sample BC08, amplified regions were significantly hypomethylated. Conversely,
91 deleted regions showed reduced hypomethylation relative to diploid regions, but this trend only
92 reached statistical significance in two of the three cases with global hypomethylation (19_326
93 and BC10). In the final case (BC11), DNA methylation overall was higher than in healthy
94 plasma, and SCNAs levels suggested this was due specifically to the high methylation of
95 cancer-derived DNA (Figure 1B). While this is an interesting case, it is not surprising given the
96 high degree of variability associated with global hypomethylation (*17*), a process that is not
97 entirely understood but is known to be affected by various chromatin modifiers that are
98 dysregulated in cancer(*18*, *19*).

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

105 association between SCNAs and hypomethylation that occurred genome-wide analysis (Figure
106 1B) was also significant in the PMD-only analysis (Supplemental Figure 1B).

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

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

134 The quantitative nature of DNA methylation allows accurate estimates of cell type mixtures from
135 reference datasets of pure cell types (*24*), including applications to cfDNA (*25*). While there is
136 currently no whole-genome methylation dataset that includes pure lung epithelial cells, such a
137 dataset was recently generated on the Illumina HumanMethylation450k (HM450k) platform (*25*).
138 We adapted the non-negative least squares (NNLS) regression method used in (*25*) to
139 deconvolute our Nanopore plasma samples into lung cell and healthy plasma cell type
140 components (Figure 1H). We had to significantly expand the number of cell-type specific marker
141 CpGs used in the MethAtlas paper (*25*), due to the relatively low degree of overlap between
142 HM450k probes and CpGs called in our Nanopore samples (the majority of Nanopore samples
143 overlapped less than 20% of HM450k probes, see Supplemental Table 1). We identified a total
144 of 4,355 lung-specific marker CpGs (Supplemental Figure 3), which covered a median of 818
145 CpGs per sample (Supplemental Table 1). For example, healthy sample BC05 overlapped
146 1,760 lung-specific CpGs, while cancer sample BC11 overlapped 1,251 (Figure 1H,
147 Supplemental Figure 4, Supplemental Table 1). These were used as input to NNLS regression

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

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

163 To verify the robustness of the NNLS deconvolution results, we performed the same analysis
164 using a mutually exclusive set of 14,654 HM450k marker probes differentially methylated
165 between TCGA LUAD tumors and healthy plasma (Figure 1L). This analysis yielded very similar
166 results to the normal lung-based analysis (Figure 1M-N), reinforcing the idea that circulating
167 tumor DNA can be detected using reference data from either the appropriate normal cell type or
168 from tumors (*3*). This analysis revealed an interesting outlier, BC10, which had a methylation-
169 based lung cell estimate that was almost 2-fold higher than the ichorCNA estimate in both the
170 normal lung based (Figure 1K) and tumor based (Figure 1N) estimates. Interestingly, the NKX2-
171 1 methylation analysis agreed with these deconvolution results, with BC10 showing the highest
172 degree of NKX2-1 demethylation of any cancer sample (Supplemental Figure 2). While it would
173 require study in a larger cohort, it is possible that this represents a case where low read
174 coverage (2.6M reads) leads SCNA analysis to fail at detecting a whole-genome doubling event,
175 and that the model could be improved by incorporating DNA methylation data.

176

177 *Nanopore preserves fragmentomic features of ctDNA*

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

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

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

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

211 Cell-free DNA circulates primarily as mono-nucleosomal fragments, and nucleosome positions
212 inferred from fragment cut sites can be used to detect cell-of-origin (reviewed in (*1*)). Bisulfite
213 conversion used with Illumina-based sequencing can degrade these fragmentation patterns
214 (*30*). We have previously showed that mono-nucleosomal fragment lengths are largely
215 preserved in Nanopore cfDNA sWGS (*11*), but we wanted to further investigate the ability of
216 Nanopore to reveal biologically relevant nucleosome structure based on fragmentation patterns.
217 CTCF binding sites present the best model for nucleosome organization - they position 10
218 phased nucleosomes on either side of a central binding site, which itself lies within a 100bp
219 nucleosome depleted region (*31*). The Nanopore fragmentation pattern around CTCF binding
220 sites recapitulated this structure (Figure 2E, top), and reproduced the pattern based on our
221 deeper Illumina WGS sequencing (Figure 2E, bottom). The CTCF binding site is also known to
222 sit at the center of a 400bp demethylated region, and this was also recapitulated from 5mC
223 levels from our Nanopore sequencing (Figure 2F, top). We verified this pattern in lung
224 adenocarcinoma tumors using bisulfite-seq (WGBS) data from TCGA (*17*) (Figure 2G). CTCF
225 binding is largely not cell-type specific, and thus we did not observe any differences between
226 healthy and cancer samples in this analysis. We tried the same fragment coverage analysis for
227 the lung-specific NKX2-1 binding sites discussed earlier, but we were not able to detect any
228 nucleosome structure (Supplemental Figure 6). It is possible that higher read depth will be
229 necessary to detect nucleosomal fragmentation patterns from cell-type specific components of
230 cfDNA, but new computational methods such as Griffin (*32*) may allow for more sensitive
231 detection in the future.

232

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

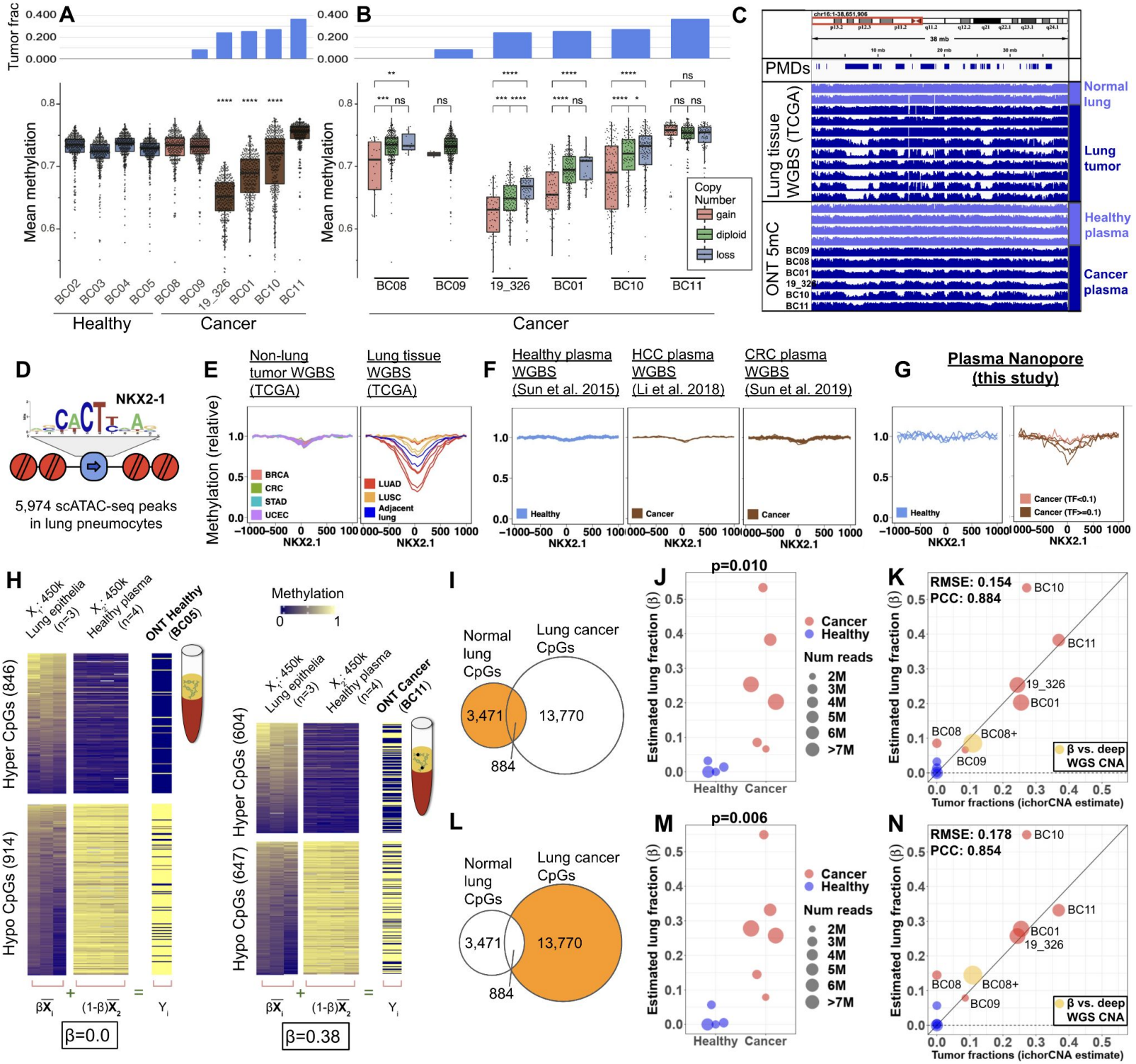
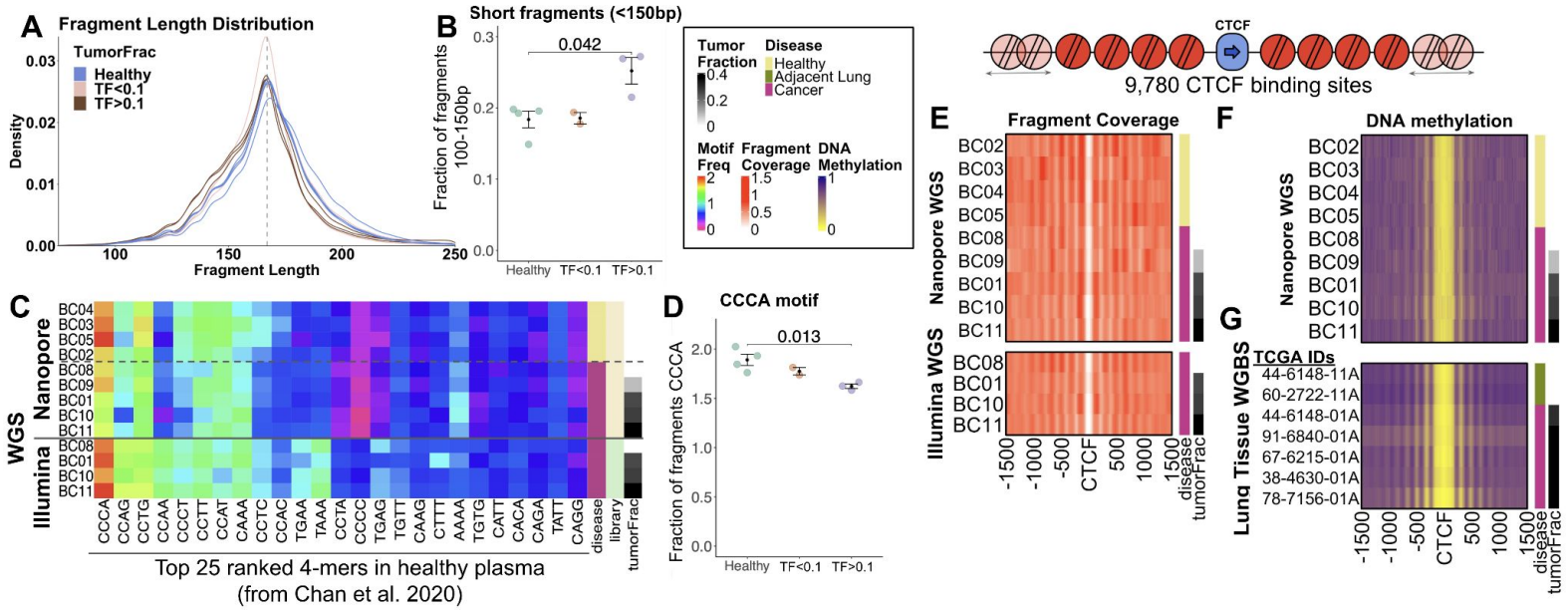


Figure 2: Nanopore preserves fragmentomic features of ctDNA. (A) Fragment length density of all samples, colored by tumor fraction estimated by ichorCNA. (B) Fragment length ratio calculated as the number of short fragments (100-150bp, same cutoff used in (*26*)) divided by the total number of total 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



355 ## References

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

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

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

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

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

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

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

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

386 9. L. Djirackor, S. Halldorsson, P. Niehusmann, H. Leske, L. P. Kuschel, J. Pahnke, B. J. Due-
387 Tønnessen, I. A. Langmoen, C. J. Sandberg, P. Euskirchen, E. O. Vik-Mo, EPCT-15.
388 RAPID EPIGENOMIC CLASSIFICATION OF BRAIN TUMORS ENABLES
389 INTRAOPERATIVE NEUROSURGICAL RISK MODULATION. *Neuro. Oncol.* **23**, i50–i50
390 (2021).

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

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

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

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

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

410 14. K. C. A. Chan, P. Jiang, C. W. M. Chan, K. Sun, J. Wong, E. P. Hui, S. L. Chan, W. C.
411 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,
412 P. B. S. Lai, H. Sun, R. W. K. Chiu, Y. M. D. Lo, Noninvasive detection of cancer-
413 associated genome-wide hypomethylation and copy number aberrations by plasma DNA
414 bisulfite sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18761–18768 (2013).

415 15. H.-N. Nguyen, N.-P. T. Cao, T.-C. Van Nguyen, K. N. D. Le, D. T. Nguyen, Q.-T. T.
416 Nguyen, T.-H. T. Nguyen, C. Van Nguyen, H. T. Le, M.-L. T. Nguyen, T. V. Nguyen, V. U.
417 Tran, B. A. Luong, L. G. H. Le, Q. C. Ho, H.-A. T. Pham, B. T. Vo, L. T. Nguyen, A.-T. H.
418 Dang, S. D. Nguyen, D. M. Do, T.-T. T. Do, A. V. Hoang, K. T. Dinh, M.-D. Phan, H. Giang,
419 L. S. Tran, Liquid biopsy uncovers distinct patterns of DNA methylation and copy number
420 changes in NSCLC patients with different EGFR-TKI resistant mutations. *Sci. Rep.* **11**, 1–
421 12 (2021).

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

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

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

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

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

439 doi:10.1126/science.aav1898.

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

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

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

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

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

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

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

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

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

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

483 30. F. Erger, D. Nörling, D. Borchert, E. Leenen, S. Habbig, M. S. Wiesener, M. P. Bartram, A.
484 Wenzel, C. Becker, M. R. Toliat, P. Nürnberg, B. B. Beck, J. Altmüller, cfNOMe — A single
485 assay for comprehensive epigenetic analyses of cell-free DNA. *Genome Med.* **12**, 1–14
486 (2020).

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

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

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

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

501

502 **SUPPLEMENTARY FIGURES AND METHODS**

503

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

506

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

509

510 1. Department of Developmental Biology and Cancer Research, Hebrew University of

511 Jerusalem, Faculty of Medicine. Jerusalem, Israel

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

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

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

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

518

519 * Equal contribution

520 # Jointly supervised project

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

522

523

524

525

526

527

528 ## Methods

529 *Plasma cfDNA samples, library construction, and sequencing.* Samples, library construction and
530 sequencing were described in our initial publication of these sequences (*1*). Notably, one
531 sample (19_326) was produced using a different library kit (SQK-LSK109 vs. NBD-
532 EXP104+SQK-LSK109 for all other samples). This is the singleplex library kit, which results in
533 shorter adapter-ligated templates overall (because adapters are shorter) and thus responds
534 differently to the equivalent clean up bead concentration. Furthermore, the multiplex libraries (all
535 except for 19_326) are pooled and an additional bead cleanup step is performed. We also found
536 that adapter clipping performed differently in 19_326 due to the library kit differences. For these
537 reasons, fragmentomic properties are not directly comparable between 19_236 and other
538 samples. We thus analyzed 19_326 separately for all fragmentomic analyses, but included it
539 together with others for methylation and copy number analyses where small differences in
540 fragment length are not expected to make a difference.

541 *Basic processing of nanopore sWGS data*. Fastq files were taken from our previous publication
542 (*1*), where they were generated using real-time high-accuracy basecalling during the GridION
543 run. These files were demultiplexed with guppy_barcoder (Version 5.0.11+2b6dbffa5) using "--
544 trim_barcodes --barcode_kits EXP-NBD104" . For singleplex sample (19_326) adapters were
545 trimmed via Porechop (https://github.com/rrwick/Porechop) using: "--discard_middle --
546 extra_end_trim 0". Alignments were performed to GRCh38 with minimap2 (Version 2.13-r850),
547 using the parameters "-ax map-ont --MD", as described in our initial study (*1*).

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

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

559 *Methylation calling of nanopore sWGS data*. 5mC was called using DeepSignal Version: 0.1.8
560 (*4*), with model.CpG.R9.4_1D.human_hx1.bn17.sn360.v0.1.7+/bn_17.sn_360.epoch_9.ckpt,
561 which was downloaded from the DeepSignal Google Drive
562 (https://drive.google.com/open?id=1zkK8Q1gyfviWWnXUBMcIwEDw3SocJg7P). We used the
563 DeepSignal call_mods (mofidication_call) output tsv file. To aid in combined methylation and
564 fragmentomic analysis, we added additional columns to this file The final 14 fields were
565 extracted from Minimap2 alignment files, matched by read id. They are as follows:

566 Column 11: flag, 12: contig, 13: start, 14: end, 15: mapping quality of the chosen
567 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.

608 To calculate "relative" methylation levels, raw methylation levels in each bin were divided by the
609 mean methylation within all bins from -1000 to -800 and +800 to +1000 across all NKX2-1 sites.
610 For NKX2-1 methylation levels in TCGA lung and non-lung samples, we downloaded TCGA
611 WGBS bedgraph files from https://zwdzwd.github.io/pmd (*5*). We used all WGBS cancer types
612 that were represented by normal tissues in the scATAC-seq atlas, as this was the atlas used to
613 define pneumocyte specific (PAL) peaks. These TGCA types included LUAD and LUSC (Lung
614 tissue from atlas), CRC (Transverse colon tissue from atlas), BRCA (Breast tissue from atlas),
615 STAD (Stomach tissue from atlas), and UCEC (Uterus tissue from atlas). Plasma WGBS of
616 HCC was downloaded from EGAD00001004317. Plasma WGBS of CRC patients was
617 downloaded from EGAD00001004568. Plasma WGBS of healthy controls was downloaded from
618 EGAD00001001602.

619 *450k healthy tissue reference atlas.* To compose the atlas of differentially-methylated probes in
620 25 human tissues and cell types, we used the data collected and tissue-specific feature
621 selection method from the MethAtlas package (https://github.com/nloyfer/meth_atlas)(*7*). The
622 script feature_selection.m was used to select Lung_cell specific CpGs for 2,000
623 hypermethylated and 2,000 hypomethylated probes. We removed any probe that did not have
624 valid (non-NA) values for 2 or more of the Lung_cell samples and 2 or more of the healthy
625 plasma samples. We plotted the values using the a custom R script available at
626 (https://github.com/methylgrammarlab/cfdna-
627 ont/deconvolution_code/cell_types_probes/plot_tumor_fractions_vs_score.R).

628 *450k TCGA tumor reference atlas.* We downloaded the Infinium 450k beta value files for TCGA
629 Lung Adenocarcinoma (LUAD) tumors using the ELMER packaged in Bioconductor (*8*). We
630 removed any probe that did not have valid (non-NA) values for 2 or more of the LUAD samples
631 and 2 or more of the healthy plasma samples. We then performed a t-test to compare the
632 methylation beta values of these Lung_specific probes to the four plasma cfDNA samples from
633 the MethAtlas paper (*7*), requiring a Benjamini-Hochberg corrected FDR of <0.01 and an
634 absolute beta value difference of 0.3 or greater. We plotted the values using the a custom R
635 script available at (https://github.com/methylgrammarlab/cfdna-
636 ont/deconvolution_code/TCGA_probes/plot_tumor_fractions_vs_score.R file).

637 *450k cell type deconvolution.* First, CpGs covering either the forward or reverse strand of each
638 CpG on the Infinium 450k array were extracted from each Nanopore beta value file (averaged
639 by taking the total number of methylated reads on either strand divided by the total number of
640 methylated+unmethylated reads on both strands) to produce a beta value vector Y. Each of
641 these files was intersected with the normal Lung_cell specific probes as described above (an
642 example of this is shown in Fig 1A). For each probe, the 450k beta values were averaged to
643 produce a single Lung-specific beta value $\underline{X_1}$. The same was done for the four plasma cfDNA
644 samples from (*7*) to yield a healthy cfDNA beta value $\underline{X_2}$. We used the Lawson-Hanson
645 algorithm for non-negative least squares (NNLS) (https://cran.r-project.org/web/packages/nnls)
646 to perform non-negative least squares regression as in (*7*). Specifically, we identified non-
647 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

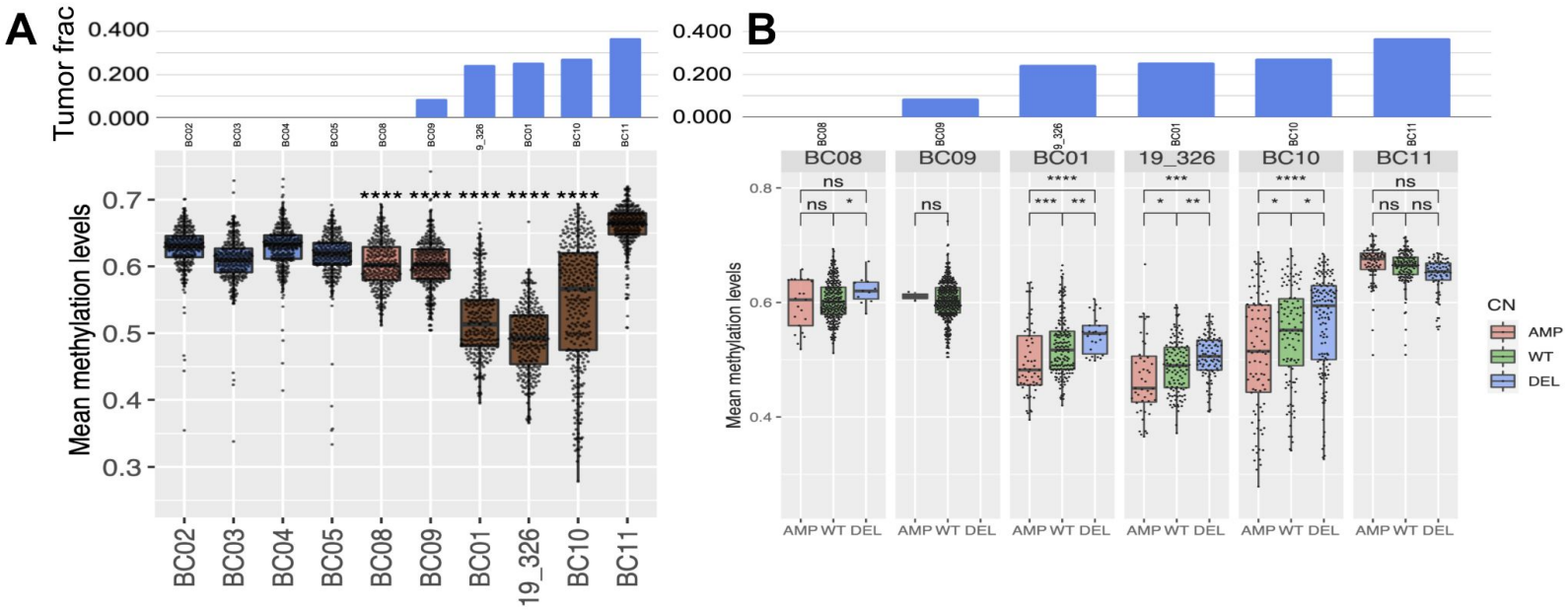
686 ## Supplemental figures

687

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

696

Supplemental Figure 1: Hypomethylation within PMDs

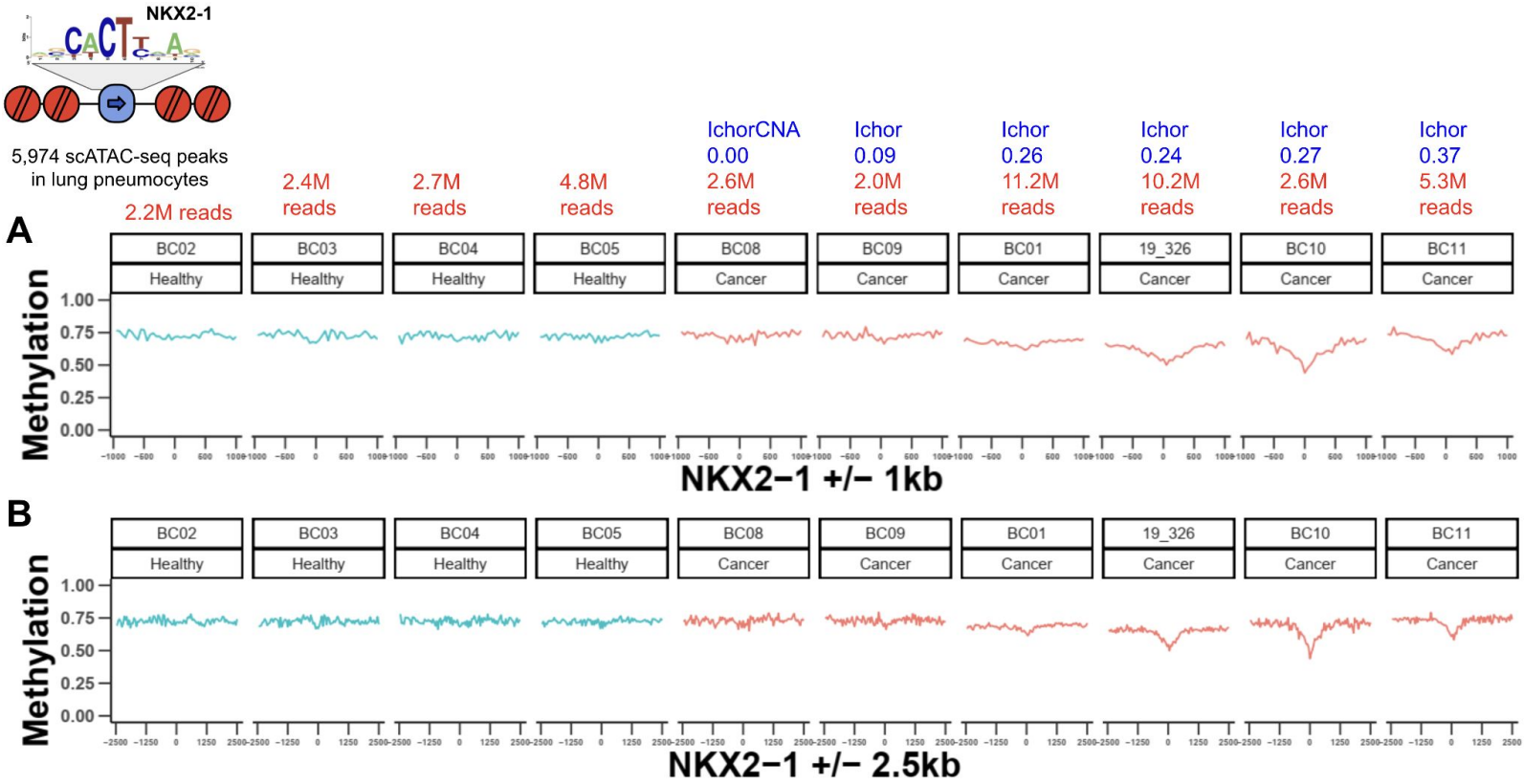


697 **Supplemental Figure 2. Methylation at Lung-specific NKX2-1 binding sites**. Illustration of
698 predicted NKX2-1 binding sites within single-cell ATAC-seq peaks specific to primary lung
699 pneumocytes. (A) Relative methylation level of healthy controls and lung cancer samples from
700 ONT cfDNA WGS within -1kb to +1kb of pneumocyte-specific NKX2-1 sites. Cancer cases are
701 ordered by tumor fraction estimates from ichorCNA. (B) Same analysis, but within -2.5kb to
702 +2.5kb of pneumocyte-specific NKX2-1 sites.

703

704

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

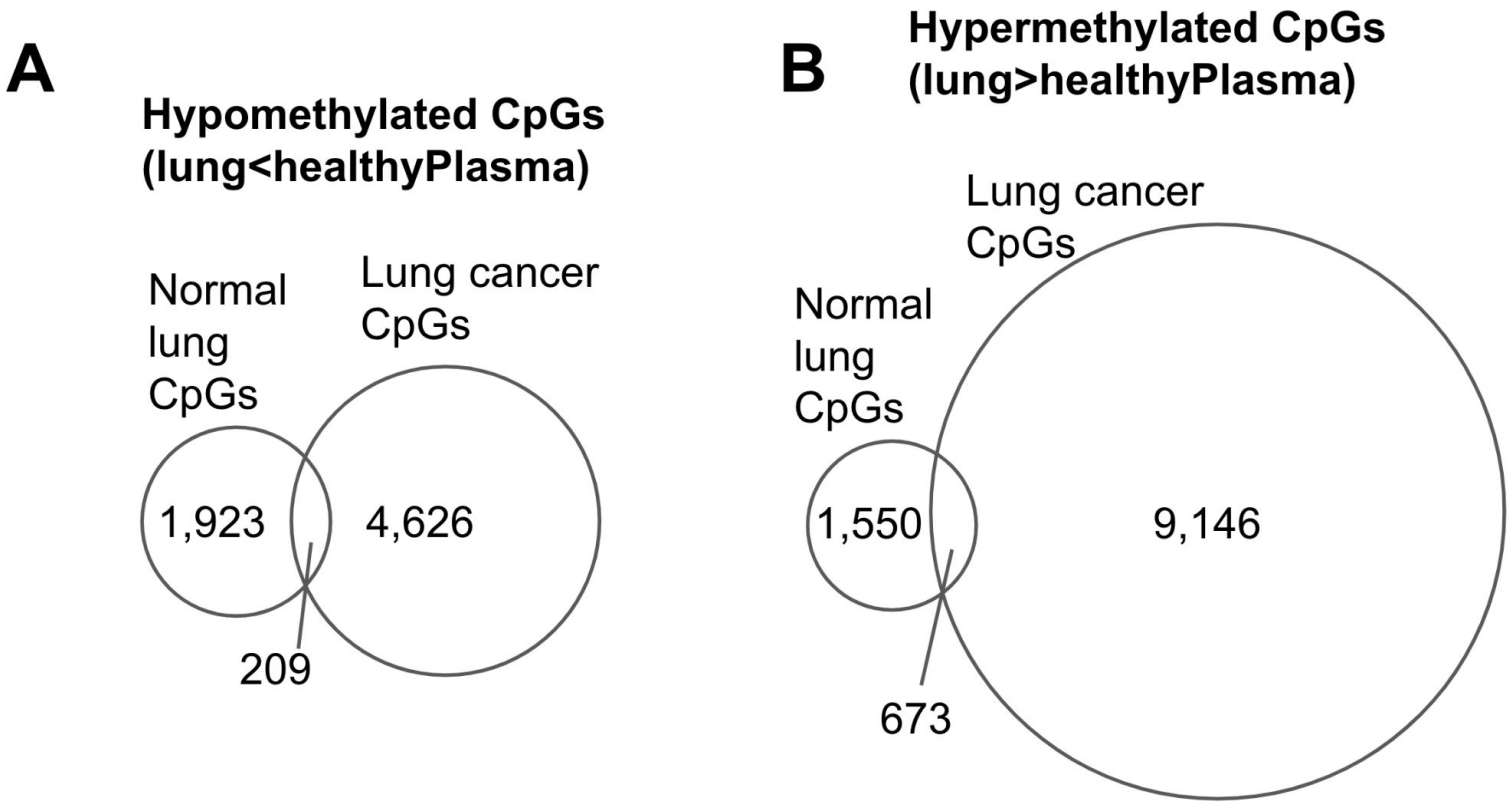


705 **Supplemental Figure 3. Deconvolution of ctDNA using HumanMethylation450k reference**
706 **data.** Estimated lung fraction β plotted for all Nanopore plasma samples. (A) Number of probes
707 hypomethylated in Normal lung cells relative (red) or Lung cancer cells (blue) relative to healthy
708 plasma 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 ctDNA using HumanMethylation450k reference data

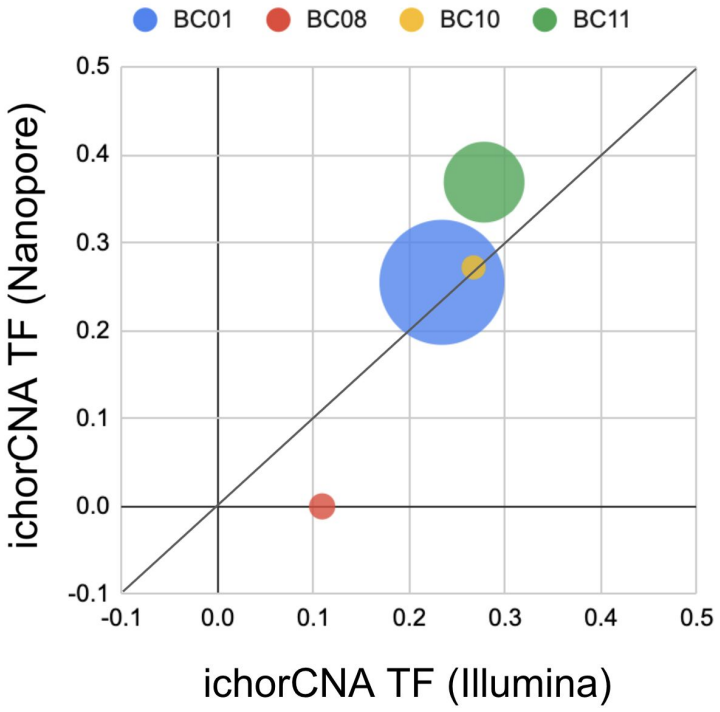


712 **Supplemental Figure 4. ichorCNA tumor fraction estimates from deep Illumina WGS.** (B)
713 ichorCNA (*12*) tumor fraction estimates for samples with matched Nanopore and Illumina WGS
714 data. Size of each bubble is the number of uniquely aligned reads in the Nanopore library. The
715 median number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

717

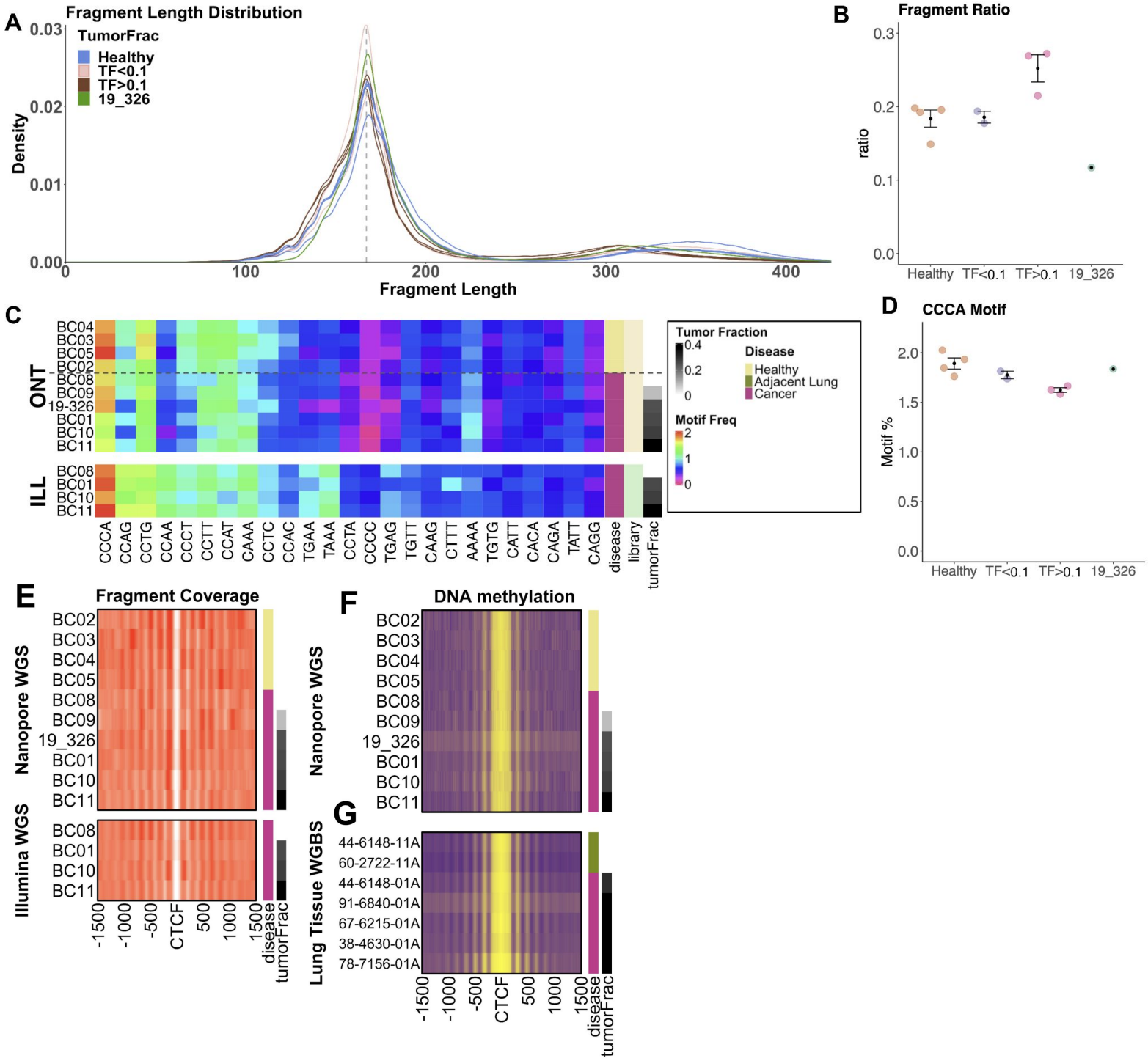
718

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



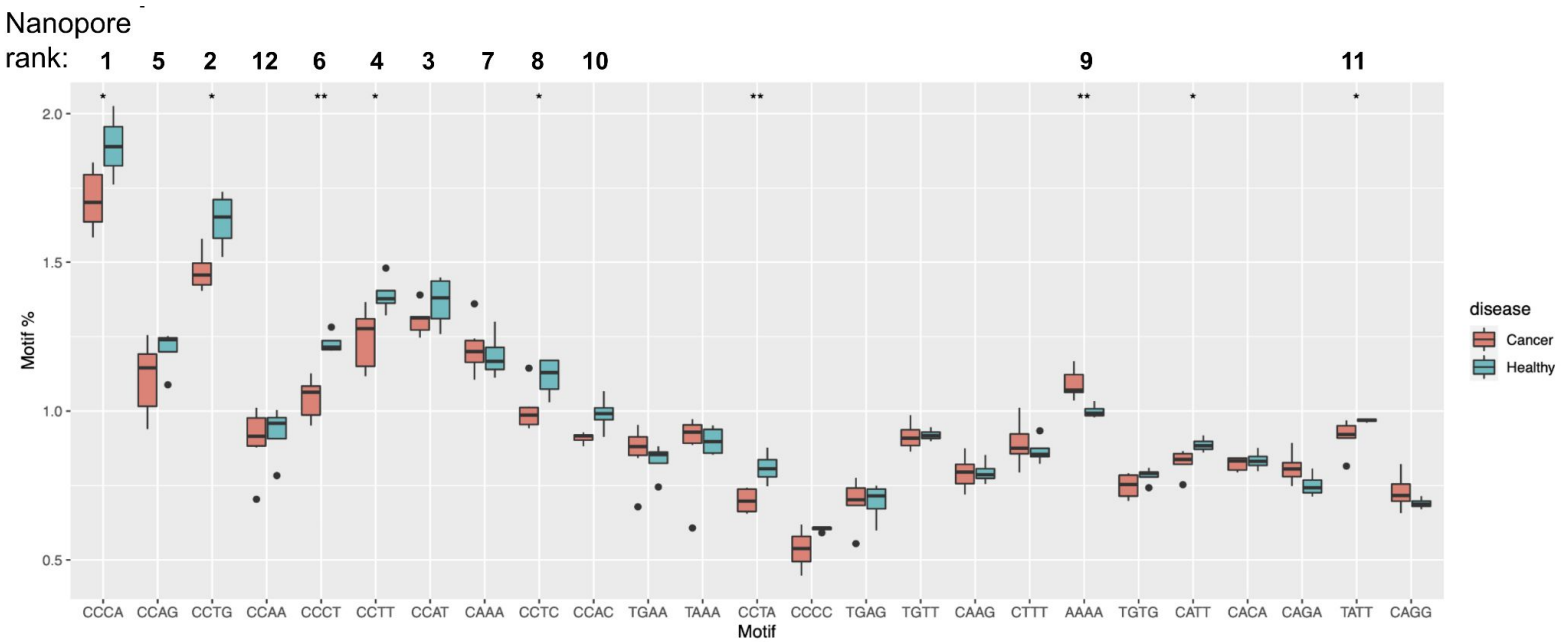
719 **Supplemental Figure 5. Fragment analysis of sample 19_326.** Sample 19_326 used different bead
720 selection parameters and a different library version, and thus is not included in the main analysis. Here,
721 we show 19_326 in context in all fragmentomic analyses. (A-D) General fragment properties. (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



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

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

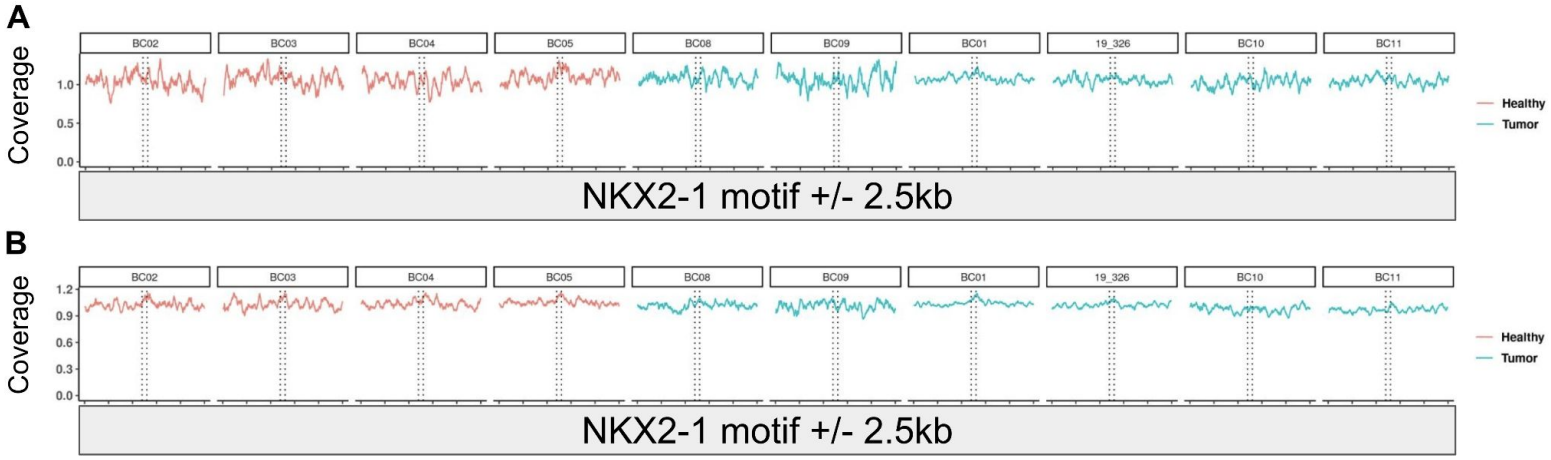


Top 25 ranked 4-mers in healthy plasma (Chan et al. 2020)

739 **Supplemental Figure 7. Fragment analysis of Lung-specific NKX2-1 binding sites.** (A)
740 Analysis of 5,974 lung-specific regions aligned to predicted NKX2-1 binding sites contained
741 within single-cell ATAC-seq peaks specific to primary lung pneumocytes (same sites as Figure
742 1H-K). Plot shows fragment coverage relative to average coverage across the genome,
743 including only fragments of length 130-155bp. (B) Same plot, but using 28,298 predicted NKX2-
744 1 binding sites from ATAC-seq peaks of TCGA Lung adenocarcinoma samples instead of
745 ATAC-seq peaks from normal lung pneumocytes (same sites from Supplemental Figure 3).

746

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



747 ## References

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

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

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

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

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

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

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

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

781 9. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell*.
782 **33**, 676–689.e3 (2018).

783 10. F. Mouliere, D. Chandrananda, A. M. Piskorz, E. K. Moore, J. Morris, L. B. Ahlborn, R. Mair,
784 T. Goranova, F. Marass, K. Heider, J. C. M. Wan, A. Supernat, I. Hudecova, I. Gounaris, S.
785 Ros, M. Jimenez-Linan, J. Garcia-Corbacho, K. Patel, O. Østrup, S. Murphy, M. D.
786 Eldridge, D. Gale, G. D. Stewart, J. Burge, W. N. Cooper, M. S. van der Heijden, C. E.
787 Massie, C. Watts, P. Corrie, S. Pacey, K. M. Brindle, R. D. Baird, M. Mau-Sørensen, C. A.
788 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