Integrated fragmentomic profile of capture-based low-pass sequencing data enables pan-cancer detection via cfDNA

Zhidong Zhang[1,†], Xuenan Pi[1,†], Chang Gao[1], Jun Zhang[2], Lin Xia[1], Xiaoqin Yan[2], Xinlei Hu[1], Ziyue Yan[2], Shuxin Zhang[3], Ailin Wei[7], Yuer Guo[1], Jingfeng Liu[5], Ang Li[4], Xiaolong Liu[5], Wei Zhang[6], Yanhui Liu[3], Dan Xie1[*]

[1]National Frontier Center of Disease Molecular Network, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, No. 37 Guoxue Alley, Chengdu 610041, Sichuan Province, P. R. China.

[2]Tailai Inc., Shanghai 200233, P. R. China.

[3]Department of Neurosurgery, West China Hospital, Sichuan University, No. 37 Guoxue Alley, Chengdu 610041, Sichuan Province, P. R. China.

[4]Department of Pancreatic Surgery, West China Hospital, Sichuan University, No. 37 Guoxue Alley, Chengdu 610041, Sichuan Province, P. R. China.

[5]Mengchao Hepatobiliary Hospital of Fujian Medical University, Xihong Road 312, Fuzhou 350025, Fujian Province, P. R. China.

[6]Department of Respiratory and Critical Care Medicine, First Affiliated Hospital of the Second Military Medical University, Shanghai 200433, Shanghai, P. R. China.

[7]Guang'an People's Hospital, Guang'an, China

*Correspondence to: Dr Dan Xie, National Frontier Center of Disease Molecular Network, State Key Laboratory of Biotherapy, West China Hospital,

Sichuan University, No. 37 Guoxue Alley, Chengdu 610041, Sichuan Province, P. R. China. Tel: +86-28-8558-2944; Fax: +86-28-8558-2944; E-mail: danxie@scu.edu.cn (D. Xie).

† These authors contributed equally to this work as joint first authors.

Abstract

Using epigenetic markers and fragmentomics of cell-free DNA for cancer detection has been proved applicable. We further combine the two features and explore the diagnostic potential of the features on pan-cancer detection. cfDNA fragmentomic features of 396 low-pass 5hmC sequencing cfDNA data from four common cancer types and controls and corresponding WGS data were investigated. We detected pan-cancer preferred end marker specificity between WGS data and low-pass 5hmC sequencing data. We identified aberrant ultra-long fragments of cancer samples in 5hmC sequencing data, both in size and coverage profile, and showed its dominant role in cancer prediction. We built an integrated model including 63 features for pan-cancer detection with high sensitivity and specificity (88.52% and 82.35%, respectively). We showed that fragmentomic information in 5hmC sequencing data is an ideal marker for cancer detection and that it shows high performance in low-pass sequencing data.

Introduction

Displaying a trend of increase, more than 19.3 million individuals were diagnosed with cancer, and 10.0 million individuals died of cancer worldwide in 2020 (1). Currently, tissue biopsy is the most widely used method for cancer diagnosis and staging. However, owing to being an invasive examination method, tissue biopsy has difficulties in tracking clonal evolution, monitoring treatment response, and early detection of cancer and recurrence (2). Therefore, the need for the development of accurate liquid biopsy methods is urgent, with the potential to detect cancer and identify resistance mechanisms early, to quantify minimal residual disease, and to monitor treatment responses (2, 3).

Circulating cell-free tumor DNA (ctDNA) is a kind of tumor-derived material in the blood of patients with cancer and is mainly derived from cancer cells via apoptosis or necrosis (3, 4). ctDNA carries cancer-specific genetic mutations, epigenetic alterations, and fragmentomic aberrations, allowing for cancer detection and tissue-of-origin prediction in a non-invasive manner. Nevertheless, as cancer is a highly heterogeneous disease, biomarkers based on genetic and epigenetic aberrations often have limitations in practical applications. Recent studies have shown that cell-free DNA (cfDNA) fragmentation is a non-random process (5, 6), and it is thus possible to develop a generic approach for non-invasive cancer detection using cfDNA fragmentomics.

The fragmentomics analysis of cfDNA encompasses the sizes (7), coverage

67 (7), nucleosome footprints (8), end-points (9, 10) and end-motifs (11) of cfDNA.
68 For example, the size distribution of cfDNA fragments can indicate specific
69 cellular biological changes (12, 13). Coverage in cfDNA sequencing can be
70 used to analyze nucleosome positions (8, 14) and identify expressed genes
71 (14). Fragmentation patterns of cfDNA can be characterized by deferentially
72 phased fragment end signals, which were preferentially found in tissue-
73 specific open chromatin regions (15). Practically, Jiang et al. showed that
74 cfDNA fragments ending with particular genomic coordinates or motifs had
75 higher degrees of association with hepatocellular carcinoma (HCC) (10, 11).
76 Recently, the epigenetic marker 5-hydroxymethylcytosine (5hmC) has
77 attracted attention in the field of cancer research due to its regulatory role in
78 tissue-specific gene expression (16). The cfDNA captured by 5hmC comprises
79 a subset of all circulating cfDNA and, therefore, may contain fragmentomic
80 information. Integrative multi-omics analysis including the 5hmC profile of
81 cfDNA improves the sensitivity and specificity of cancer detection, as
82 demonstrated by recent studies (17, 18).
83 To highlight the potential of 5hmC sequencing data in cancer detection, we
84 obtained a large cohort of cancer samples, including HCC, pancreatic ductal
85 adenocarcinoma (PDAC), lung adenocarcinoma (LUAD), and glioblastoma
86 (GBM). We then established a pan-cancer preferred end map based on WGS
87 data from these four cancer types and checked their characteristics in low-
88 pass 5hmC sequencing data. We hypothesized that the fragmentomic

89 information would be consistent between WGS and 5hmC sequencing data
90 and tested it by comparison of each feature. We built classification models
91 and investigated the effect of combining 5hmC signatures with fragmentomic
92 information in pan-cancer screening.
93

94 ## Methods

95 ### Sample collection and study design

96 In total, low-pass (mean mapped reads: 38.32 ± 8.02 million) 5hmC-seq data
97 from 85 healthy controls and 311 cancer patients were included in this study.
98 We obtained 5hmC sequencing data from 33 LUAD (18), 74 PDAC (37), 132
99 HCC (38), and 85 control (37) samples from three publications, and 72 GBM
100 samples from an unpublished paper.
101 Plasma samples from 15 healthy controls and 59 cancer patients were
102 collected and subjected to WGS (mean: 14×, range: 8.6-23.7×) to establish
103 the preferred end marker set, while the low-pass WGS data from 19 healthy
104 controls and 98 cancer patients (mean: 3×, range: 2.3-6.2×) were obtained to
105 cross-check with low-pass 5hmC sequencing data. 13 of the samples with
106 low-pass WGS sequencing had above 10x WGS sequencing data. The
107 demographics and clinical characteristics of the cohort are summarized in
108 Supplementary Tables 1 and 2. The design of the study is shown in Figure 1.
109

110 ### Plasma sample collection and cfDNA WGS sequencing

111 We collected plasma samples from 46 GBM patients, 29 PDAC patients and
112 34 controls from West China Hospital. Plasma samples of 37 HCC patients
113 and 32 LUAD patients were collected from Mengchao Hepatobiliary Hospital
114 of Fujian Medical University and the First Affiliated Hospital of the Second
115 Military Medical University, respectively. For every subject, the QIAseq cfDNA
116 All-in-one Kit (Qiagen) was applied on the 2 ml plasma sample for the cfDNA
117 extraction and library construction. Pair-end 150 bp sequencing of the libraries
118 was performed on the Illumina Novaseq 6000 platform.
119
120 **cfDNA 5hmC profiling and sequencing**
121 cfDNA extraction and 5hmC library construction was performed as previously
122 described (28). Firstly, cfDNA was extracted from 2 ml plasma sample using
123 QIAamp Circulating Nucleic Acid Kit (QIAGEN Inc., Valencia, CA, USA). Then,
124 cfDNA (5–10 ng) ligated with sequencing adaptors was incubated in a 25 μL
125 reaction solution containing HEPES buffer (50 mM, pH 8.0), 25 mM MgCl2, 60
126 μM N3-UDP-Glc (ActiveMotif, Carlsbad, CA, USA), and 12.5 U β-
127 glucosyltransferase (NEB, Beverly, MA, USA) for 2 h at 37 °C. Next, 2.5 μL
128 DBCO-PEG4-biotin (Sigma, Carlsbad, CA, USA) was directly added and
129 incubated for 2 h at 37 °C. 10 μg sheared salmon sperm DNA (Life
130 Technologies, USA) was added, then the Micro Bio-Spin 30 Column (Bio-Rad,
131 Hercules, CA, USA) was used to purify the DNA following the instruction, and
132 the final volume was adjusted to 25μL. After that, the purified DNA was

133 incubated with 5 μL C1 streptavidin beads (Life Technologies, USA) in buffer 1
134 (5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl and 0.2% Tween 20) for 30 min.
135 The beads were subsequently undergone three 5-min washes each with
136 buffer 1, buffer 2 (buffer 1 without NaCl), buffer 3 (buffer 1 with pH 9) and
137 buffer 4 (buffer 3 without NaCl). Then, the beads were resuspended in water
138 and amplified with 11 cycles of PCR amplification (initial denaturing at 98 °C
139 for 45 s, followed by 11 cycles of denaturing at 98 °C for 15 s, annealing at
140 60 °C for 30 s, extension at 72 °C for 30 s, and a final extension at 17 °C for 5
141 min). The amplified products were purified using 0.8 × AMPure XP beads
142 (Beck-man Coulter, Fullerton, CA, USA). Pair-end 150 bp sequencing was
143 performed on the Illumina Novaseq 6000 platform.
144

145 **Fragmentomic profiling of cfDNA and cancer prediction analysis**

146 The details of the preferred end, 5hmC signatures, size profile and coverage
147 profile calculation are described in the Supplementary Material. We used a
148 random forest model to distinguish healthy people from cancer patients using
149 fragmentomic features. To estimate the prediction error, we used five cross-
150 validations. Near-zero variance features were removed. The training,
151 validation, and test sets account for 60%, 20%, and 20% of the data,
152 respectively. The samples were selected randomly in a balanced way to keep
153 the ratio of the number of cancer to non-cancer samples similar in the training,
154 validation, and test subsets. Feature importance was calculated using the

training data in each cross-validation run, and we sorted the features according to the mean value of feature importance. The number of features used in the final model was obtained by the highest AUC value in the validation set. To build an integrative model, we fitted the features selected by every feature-alone model into one random forest model, and performed feature selection, training, testing as described above. Random forest machine learning was implemented using the python package sklearn.ensemble.RandomForestClassifier with the following parameters: n_estimators = 100, criterion="gini".

Statistics

All statistical analyses were performed using R version 3.6.1 and python 3.7.0. All tests were two-sided, and P values < 0.05 were considered statistically significant. The ROC-AUC plot and AUC value were implemented using sklearn.metrics.plot_roc_curve.

Ethics

This study was approved by the Ethics Committee of Sichuan Cancer Hospital (SCCHEC-02-2016-005). The written informed consent was obtained from all participants.

Results

Constructing the cfDNA preferred end map

cfDNA fragments ending at specific genomic coordinates that were statistically more abundant than the Poisson distribution was an essential fragmentation signature referred to as the preferred end (10). We utilized WGS data from the four cancer types to build a preferred end map and checked the feature using low-pass data. The basis of utilizing cfDNA preferred end coordinates to indicate the occurrence and location of cancer was its ability to locate nucleosomes, the position of which is related to cell identity (19). To confirm this, we defined upstream (U) ends and downstream (D) ends by their genomic coordinate order on the reference genome calculated via alignment. We surveyed the distance distribution between adjacent U ends, D ends, and nucleosome centers, and the result suggested the preferred ends were enriched on the DNA linker (Supplementary Figure 1a-f). We further checked the distribution of the predicted nucleosome location based on non-malignant tissues at chr12: 34517269-34519122, and the preferred U and D ends of WGS data from controls (n=15) and HCC patients (n=15). As expected, the preferred U and D ends of the controls were enriched at the midpoint between two nucleosomes, which is the location of the DNA linker (Figure 2a), and U ends were found downstream of their nearest D ends. However, in HCC

199 patients, U ends were upstream of their nearest D ends, and only D ends
200 were found at one midpoint between two nucleosomes (Figure 2b), which
201 indicated inconsistent nucleosome positions between HCC patients with
202 controls. We observed the same abnormal preferred end distribution in GBM,
203 LUAD and PDAC patients (Supplementary Figure 1g-i).
204 By using WGS data, we constructed cfDNA fragment preferred end maps for
205 the four cancer types and control (Figure 2c-d). Preferred end enrichment was
206 observed in intron and open chromatin regions (Supplementary Figure 2).
207 Additionally, we noticed that for every cancer type, both U and D preferred
208 ends were more enriched in health-specific open chromatin regions than in
209 cancer-specific open chromatin regions (Supplementary Figure 2). One
210 explanation was the nucleosome pattern in the health-specific open chromatin
211 regions was unstable, as the enrichment of both U and D preferred ends in
212 healthy cfDNA were significantly lower than cancer patients (Mann–Whitney,
213 $P<0.05$). Another was the heterogeneous nature of cfDNA, as cfDNA can be
214 derived from various tissue sources. To search for preferred ends specific to
215 each of the four cancer types, we determined genome-wide cancer-specific
216 preferred U/D end coordinates and health-specific preferred U/D end
217 coordinates based on the Youden index. In total, we got 81,083 D and 80,505
218 U preferred ends. Intriguingly, compared with all repeatable ends (preferred
219 ends appeared in at least two samples in each cancer type), cancer-specific
220 preferred U/D ends were no longer enriched in health-specific open chromatin

221 regions, but they showed enrichment in promoter regions and cancer-specific
222 open chromatin regions (Figure 2e-f, Supplementary Figure 3-4). Furthermore,
223 health-specific preferred U/D ends showed enrichment in intergenic regions
224 and were not found in promoters, as cancer-specific preferred ends were
225 (Supplementary Figure 4). To cross check the cancer specificity of the
226 preferred ends in low-pass WGS data, we calculated a ratio, which was the
227 ratio of detected cancer-specific preferred end numbers to health-specific
228 preferred end numbers for each sample. The results showed significantly
229 higher ratios of cancer-specific to health-specific preferred U/D ends in the
230 three cancer types (except GBM) than in the control (Figure 2g,
231 Mann–Whitney, p value<0.001). In addition to the difference between cancer
232 and health, statistically significant differences in the ratios between cancer
233 types were observed (Supplementary Figure 5).
234
235 **Fragmentomic features in 5hmC data**
236 We checked the batch effect of the 5hmC sequencing data based on size
237 profile and coverage profile, and found no clustering of samples according to
238 cancer types (Supplementary Figure 6).
239 The cfDNA captured by 5hmC marks comprised a subset of DNA in whole
240 plasma. In theory, one sequenced fragment should contain at least one 5hmC.
241 Hence, we investigated the consistency of the size profile between WGS data
242 and 5hmC sequencing data. According to the results, fragment sizes of 5hmC

243 cfDNA in the four cancer samples were significantly shorter than those in the
244 control (Mann–Whitney test, p value<0.001, pan-cancer Median N50 = 174 bp,
245 control Median N50 = 210 bp), though the difference between cancer and
246 control groups in WGS data was only observed for HCC and LUAD (Figure
247 3a). Next, we confirmed that the cfDNA size profile of the 5hmC sequencing
248 data was consistent with that of the WGS data in the size range of 0-220 bp
249 (Figure 3b), with both data types displaying a dominant peak at 167 bp
250 (Supplementary Figure 7a). A difference between the cancer and control
251 samples in both short (100-150 bp) and long (151-220 bp) cfDNA fragments
252 (7) was found in the 5hmC data, as in the WGS data (Supplementary Figure
253 7b-c), though the 5hmC data showed a unique secondary peak at ~320 bp
254 (Figure 3c, Supplementary Figure 7a). In the WGS data, cfDNA fragments
255 ranging from 0-220 bp accounted for nearly 100% of fragments ranging from
256 0-800 bp, but the percent value of the same size interval in the 5hmC
257 sequencing data ranged from 31.8% to 99.6% (Figure 3f). Thus, we confirmed
258 the presence of ultra-long fragments (221-500 bp) in the 5hmC cfDNA. We
259 suspected that the existence of ultra-long fragments may have resulted from
260 the capture-based technique applied in 5hmC sequencing, and we
261 consequently assessed the relationship between fragment length and 5hmC
262 peak number. In ultra-long fragments, the percentage of fragments with more
263 than one peak was significantly higher than that in non-ultra-long fragments
264 (0-220 bp) (Mann–Whitney test, p value<0.001) (Figure 3d). Indeed,

265 enrichment analysis of the ultra-long fragments with more than one peak
266 revealed primary enrichment at CpG islands (Supplementary Figure 7d). This
267 indicated that the capture-based sequencing technique is more likely than
268 WGS to capture ultra-long cfDNA fragments. Another finding was that in all
269 cancer types, the percentage of ultra-long fragments was significantly lower
270 than that in the control (Supplementary Figure 7b). We examined the
271 percentage of ultra-long fragments with more than one 5hmC peak in the four
272 cancer types and the control and found that the value was significantly lower
273 in the former (Figure 3e). As reported by a recent study, the stability of
274 circulating DNA derives mostly from the nucleosome structure (21). We hence
275 evaluated the proportion of 5hmC cfDNA fragments at 146 bp, 166 bp, 312 bp,
276 and 332 bp, which indicates the length of one mono-nucleosome and one di-
277 nucleosome (plus the linker size). The result showed a higher proportion of
278 mono-nucleosome-sized cfDNA fragments and a lower proportion of di-
279 nucleosome-sized cfDNA fragments in the cancer group than in the control
280 (Supplementary Figure 7e).
281 We also explored the cancer specificity of the pan-cancer preferred end set
282 built upon WGS data in the 5hmC sequencing data by calculating the ratio of
283 cancer-specific to health-specific preferred U/D ends. The ratios were
284 significantly increased in cfDNA samples in three cancer types (except PDAC)
285 compared with healthy individuals (Figure 3g, Supplementary Figure 5c-d,
286 Mann–Whitney, p value<0.005). Coverage profiles in 5hmC sequencing data

287 also showed a higher variance than WGS data in samples from control and
288 cancer (Supplementary Figure 8). We explored the coverage patterns of short,
289 long, and ultra-long cfDNA fragments at the 5 Mb bins of the genome. Similar
290 to WGS data (Supplementary Material), cancer patients exhibited multiple
291 unstable genomic regions: coverage of short/long/ultra-long fragments was
292 inconsistent in cancer patients but consistent in controls (Figure 4a,
293 Supplementary Figure 9a). We analyzed the genome-wide coverage profile
294 correlation of cancer patients to controls in the 5hmC data and found that the
295 coverage profiles were consistent among controls and that the correlation
296 value in cancer patients was significantly lower (Wilcoxon rank-sum test, p
297 value<0.005). Importantly, the largest difference between cancer samples and
298 controls was found with respect to the coverage profile of ultra-long fragments
299 (Wilcoxon rank-sum test, p value<0.005) (Figure 4b).
300

301 **Genomic distribution of 5hmC in cancer and control cohorts**

302 To gain an understanding of the genomic regions associated with
303 hydroxymethylation in cfDNA, we first determined 5hmC-enriched loci in the
304 control and four types of cancer, which were detected as peaks via MACS2
305 (22). 5hmC-modified regions among samples were compared in 1 kb bins on
306 the reference genome, and thus a consensus list of the absence and
307 presence of 5hmC-modified peaks among the samples was obtained. 5hmC
308 signatures were enriched over genic features, most significantly in the

309 promoter, 5'UTRs, 3'UTRs, exons, and transcription end sites (TESs) (Figure
310 5a). Comparison of 5hmC signature enrichment in the four types of cancers
311 with the control revealed significant differences, whereby 5hmC peak
312 enrichment was lower in all cancer types than in the control over promoters,
313 5'UTRs, exons, and CpG islands (Wilcoxon rank-sum test, p value<0.001)
314 (Figure 5a).
315 Differential analysis of 5hmC-modified peaks between cancer patients and
316 controls by Fisher's exact test detected 1,010, 6,395 and 773 differentially
317 modified 5hmC peaks in GBM, HCC and PDAC, respectively. Moreover, this
318 pan-cancer 5hmC peak set was able to separate most cancer samples from
319 the control (Figure 5b). KEGG enrichment analysis of genes located within the
320 differentially modified 5hmC peaks revealed associated oncogenesis
321 pathways for each cancer type (Supplementary Figure 10), such as the
322 neurotrophin signaling pathway and platelet activation for GBM (23, 24), the
323 MAPK signaling pathway and focal adhesion for HCC (25, 26), and the FoxO
324 signaling pathway and insulin signaling pathway for PDAC (27).
325
326 **Cancer detection by combining fragmentomic features and 5hmC signatures using**
327 **5hmC sequencing data**
328 As 5hmC sequencing data retain both 5hmC signatures and fragmentomic
329 information, it is expected to theoretically show high sensitivity and specificity
330 in cancer detection. A receiver operator characteristic (ROC) curve was used

331 to evaluate the performance of the classifier. In total, 53 5hmC signatures
332 were selected, and the AUC of the classifier was 0.876 in the validation set
333 (sensitivity = 82.26%, specificity = 82.35%) and 0.872 in the test set
334 (sensitivity = 81.97%, specificity = 82.35%) (Figure 5d-e, Supplementary
335 Figure 11 c).
336 Overall, the fragmentomic information in the 5hmC sequencing data
337 performed well. Using the size profile as features, 40 were selected to build a
338 pan-cancer prediction model, among which 32 features were of the ultra-long
339 and longer size range (>= 220 bp) (Figure 5c). The AUC value was 0.981 in
340 the validation set (sensitivity = 93.55%, specificity = 94.12%) and 0.882 in the
341 test set (sensitivity = 71.43%, specificity = 88.24%) (Figure 5d-e). Using
342 preferred ends as features, 37 were selected to build a pan-cancer prediction
343 model. ROC analysis showed an AUC value of 0.940 in the validation set
344 (sensitivity = 83.87%, specificity = 94.12%) and 0.899 in the test set
345 (sensitivity = 73.02%, specificity = 94.12%) (Figure 5d-e, Supplementary
346 Figure 11a). Regarding the coverage profile of the 5hmC sequencing data, we
347 explored the genomic distribution of short, long, and ultra-long cfDNA
348 fragments on a 100 kb window of the genome. This model achieved an AUC
349 value of 0.946 in the validation set (sensitivity = 80.65%, specificity = 94.12%)
350 and 0.882 in the test set (sensitivity = 71.43%, specificity = 88.24%) (Figure
351 5d-e).
352 Finally, we constructed an integrated cancer screening model by combining

353 fragmentomic features and 5hmC signatures in 5hmC data. Sixty-three
354 features, including 10 5hmC signatures, 24 size profile features, 21 coverage
355 profile features, and 8 preferred end features, were selected to build the pan-
356 cancer prediction model. The AUC value was 0.927 in the validation set
357 (sensitivity = 93.44%, specificity = 88.24%) and 0.920 in the test set
358 (sensitivity = 88.52%, specificity = 82.35%) (Figure 5d-e). The performance of
359 single-cancer detection is depicted in Supplementary Figure 11b.
360 With data from the four cancer types, we further explored tissue-of-origin
361 prediction. The performance of the random forest model with 5hmC
362 fragmentomic features and the integrated model are shown in Supplementary
363 Tables 3-5.
364

365 **Discussion**

366 cfDNA 5hmC signatures are reported to have potential for cancer detection,
367 such as in lung cancer (28), hepatocellular carcinoma (28), colon cancer (29),
368 gastric cancer (29), and pancreatic cancer (17, 30). The large-scale cohort of
369 5hmC sequencing data utilized in this study additionally shows the potential of
370 cfDNA 5hmC signatures for cancer detection in glioblastoma and pan-cancer.
371 Besides, we further examined the classification effect using fragmentomic
372 features in tissue-of-origin prediction of cancer. Most importantly, we explored
373 cfDNA fragmentomic information in 5hmC sequencing data and found that it is
374 possible to detect cfDNA hydroxymethylation and fragmentomic markers

375 simultaneously. The size profile, preferred end, and coverage profile in 5hmC
376 sequencing data showed a large difference between cancer patients and
377 healthy individuals. The integrated model covering the 5hmC signature, size
378 profile, preferred end, and coverage profile contained more information than
379 the model with the 5hmC signature alone, as revealed by higher sensitivity
380 and specificity in pan-cancer detection.
381
382 Previous work reported a difference in the size profile between specific cancer
383 patients or pan-cancer patients and healthy controls (12, 13, 31), though the
384 analyses only focused on lengths less than 320 bp. Here, we characterized
385 ultra-long fragments in 5hmC sequencing data and identified their size and
386 coverage profile aberrations in cancer samples. Although it is commonly
387 acknowledged that ctDNA tends to be more fragmented than cfDNA in normal
388 tissue (12, 31), we found an even larger size difference in cfDNA between
389 cancer samples and controls in 5hmC data owing to ultra-long fragments.
390 Hence, we further expanded the size range to 800 bp by setting 80 adjacent,
391 non-overlapping bins to capture 10-bp interval signals (32). The coverage
392 profile is another major improvement we made with regard to previously
393 reported models based on fragmentation information (9). Indeed, only the
394 ratio of short to long cfDNA fragments has been analyzed in previously
395 reported models, with information on cfDNA ultra-long fragments being lost.
396 Our results suggest that the coverage distribution of size-selected cfDNA

397 fragments in ultra-long fragments has the strongest inconsistency between
398 control and cancer samples (Figure 4a). Of note, the classification models
399 selected the majority of features from ultra-long and longer size range
400 fragments in the coverage profile model (34/42), size profile model (32/40)
401 (Figure 5c, Supplementary Figure 12c) and integrated model (41/63),
402 suggesting that ultra-long features are a major contribution to the
403 classification model. Further research on the mechanism of the aberration of
404 the captured ultra-long fragments in cfDNA from cancer samples may provide
405 insights into the fundamental biological properties of plasma cfDNA from
406 patients with cancer.
407
408 In our work, we constructed a single-base resolution preferred end set based
409 on a set of WGS data. Our results indicated that the preferred end set had
410 good generalization ability because the differences between patients with
411 cancer and controls were also shown in other independent data sets, e.g.,
412 low-pass WGS data and 5hmC data. It has been reported that referred end
413 coordinates differ between HCC patients and healthy people at a whole-
414 genome scale (10). Apart from HCC, our results suggest the cfDNA preferred
415 end can also be used to distinguish patients with GBM, LUAD, and PDAC. We
416 herein discussed the distribution of cfDNA preferred ends in genomic features,
417 open chromatin regions, and nucleosome structure. The distribution of cfDNA
418 preferred ends can reflect nucleosome positioning, though there is no

419 apparent preference in genomic features and open chromatin regions,
420 consistent with previous work (15). However, informative cancer-specific
421 preferred ends were found to be significantly enriched in promoter and far
422 away from health-specific open chromatin regions (Figure 2e-f,
423 Supplementary Figure 4c-d, Supplementary Figure 5c-d).
424
425 In summary, our results indicate that cfDNA fragmentomic analysis with a
426 nonlinear classification algorithm using low-pass 5hmC sequencing data may
427 provide a simple, low-cost, and highly effective method for cancer detection.
428 The success of utilizing ultra-long fragments in 5hmC sequencing data as
429 biomarkers indicates the potential of other capture-based sequencing
430 approaches, such as cell-free methylated DNA immunoprecipitation and high-
431 throughput sequencing (cfMeDIP–seq) (33), in the diagnosis of cancer.
432 Notably, third-generation sequencing methods are expected to detect long
433 cfDNA molecules, which have been utilized in non-invasive prenatal testing
434 (34). As third-generation sequencing methods can detect sequence context
435 and epigenetic modification at the same time, they can provide simultaneous
436 analysis of genome-wide genetic, fragmentomic and epigenetic detection (35,
437 36). Moreover, standardized and effective workflows for analysis of cfDNA
438 fragments need to be developed.
439
440 **Data availability**

441 The 5hmC sequencing data for controls, LUAD, HCC and PDAC were publicly
442 available as described in the Method section. The 5hmC sequencing data for
443 GBM can be accessed at https://ngdc.cncb.ac.cn/gsa-human/s/15mfS98o.
444

445 **Acknowledgments**

446 This study was supported by the 1·3·5 project for disciplines of excellence,
447 West China Hospital, Sichuan University (ZYYC20006) to D. Xie; Thousand
448 Talents Program of the West China Hospital (0040205401F58) to D. Xie;
449 Sichuan Provincial Foundation of Science and Technology (2020YFS0051) to
450 D. Xie, (2017SZ0006) to Y. Liu; Clinical Research Innovation Project, West
451 China Hospital, Sichuan University (19HXCX009) to Y. Liu; the San Hang
452 Program of the Second Military Medical University, Medical basic research
453 project of the First Affiliated Hospital, the Second Military Medical University
454 (2021JCMS16) to W. Zhang; the Science and technology project of Sichuan
455 Province (2021YFS0109) to A. Li; Post-Doctor Research Project, West China
456 Hospital, Sichuan University (20HXBH035) to S. Zhang; the high quality
457 development of Guang 'an People's Hospital (21FZ003) to A. Wei.

458 **Author Contributions**

459 Conception and design: Dan Xie, Zhidong Zhang, Xuenan Pi and Lin Xia.
460 Collection and assembly of data: Jun Zhang, Xiaoqin Yan, Shuxing Zhang,
461 Ailin Wei, Jingfeng Liu, Ang Li, Xiaolong Liu, Wei Zhang and Yanhui Liu. Data
462 analysis and interpretation: Zhidong Zhang, Xuenan Pi, Chang Gao, Xinlei Hu,

463 Xiyue Yan and Yuer Guo. Manuscript writing: Zhidong Zhang and Xuenan Pi.
464 All authors read and approved the final version of the manuscript.

465 **Competing interests**

466 The authors declare no competing interests.

467

468 **References**

469 1. Sung, H., et al. Global cancer statistics 2020: GLOBOCAN estimates of
470 incidence and mortality worldwide for 36 cancers in 185 countries. *CA*
471 *Cancer. J. Clin.* **0**, 1–41 (2021).
472 2. Heitzer, E., Haque, I. S., Roberts, C. E. S. & Speicher, M. R. Current
473 and future perspectives of liquid biopsies in genomics-driven oncology.
474 *Nat. Rev. Genet.* **20**, 71–88 (2019).
475 3. Siravegna, G., Marsoni, S., Siena, S. & Bardelli, A. Integrating liquid
476 biopsies into the management of cancer. *Nat. Rev. Clin. Oncol.* **14**,
477 531–48 (2017).
478 4. Jahr, S., et al. DNA fragments in the blood plasma of cancer patients:
479 quantitations and evidence for their origin from apoptotic and necrotic
480 cells. *Cancer. Res.* **61**, 1659-65 (2001).
481 5. Serpas, L., et al. Dnase1l3 deletion causes aberrations in length and
482 end-motif frequencies in plasma DNA. *Proc. Natl. Acad. Sci. U. S. A.*
483 **116**, 641-49 (2019).
484 6. Han, D. S. C. , et al. The Biology of Cell-free DNA Fragmentation and

485 the Roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* **106**,
486 202-14 (2020).
487 7. Cristiano, S., et al. Genome-wide cell-free DNA fragmentation in
488 patients with cancer. *Nature.* **570**, 385-89 (2019).
489 8. Snyder, M. W., Kircher M., Hill A. J., Daza R. M. & Shendure J. Cell-free
490 DNA Comprises an In Vivo Nucleosome Footprint that Informs Its
491 Tissues-Of-Origin. *Cell.* **164**, 57-68 (2016).
492 9. Chan, K. C., et al. Second generation noninvasive fetal genome
493 analysis reveals de novo mutations, single-base parental inheritance,
494 and preferred DNA ends. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E8159-68
495 (2016).
496 10. Jiang, P., et al. Preferred end coordinates and somatic variants as
497 signatures of circulating tumor DNA associated with hepatocellular
498 carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E10925-33 (2018).
499 11. Jiang, P., et al. Plasma DNA End-Motif Profiling as a Fragmentomic
500 Marker in Cancer, Pregnancy, and Transplantation. *Cancer. Discov.* **10**,
501 664-73 (2020).
502 12. Mouliere, F., et al. Enhanced detection of circulating tumor DNA by
503 fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018).
504 13. Jiang, P., et al. Lengthening and shortening of plasma DNA in
505 hepatocellular carcinoma patients. *Proc. Natl. Acad. Sci. U. S. A.* **112**,
506 E1317-25 (2015).

507 14. Ulz, P., et al. Inferring expressed genes by whole-genome sequencing
508 of plasma DNA. *Nat. Genet.* **48**, 1273-8 (2016).
509 15. Sun, K., et al. Orientation-aware plasma cell-free DNA fragmentation
510 analysis in open chromatin regions informs tissue of origin. *Genome*
511 *Res.* **29**, 418-27 (2019).
512 16. He, B., et al. Tissue-specific 5-hydroxymethylcytosine landscape of the
513 human genome. *Nat. Commun.* **12**, 4249 (2021).
514 17. Chen, L., et al. Genome-scale profiling of circulating cell-free DNA
515 signatures for early detection of hepatocellular carcinoma in cirrhotic
516 patients. *Cell. Res.* **31**, 589-92 (2021).
517 18. [dataset]* Hu, X., et al. Integrated 5-hydroxymethylcytosine and
518 fragmentation signatures as enhanced biomarkers in lung cancer. *Clin.*
519 *Epigenetics.* **14**, 15 (2022).
520 19. Roadmap Epigenomics Consortium, et al. Integrative analysis of 111
521 reference human epigenomes. *Nature.* **518**, 317-30 (2015).
522 20. Gaffney, D. J., et al. Controls of nucleosome positioning in the human
523 genome. *PLoS. Genet.* **8**, e1003036 (2012).
524 21. Sanchez, C., et al. Circulating nuclear DNA structural features, origins,
525 and complete size profile revealed by fragmentomics. *JCI. Insight.* **6**,
526 e144561 (2021).
527 22. Zhang, Y., et al. Model-based analysis of ChIP-Seq (MACS). *Genome*
528 *Biol.* **9**, R137 (2008).

529 23. Lawn, S., et al. Neurotrophin signaling via TrkB and TrkC receptors
530 promotes the growth of brain tumor-initiating cells. *J. Biol. Chem.* **290**,
531 3814-24 (2015).
532 24. Nolte, I., Przibylla, H., Bostel, T., Groden, C. & Brockmann, M. A.
533 Tumor-platelet interactions: glioblastoma growth is accompanied by
534 increasing platelet counts. *Clin. Neurol. Neurosurg.* **110**, 339-42 (2008).
535 25. Moon, H. & Ro, S. W. MAPK/ERK Signaling Pathway in Hepatocellular
536 Carcinoma. *Cancers (Basel).* **13**, 3026 (2021).
537 26. Gnani, D., et al. Focal adhesion kinase depletion reduces human
538 hepatocellular carcinoma growth by repressing enhancer of zeste
539 homolog 2. *Cell. Death. Differ.* **24**, 889-902 (2017).
540 27. Li, J., et al. Knockdown of FOXO3a induces epithelial-mesenchymal
541 transition and promotes metastasis of pancreatic ductal
542 adenocarcinoma by activation of the β-catenin/TCF4 pathway through
543 SPRY2. *J. Exp. Clin. Cancer. Res.* **38**, 38 (2019).
544 28. Song, C. X., et al. 5-Hydroxymethylcytosine signatures in cell-free DNA
545 provide information about tumor types and stages. *Cell. Res.* **27**, 1231-
546 42 (2017).
547 29. Li, W., et al. 5-Hydroxymethylcytosine signatures in circulating cell-free
548 DNA as diagnostic biomarkers for human cancers. *Cell. Res.* **27**, 1243-
549 57 (2017).
550 30. Guler, G. D., et al. Detection of early stage pancreatic cancer using 5-

551 hydroxymethylcytosine signatures in circulating cell free DNA. *Nat.*
552 *Commun.* **11**, 5270 (2020).
553 31. Lam, W. K. J., et al. Sequencing-based counting and size profiling of
554 plasma Epstein-Barr virus DNA enhance population screening of
555 nasopharyngeal carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5115-
556 24 (2018).
557 32. Lo, Y. M., et al. Maternal plasma DNA sequencing reveals the genome-
558 wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**,
559 61ra91 (2010).
560 33. Shen, S. Y., et al. Sensitive tumour detection and classification using
561 plasma cell-free DNA methylomes. *Nature.* **563**, 579-83 (2018).
562 34. Yu, S. C. Y., et al. Single-molecule sequencing reveals a large
563 population of long cell-free DNA molecules in maternal plasma. *Proc.*
564 *Natl. Acad. Sci. U. S. A.* **118**, e2114937118 (2021).
565 35. Tse, O. Y. O., et al. Genome-wide detection of cytosine methylation by
566 single molecule real-time sequencing. *Proc. Natl. Acad. Sci. U. S. A.*
567 **118**, e2019768118 (2021).
568 36. Liu, Q., et al. Detection of DNA base modifications by deep recurrent
569 neural network on Oxford Nanopore sequencing data. *Nat. Commun.* **10**,
570 2449 (2019).
571 37. [dataset]* Cao, F., et al. Integrated epigenetic biomarkers in circulating
572 cell-free DNA as a robust classifier for pancreatic cancer. *Clin.*

573 *Epigenetics.* **12**, 112 (2020).

574 38. [dataset]* Cai, Z., et al. Liquid biopsy by combining 5-
575 hydroxymethylcytosine signatures of plasma cell-free DNA and protein
576 biomarkers for diagnosis and prognosis of hepatocellular carcinoma.
577 *ESMO Open.* **6**, 100021 (2021).

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

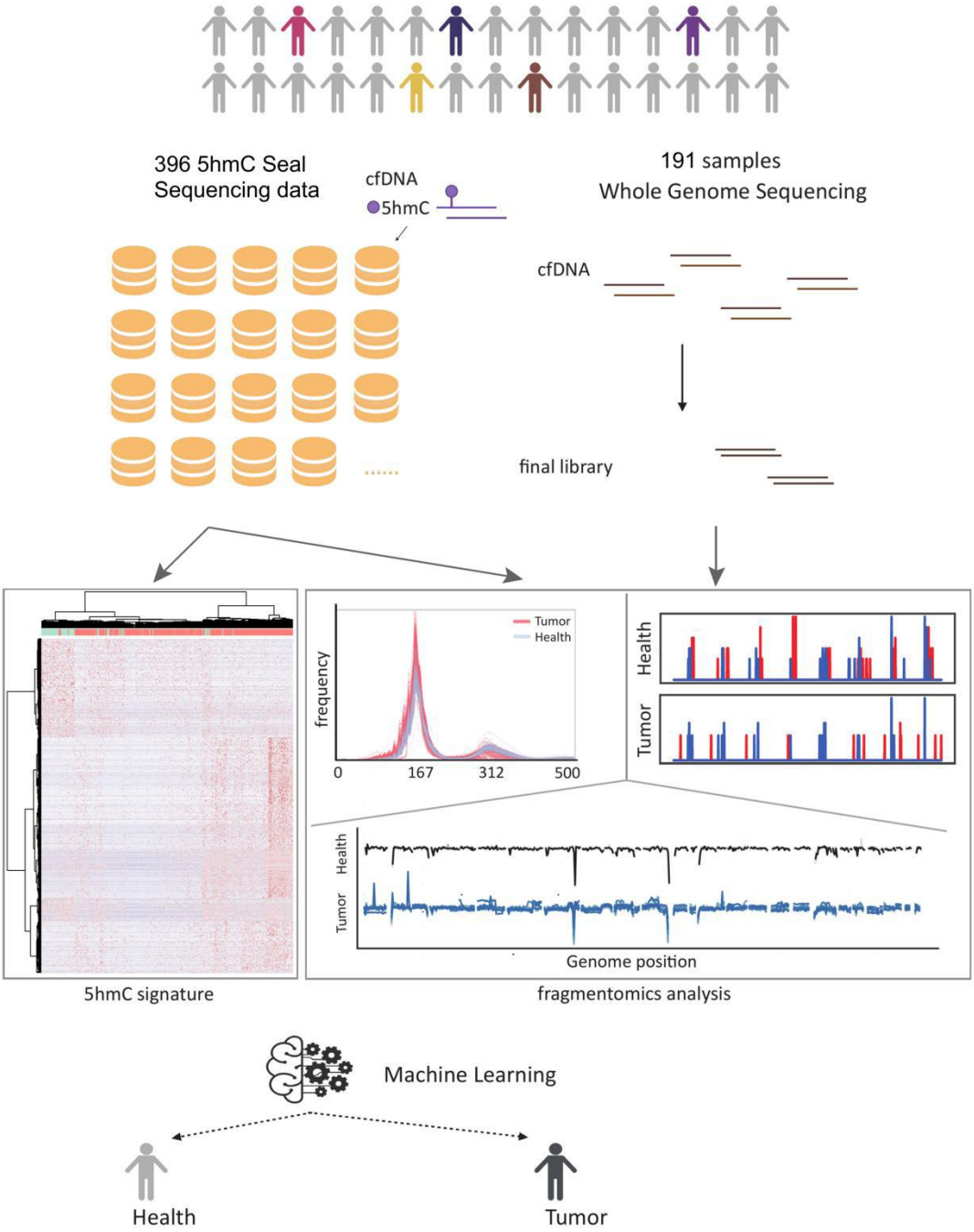
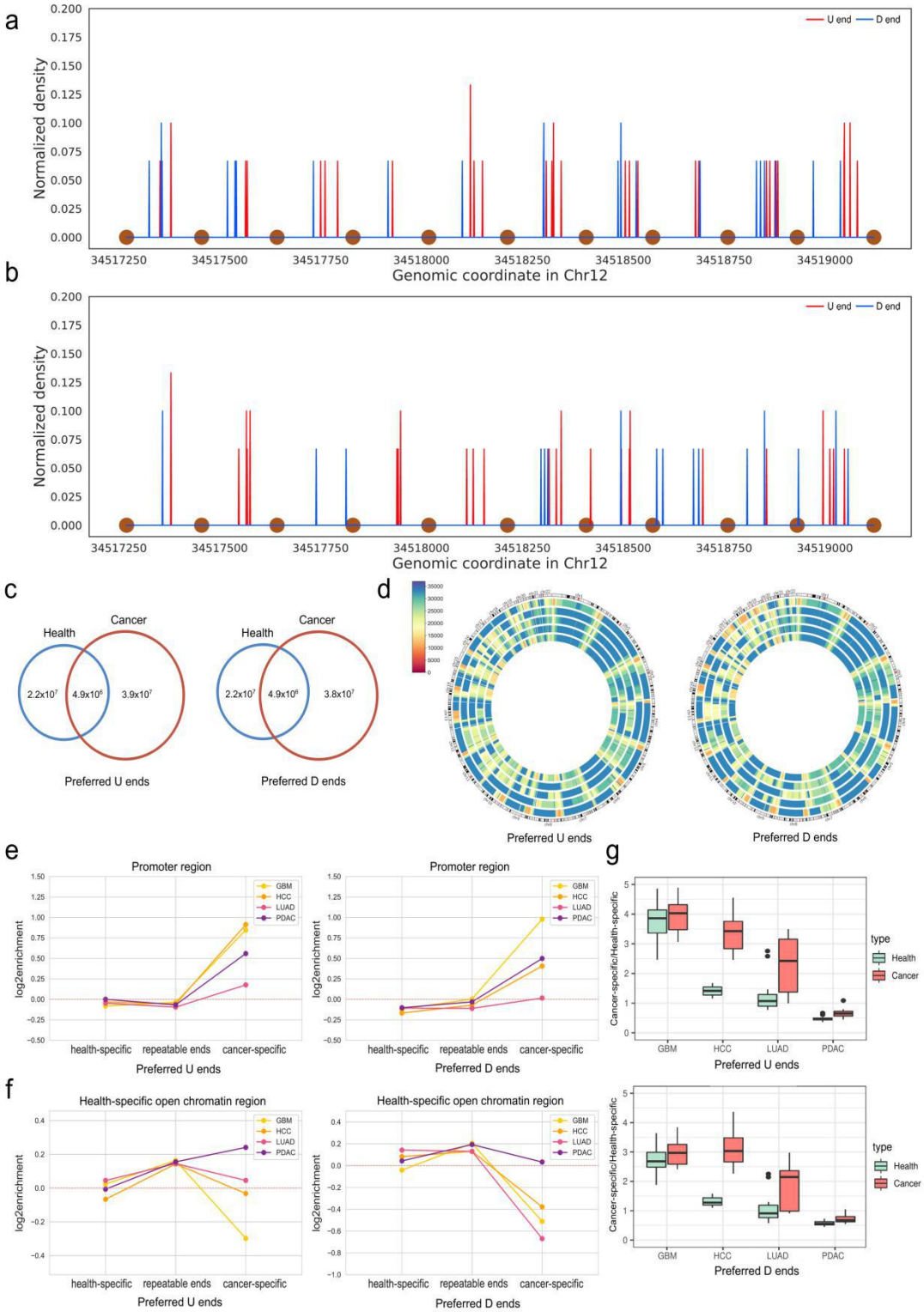


Figure 1 Overview of the study design.

cfDNA from cancer patients and controls were sequenced with WGS. Available 5hmC sequencing data from the same four cancer types and controls were combined with WGS data to get the accurate fragmentomic information and upon it a machine learning model was built to distinguish

601 healthy controls and cancer patients.



602

603 **Figure 2 Genome-wide cfDNA fragment preferred end map construction.**

604 cfDNA preferred end signals in a nucleosome array region (chr12: 34517269-

605 34519122) in healthy (a) and HCC (b) subjects. Brown dots at the bottom
606 represent the predicted nucleosome center loci reported in a previous study (8,
607 22). (c) Venn diagram showing the intersection of the preferred U and D ends
608 between the cancer cohort and the control cohort. (d) Circle diagram showing
609 the density of preferred ends in 1 Mb windows in four types of cancer and in
610 healthy controls (the order of sample sources from outer ring to inner ring are
611 control, GBM, HCC, LUAD and PDAC). (e) Enrichment of preferred U/D ends
612 in promoter regions. (f) Enrichment of preferred U/D ends in health-specific
613 open chromatin regions. (g) Boxplots showing the ratios of the cancer-specific
614 to health-specific preferred U/D ends in low-pass WGS data for the four
615 cancer types (Wilcoxon rank-sum test).

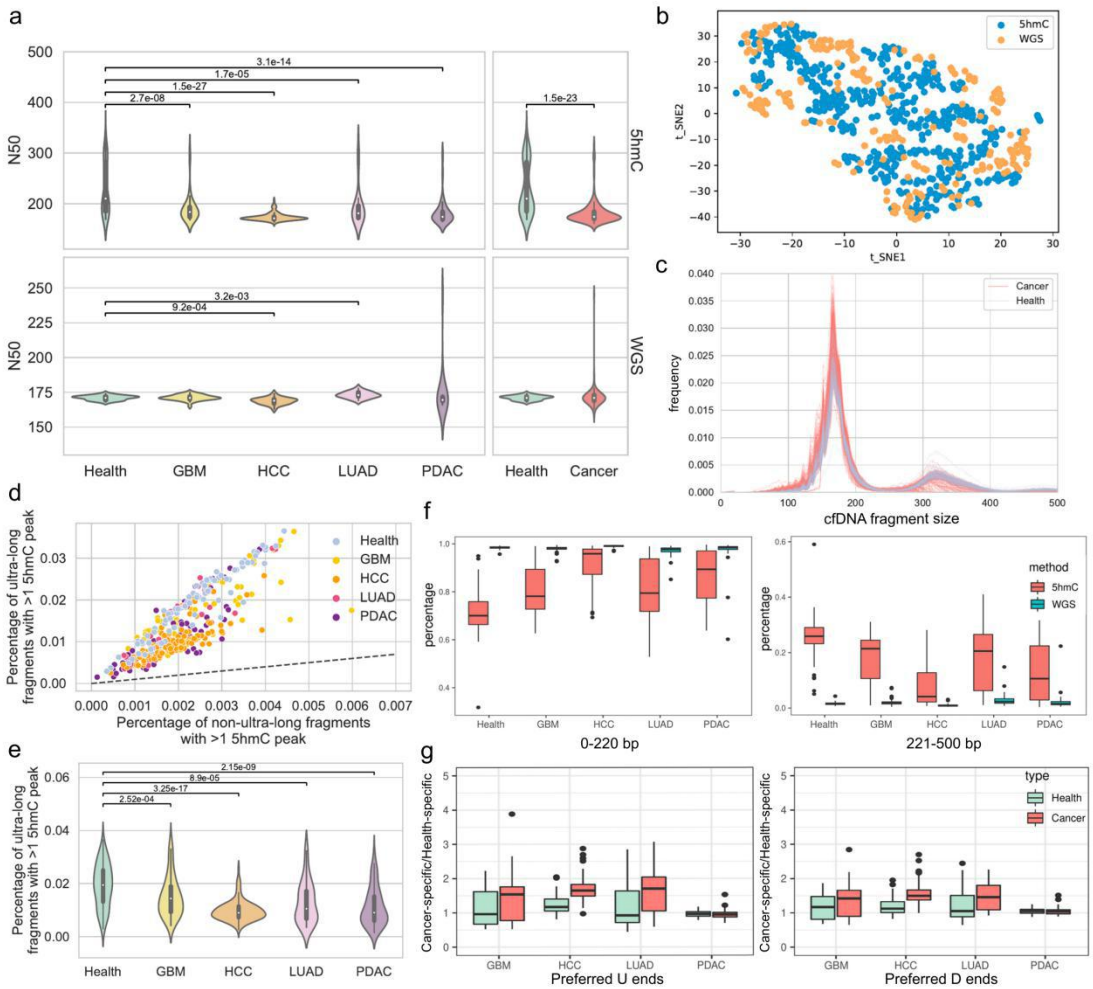


Figure 3 cfDNA fragmentomic information in 5hmC sequencing data.

(a) Violin plots showing fragment size profile comparison between cancer samples and controls in terms of N50 (Mann–Whitney test). (b) tSNE plot showing 5hmC sequencing data and WGS data by fragments below 220 bp. (c) The cfDNA size profile of 5hmC sequencing data. (d) Comparison of the percentage of fragments with more than one 5hmC peak between ultra-long and non-ultra-long fragments in every sample. The dashed line indicates the diagonal line. (e) Violin plots showing the percentage of ultra-long fragments with more than one 5hmC peak in cancer samples and controls (Mann–Whitney test). (f) Boxplots showing the percentage of fragments

627 ranging from 0-220 bp and 221-500 bp among fragments ranging from 0-800
628 bp in WGS data and 5hmC sequencing data (Wilcoxon rank-sum test). (g)
629 Boxplots showing the ratios of the cancer-specific to health-specific preferred
630 U/D ends in each of the four cancer types (Wilcoxon rank-sum test).

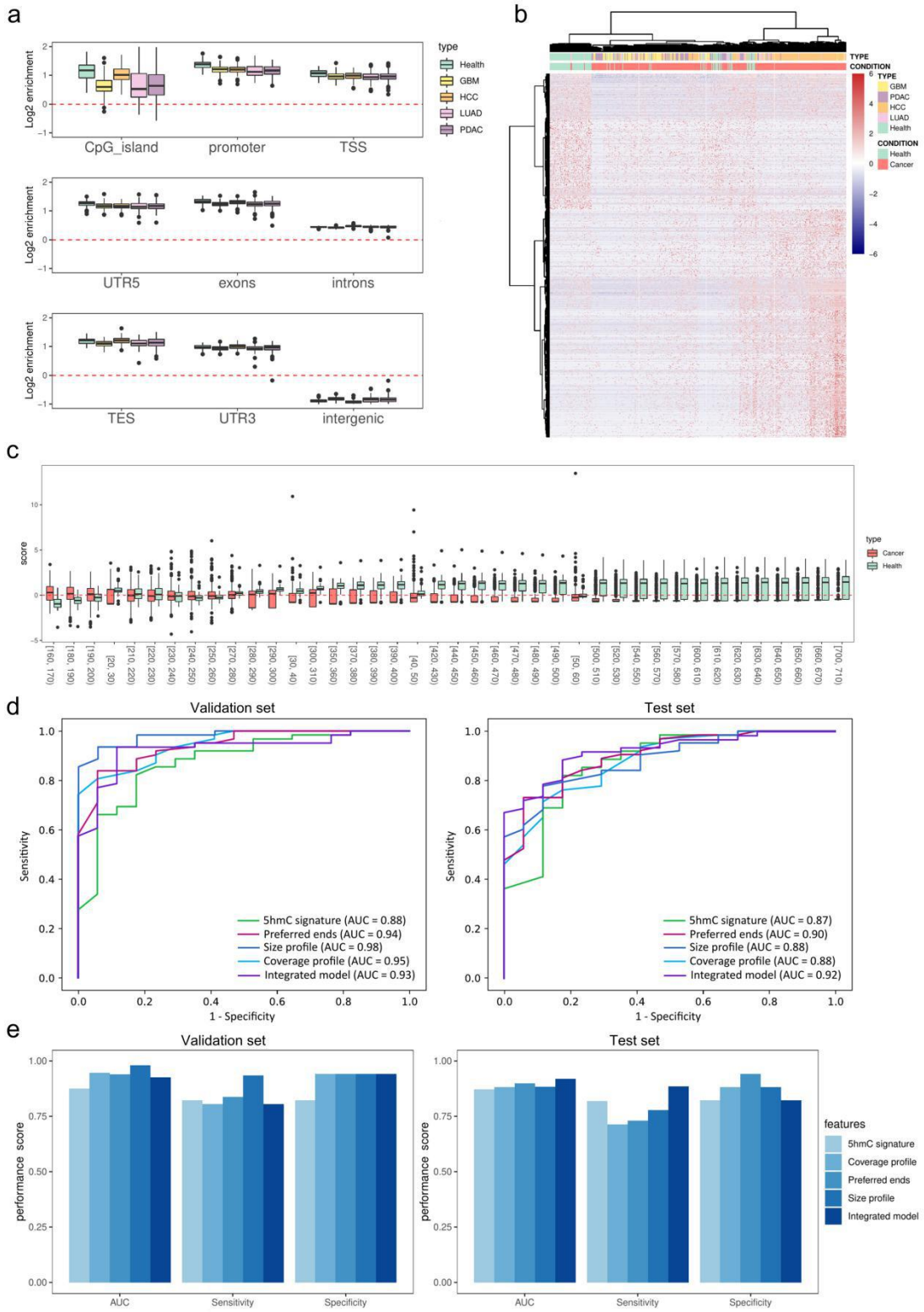
Figure 4 Size-selected coverage profile and correlation to median health of cfDNA fragments in 5hmC sequencing data.

(a) Genome-wide coverage profile of short, long, and ultra-long cfDNA fragments; color indicates sample-wise correlation to median health. (b)

636 Coverage profile correlation to median healthy in control and four types of

637 cancer in short, long, and ultra-long cfNDA fragments (Wilcoxon rank-sum test,

638 p value<0.005).



639

640 **Figure 5 Performance of cancer detection combining fragmentomic**
641 **features and 5hmC signatures using 5hmC sequencing data.**
642 (a) Enrichment of 5hmC in genomic features in control and four types of
643 cancer (Wilcoxon rank-sum test, p value<0.001). (b) Heatmap of differentially
644 modified 5hmC peaks. (c) Selected size profile features in the size profile
645 model. (d) ROC curves for the validation set and test set in pan-cancer
646 diagnosis. (e) Performance evaluation of validation set and test set in pan-
647 cancer diagnosis.
648
649