Integrated fragmentomic profile and 5-Hydroxymethylcytosine 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 proven applicable. We further combine the two features and explore the diagnostic potential of the features on pan-cancer detection. We extracted cfDNA fragmentomic features from 191 whole-genome sequencing data and investigated them in 396 low-pass 5hmC sequencing data from four common cancer types and controls. We identified aberrant ultra-long fragments (220-500bp) of cancer samples in 5hmC sequencing data, both in size and coverage profile, and showed its dominant role in cancer prediction. Since cfDNA hydroxymethylation and fragmentomic markers can be detected simultaneously in low-pass 5hmC sequencing data, we built an integrated model including 63 features of both fragmentomic features and hydroxymethylation signatures 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

45 sequencing data.

46

47 **Introduction**

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

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

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

(GBM). We then established a pan-cancer preferred end map based on WGS data from these four cancer types and checked their characteristics in low-pass 5hmC sequencing data. We hypothesized that the fragmentomic information would be consistent between WGS and 5hmC sequencing data and tested it by comparison of each feature. We built classification models and investigated the effect of combining 5hmC signatures with fragmentomic information in pan-cancer screening.

Methods

Sample collection and study design

In total, low-pass (mean mapped reads: 38.32±8.02 million) 5hmC-seq data from 85 healthy controls and 311 cancer patients were included in this study. We obtained 5hmC sequencing data from 33 LUAD (18), 74 PDAC (37), 132 HCC (38), and 85 control (37) samples from three publications, and 72 GBM samples from an unpublished paper.

Plasma samples from 15 healthy controls and 59 cancer patients were collected and subjected to WGS (mean: 14×, range: 8.6-23.7×) to establish the preferred end marker set, while the low-pass WGS data from 19 healthy controls and 98 cancer patients (mean: 3×, range: 2.3-6.2×) were obtained to cross-check with low-pass 5hmC sequencing data. 13 of the samples with low-pass WGS sequencing had above 10x WGS sequencing data. The demographics and clinical characteristics of the cohort are summarized in

111 Supplementary Tables 1 and 2. The design of the study is shown in Figure 1.

112

113 **Plasma sample collection and cfDNA WGS sequencing**

114 We collected plasma samples from 46 GBM patients, 29 PDAC patients and
115 34 controls from West China Hospital. Plasma samples of 37 HCC patients
116 and 32 LUAD patients were collected from Mengchao Hepatobiliary Hospital
117 of Fujian Medical University and the First Affiliated Hospital of the Second
118 Military Medical University, respectively. For every subject, the QIAseq cfDNA
119 All-in-one Kit (Qiagen) was applied on the 2 ml plasma sample for the cfDNA
120 extraction and library construction. Pair-end 150 bp sequencing of the libraries
121 was performed on the Illumina Novaseq 6000 platform.

122

123 **cfDNA 5hmC profiling and sequencing**

124 cfDNA extraction and 5hmC library construction was performed as previously
125 described (28). Firstly, cfDNA was extracted from 2 ml plasma sample using
126 QIAamp Circulating Nucleic Acid Kit (QIAGEN Inc., Valencia, CA, USA). Then,
127 cfDNA (5–10 ng) ligated with sequencing adaptors was incubated in a 25 μL
128 reaction solution containing HEPES buffer (50 mM, pH 8.0), 25 mM MgCl2, 60
129 μM N3-UDP-Glc (ActiveMotif, Carlsbad, CA, USA), and 12.5 U β-
130 glucosyltransferase (NEB, Beverly, MA, USA) for 2 h at 37 °C. Next, 2.5 μL
131 DBCO-PEG4-biotin (Sigma, Carlsbad, CA, USA) was directly added and
132 incubated for 2 h at 37 °C. 10 μg sheared salmon sperm DNA (Life

133 Technologies, USA) was added, then the Micro Bio-Spin 30 Column (Bio-Rad,
134 Hercules, CA, USA) was used to purify the DNA following the instruction, and
135 the final volume was adjusted to 25μL. After that, the purified DNA was
136 incubated with 5 μL C1 streptavidin beads (Life Technologies, USA) in buffer 1
137 (5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl and 0.2% Tween 20) for 30 min.
138 The beads were subsequently undergone three 5-min washes each with
139 buffer 1, buffer 2 (buffer 1 without NaCl), buffer 3 (buffer 1 with pH 9) and
140 buffer 4 (buffer 3 without NaCl). Then, the beads were resuspended in water
141 and amplified with 11 cycles of PCR amplification (initial denaturing at 98 °C
142 for 45 s, followed by 11 cycles of denaturing at 98 °C for 15 s, annealing at
143 60 °C for 30 s, extension at 72 °C for 30 s, and a final extension at 17 °C for 5
144 min). The amplified products were purified using 0.8 × AMPure XP beads
145 (Beck-man Coulter, Fullerton, CA, USA). Pair-end 150 bp sequencing was
146 performed on the Illumina Novaseq 6000 platform.
147
148 **Fragmentomic profiling of cfDNA and cancer prediction analysis**
149 The details of the preferred end, 5hmC signatures, size profile and coverage
150 profile calculation are described in the Supplementary Material. We used a
151 random forest model to distinguish healthy people from cancer patients using
152 fragmentomic features. To estimate the prediction error, we used five cross-
153 validations. Near-zero variance features were removed. The training,
154 validation, and test sets account for 60%, 20%, and 20% of the data,

respectively. The samples were selected randomly in a balanced way to keep the ratio of the number of cancer to non-cancer samples similar in the training, 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

177 participants.

178

179

180 ## Results

181

182

183 ### Constructing the cfDNA preferred end map

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

199 preferred U and D ends of the controls were enriched at the midpoint between
200 two nucleosomes, which is the location of the DNA linker (Figure 2a), and U
201 ends were found downstream of their nearest D ends. However, in HCC
202 patients, U ends were upstream of their nearest D ends, and only D ends
203 were found at one midpoint between two nucleosomes (Figure 2b), which
204 indicated inconsistent nucleosome positions between HCC patients with
205 controls. We observed the same abnormal preferred end distribution in GBM,
206 LUAD and PDAC patients (Supplementary Figure 1g-i).
207 By using WGS data, we constructed cfDNA fragment preferred end maps for
208 the four cancer types and control (Figure 2c-d). Preferred end enrichment was
209 observed in intron and open chromatin regions (Supplementary Figure 2).
210 Additionally, we noticed that for every cancer type, both U and D preferred
211 ends were more enriched in health-specific open chromatin regions than in
212 cancer-specific open chromatin regions (Supplementary Figure 2). One
213 explanation was the nucleosome pattern in the health-specific open chromatin
214 regions was unstable, as the enrichment of both U and D preferred ends in
215 healthy cfDNA were significantly lower than cancer patients (Mann–Whitney,
216 $P<0.05$). Another was the heterogeneous nature of cfDNA, as cfDNA can be
217 derived from various tissue sources. To search for preferred ends specific to
218 each of the four cancer types, we determined genome-wide cancer-specific
219 preferred U/D end coordinates and health-specific preferred U/D end
220 coordinates based on the Youden index. In total, we got 81,083 D and 80,505

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

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

265 peak number. In ultra-long fragments, the percentage of fragments with more
266 than one peak was significantly higher than that in non-ultra-long fragments
267 (0-220 bp) (Mann–Whitney test, p value<0.001) (Figure 3d). Indeed,
268 enrichment analysis of the ultra-long fragments with more than one peak
269 revealed primary enrichment at CpG islands (Supplementary Figure 7d). This
270 indicated that the capture-based sequencing technique is more likely than
271 WGS to capture ultra-long cfDNA fragments. Another finding was that in all
272 cancer types, the percentage of ultra-long fragments was significantly lower
273 than that in the control (Supplementary Figure 7b). We examined the
274 percentage of ultra-long fragments with more than one 5hmC peak in the four
275 cancer types and the control and found that the value was significantly lower
276 in the former (Figure 3e). As reported by a recent study, the stability of
277 circulating DNA derives mostly from the nucleosome structure (21). We hence
278 evaluated the proportion of 5hmC cfDNA fragments at 146 bp, 166 bp, 312 bp,
279 and 332 bp, which indicates the length of one mono-nucleosome and one di-
280 nucleosome (plus the linker size). The result showed a higher proportion of
281 mono-nucleosome-sized cfDNA fragments and a lower proportion of di-
282 nucleosome-sized cfDNA fragments in the cancer group than in the control
283 (Supplementary Figure 7e).
284 We also explored the cancer specificity of the pan-cancer preferred end set
285 built upon WGS data in the 5hmC sequencing data by calculating the ratio of
286 cancer-specific to health-specific preferred U/D ends. The ratios were

287 significantly increased in cfDNA samples in three cancer types (except PDAC)
288 compared with healthy individuals (Figure 3g, Supplementary Figure 5c-d,
289 Mann–Whitney, p value<0.005). Coverage profiles in 5hmC sequencing data
290 also showed a higher variance than WGS data in samples from control and
291 cancer (Supplementary Figure 8). We explored the coverage patterns of short,
292 long, and ultra-long cfDNA fragments at the 5 Mb bins of the genome. Similar
293 to WGS data (Supplementary Material), cancer patients exhibited multiple
294 unstable genomic regions: coverage of short/long/ultra-long fragments was
295 inconsistent in cancer patients but consistent in controls (Figure 4a,
296 Supplementary Figure 9a). We analyzed the genome-wide coverage profile
297 correlation of cancer patients to controls in the 5hmC data and found that the
298 coverage profiles were consistent among controls and that the correlation
299 value in cancer patients was significantly lower (Wilcoxon rank-sum test, p
300 value<0.005). Importantly, the largest difference between cancer samples and
301 controls was found with respect to the coverage profile of ultra-long fragments
302 (Wilcoxon rank-sum test, p value<0.005) (Figure 4b).
303
304 **Genomic distribution of 5hmC in cancer and control cohorts**
305 To gain an understanding of the genomic regions associated with
306 hydroxymethylation in cfDNA, we first determined 5hmC-enriched loci in the
307 control and four types of cancer, which were detected as peaks via MACS2
308 (22). 5hmC-modified regions among samples were compared in 1 kb bins on

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

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

353 and 0.882 in the test set (sensitivity = 71.43%, specificity = 88.24%) (Figure
354 5d-e).
355 Finally, we constructed an integrated cancer screening model by combining
356 fragmentomic features and 5hmC signatures in 5hmC data. Sixty-three
357 features, including 10 5hmC signatures, 24 size profile features, 21 coverage
358 profile features, and 8 preferred end features, were selected to build the pan-
359 cancer prediction model. The AUC value was 0.927 in the validation set
360 (sensitivity = 93.44%, specificity = 88.24%) and 0.920 in the test set
361 (sensitivity = 88.52%, specificity = 82.35%) (Figure 5d-e). The performance of
362 single-cancer detection is depicted in Supplementary Figure 11b.
363 With data from the four cancer types, we further explored tissue-of-origin
364 prediction. The performance of the random forest model with 5hmC
365 fragmentomic features and the integrated model are shown in Supplementary
366 Tables 3-5.
367

368 **Discussion**

369 cfDNA 5hmC signatures are reported to have potential for cancer detection,
370 such as in lung cancer (28), hepatocellular carcinoma (28), colon cancer (29),
371 gastric cancer (29), and pancreatic cancer (17, 30). The large-scale cohort of
372 5hmC sequencing data utilized in this study additionally shows the potential of
373 cfDNA 5hmC signatures for cancer detection in glioblastoma and pan-cancer.
374 Besides, we further examined the classification effect using fragmentomic

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

397 ratio of short to long cfDNA fragments has been analyzed in previously
398 reported models, with information on cfDNA ultra-long fragments being lost.
399 Our results suggest that the coverage distribution of size-selected cfDNA
400 fragments in ultra-long fragments has the strongest inconsistency between
401 control and cancer samples (Figure 4a). Of note, the classification models
402 selected the majority of features from ultra-long and longer size range
403 fragments in the coverage profile model (34/42), size profile model (32/40)
404 (Figure 5c, Supplementary Figure 12c) and integrated model (41/63),
405 suggesting that ultra-long features are a major contribution to the
406 classification model. Further research on the mechanism of the aberration of
407 the captured ultra-long fragments in cfDNA from cancer samples may provide
408 insights into the fundamental biological properties of plasma cfDNA from
409 patients with cancer.
410
411 In our work, we constructed a single-base resolution preferred end set based
412 on a set of WGS data. Our results indicated that the preferred end set had
413 good generalization ability because the differences between patients with
414 cancer and controls were also shown in other independent data sets, e.g.,
415 low-pass WGS data and 5hmC data. It has been reported that referred end
416 coordinates differ between HCC patients and healthy people at a whole-
417 genome scale (10). Apart from HCC, our results suggest the cfDNA preferred
418 end can also be used to distinguish patients with GBM, LUAD, and PDAC. We

419 herein discussed the distribution of cfDNA preferred ends in genomic features,
420 open chromatin regions, and nucleosome structure. The distribution of cfDNA
421 preferred ends can reflect nucleosome positioning, though there is no
422 apparent preference in genomic features and open chromatin regions,
423 consistent with previous work (15). However, informative cancer-specific
424 preferred ends were found to be significantly enriched in promoter and far
425 away from health-specific open chromatin regions (Figure 2e-f,
426 Supplementary Figure 4c-d, Supplementary Figure 5c-d).
427
428 In summary, our results indicate that cfDNA fragmentomic analysis with a
429 nonlinear classification algorithm using low-pass 5hmC sequencing data may
430 provide a simple, low-cost, and highly effective method for cancer detection.
431 The success of utilizing ultra-long fragments in 5hmC sequencing data as
432 biomarkers indicates the potential of other capture-based sequencing
433 approaches, such as cell-free methylated DNA immunoprecipitation and high-
434 throughput sequencing (cfMeDIP–seq) (33), in the diagnosis of cancer.
435 Notably, third-generation sequencing methods are expected to detect long
436 cfDNA molecules, which have been utilized in non-invasive prenatal testing
437 (34). As third-generation sequencing methods can detect sequence context
438 and epigenetic modification at the same time, they can provide simultaneous
439 analysis of genome-wide genetic, fragmentomic and epigenetic detection (35,
440 36). Moreover, standardized and effective workflows for analysis of cfDNA

441 fragments need to be developed.

442

443 **Data availability**

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

447

448 **Acknowledgments**

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

461 **Author Contributions**

462 Conception and design: Dan Xie, Zhidong Zhang, Xuenan Pi and Lin Xia.

463 Collection and assembly of data: Jun Zhang, Xiaoqin Yan, Shuxing Zhang,
464 Ailin Wei, Jingfeng Liu, Ang Li, Xiaolong Liu, Wei Zhang and Yanhui Liu. Data
465 analysis and interpretation: Zhidong Zhang, Xuenan Pi, Chang Gao, Xinlei Hu,
466 Xiyue Yan and Yuer Guo. Manuscript writing: Zhidong Zhang and Xuenan Pi.
467 All authors read and approved the final version of the manuscript.

468 **Competing interests**

469 The authors declare no competing interests.

470

471 **References**

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

485 end-motif frequencies in plasma DNA. *Proc. Natl. Acad. Sci. U. S. A.*
486 **116**, 641-49 (2019).
487 6. Han, D. S. C. , et al. The Biology of Cell-free DNA Fragmentation and
488 the Roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* **106**,
489 202-14 (2020).
490 7. Cristiano, S., et al. Genome-wide cell-free DNA fragmentation in
491 patients with cancer. *Nature.* **570**, 385-89 (2019).
492 8. Snyder, M. W., Kircher M., Hill A. J., Daza R. M. & Shendure J. Cell-free
493 DNA Comprises an In Vivo Nucleosome Footprint that Informs Its
494 Tissues-Of-Origin. *Cell.* **164**, 57-68 (2016).
495 9. Chan, K. C., et al. Second generation noninvasive fetal genome
496 analysis reveals de novo mutations, single-base parental inheritance,
497 and preferred DNA ends. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E8159-68
498 (2016).
499 10. Jiang, P., et al. Preferred end coordinates and somatic variants as
500 signatures of circulating tumor DNA associated with hepatocellular
501 carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E10925-33 (2018).
502 11. Jiang, P., et al. Plasma DNA End-Motif Profiling as a Fragmentomic
503 Marker in Cancer, Pregnancy, and Transplantation. *Cancer. Discov.* **10**,
504 664-73 (2020).
505 12. Mouliere, F., et al. Enhanced detection of circulating tumor DNA by
506 fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018).

507 13. Jiang, P., et al. Lengthening and shortening of plasma DNA in
508 hepatocellular carcinoma patients. *Proc. Natl. Acad. Sci. U. S. A.* **112**,
509 E1317-25 (2015).
510 14. Ulz, P., et al. Inferring expressed genes by whole-genome sequencing
511 of plasma DNA. *Nat. Genet.* **48**, 1273-8 (2016).
512 15. Sun, K., et al. Orientation-aware plasma cell-free DNA fragmentation
513 analysis in open chromatin regions informs tissue of origin. *Genome*
514 *Res.* **29**, 418-27 (2019).
515 16. He, B., et al. Tissue-specific 5-hydroxymethylcytosine landscape of the
516 human genome. *Nat. Commun.* **12**, 4249 (2021).
517 17. Chen, L., et al. Genome-scale profiling of circulating cell-free DNA
518 signatures for early detection of hepatocellular carcinoma in cirrhotic
519 patients. *Cell. Res.* **31**, 589-92 (2021).
520 18. [dataset]* Hu, X., et al. Integrated 5-hydroxymethylcytosine and
521 fragmentation signatures as enhanced biomarkers in lung cancer. *Clin.*
522 *Epigenetics.* **14**, 15 (2022).
523 19. Roadmap Epigenomics Consortium, et al. Integrative analysis of 111
524 reference human epigenomes. *Nature.* **518**, 317-30 (2015).
525 20. Gaffney, D. J., et al. Controls of nucleosome positioning in the human
526 genome. *PLoS. Genet.* **8**, e1003036 (2012).
527 21. Sanchez, C., et al. Circulating nuclear DNA structural features, origins,
528 and complete size profile revealed by fragmentomics. *JCI. Insight.* **6**,

529 e144561 (2021).
530 22. Zhang, Y., et al. Model-based analysis of ChIP-Seq (MACS). *Genome*
531 *Biol.* **9**, R137 (2008).
532 23. Lawn, S., et al. Neurotrophin signaling via TrkB and TrkC receptors
533 promotes the growth of brain tumor-initiating cells. *J. Biol. Chem.* **290**,
534 3814-24 (2015).
535 24. Nolte, I., Przibylla, H., Bostel, T., Groden, C. & Brockmann, M. A.
536 Tumor-platelet interactions: glioblastoma growth is accompanied by
537 increasing platelet counts. *Clin. Neurol. Neurosurg.* **110**, 339-42 (2008).
538 25. Moon, H. & Ro, S. W. MAPK/ERK Signaling Pathway in Hepatocellular
539 Carcinoma. *Cancers (Basel).* **13**, 3026 (2021).
540 26. Gnani, D., et al. Focal adhesion kinase depletion reduces human
541 hepatocellular carcinoma growth by repressing enhancer of zeste
542 homolog 2. *Cell. Death. Differ.* **24**, 889-902 (2017).
543 27. Li, J., et al. Knockdown of FOXO3a induces epithelial-mesenchymal
544 transition and promotes metastasis of pancreatic ductal
545 adenocarcinoma by activation of the β-catenin/TCF4 pathway through
546 SPRY2. *J. Exp. Clin. Cancer. Res.* **38**, 38 (2019).
547 28. Song, C. X., et al. 5-Hydroxymethylcytosine signatures in cell-free DNA
548 provide information about tumor types and stages. *Cell. Res.* **27**, 1231-
549 42 (2017).
550 29. Li, W., et al. 5-Hydroxymethylcytosine signatures in circulating cell-free

551 DNA as diagnostic biomarkers for human cancers. *Cell. Res.* **27**, 1243-
552 57 (2017).
553 30. Guler, G. D., et al. Detection of early stage pancreatic cancer using 5-
554 hydroxymethylcytosine signatures in circulating cell free DNA. *Nat.*
555 *Commun.* **11**, 5270 (2020).
556 31. Lam, W. K. J., et al. Sequencing-based counting and size profiling of
557 plasma Epstein-Barr virus DNA enhance population screening of
558 nasopharyngeal carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5115-
559 24 (2018).
560 32. Lo, Y. M., et al. Maternal plasma DNA sequencing reveals the genome-
561 wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**,
562 61ra91 (2010).
563 33. Shen, S. Y., et al. Sensitive tumour detection and classification using
564 plasma cell-free DNA methylomes. *Nature.* **563**, 579-83 (2018).
565 34. Yu, S. C. Y., et al. Single-molecule sequencing reveals a large
566 population of long cell-free DNA molecules in maternal plasma. *Proc.*
567 *Natl. Acad. Sci. U. S. A.* **118**, e2114937118 (2021).
568 35. Tse, O. Y. O., et al. Genome-wide detection of cytosine methylation by
569 single molecule real-time sequencing. *Proc. Natl. Acad. Sci. U. S. A.*
570 **118**, e2019768118 (2021).
571 36. Liu, Q., et al. Detection of DNA base modifications by deep recurrent
572 neural network on Oxford Nanopore sequencing data. *Nat. Commun.* **10**,

573 2449 (2019).

574 37. [dataset]* Cao, F., et al. Integrated epigenetic biomarkers in circulating
575 cell-free DNA as a robust classifier for pancreatic cancer. *Clin.*
576 *Epigenetics.* **12**, 112 (2020).

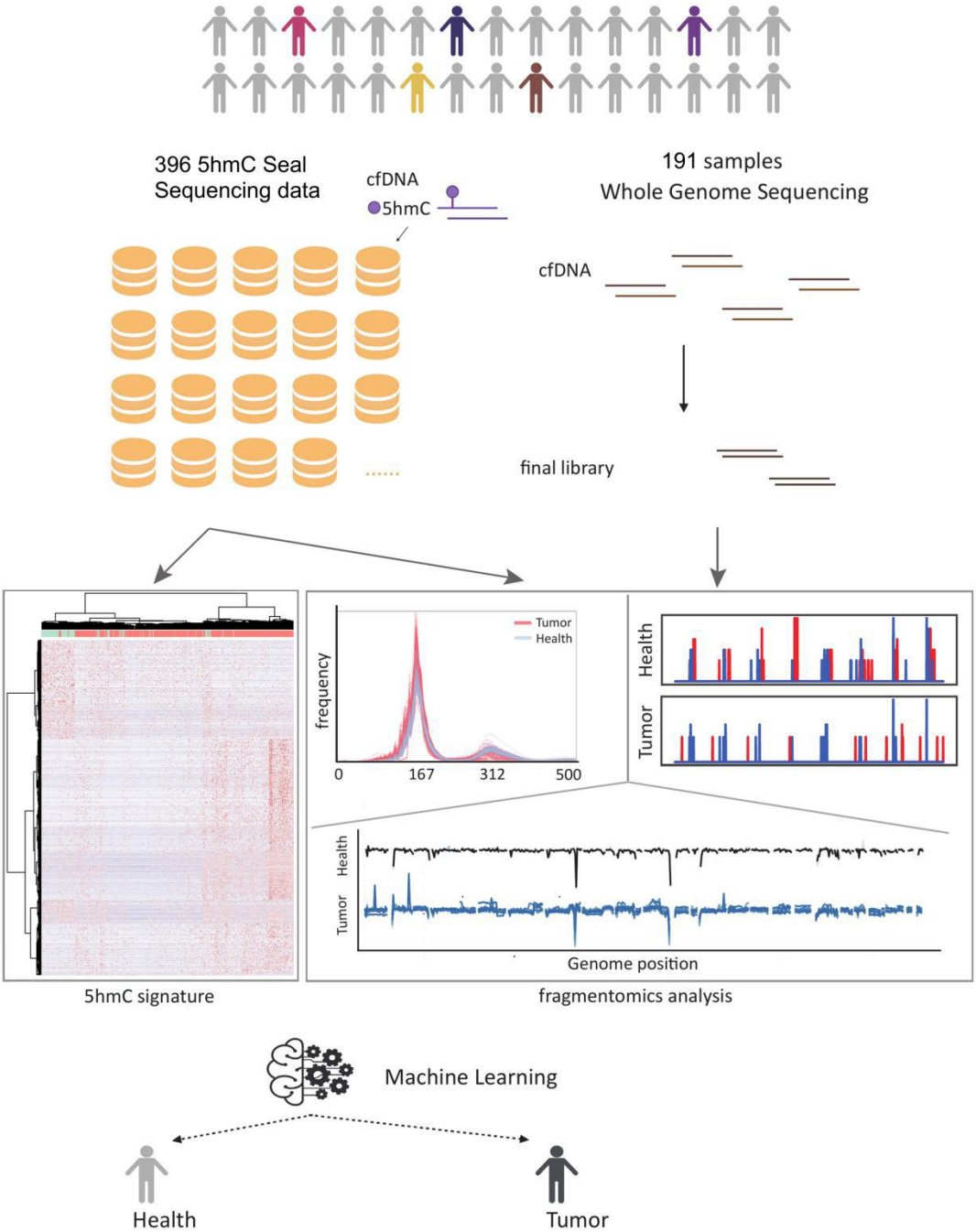
577 38. [dataset]* Cai, Z., et al. Liquid biopsy by combining 5-
578 hydroxymethylcytosine signatures of plasma cell-free DNA and protein
579 biomarkers for diagnosis and prognosis of hepatocellular carcinoma.
580 *ESMO Open.* **6**, 100021 (2021).

581
582
583
584
585
586
587
588
589
590
591
592
593
594

595

596

597

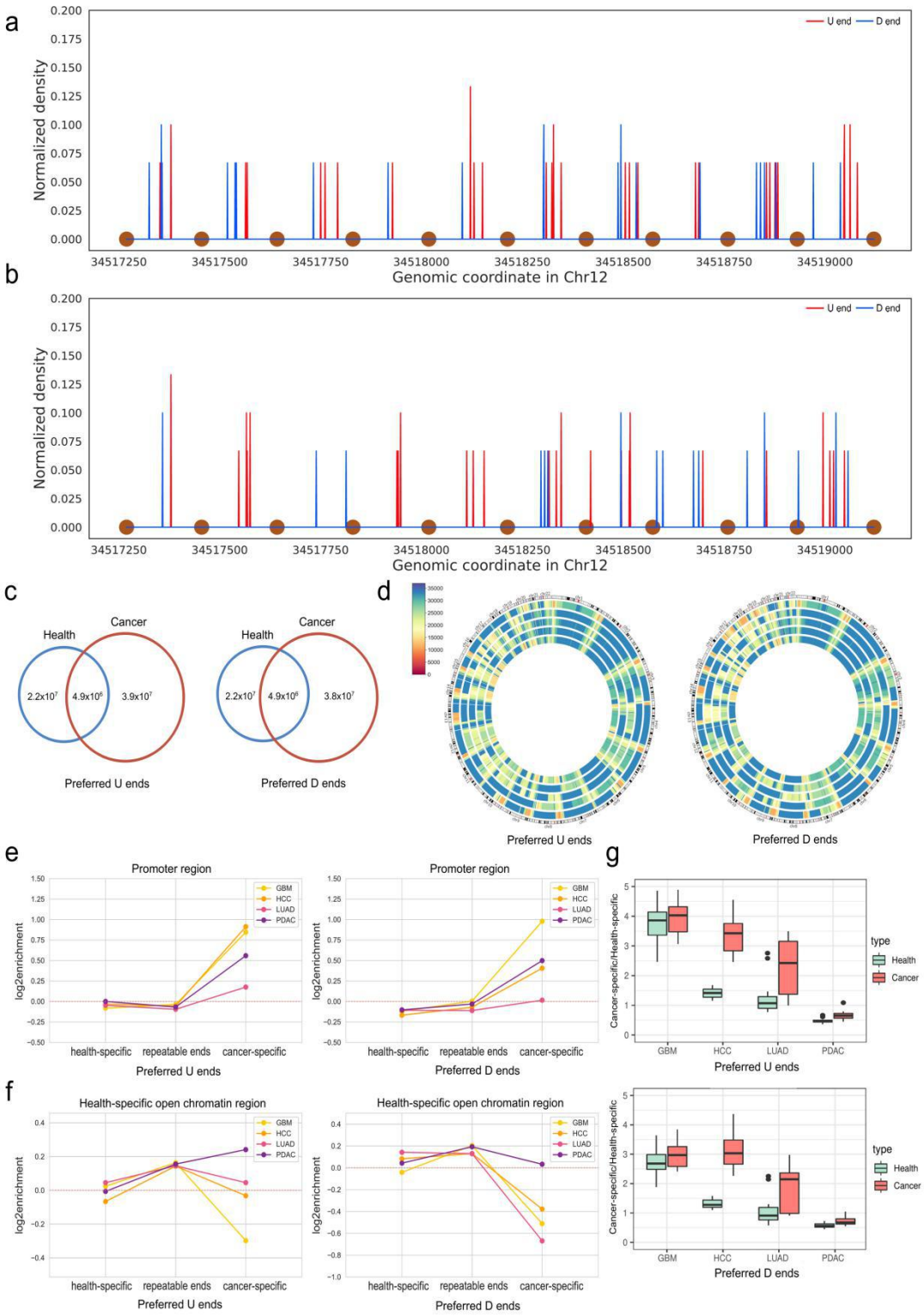


598

599 **Figure 1 Overview of the study design.**

600 cfDNA from cancer patients and controls were sequenced with WGS.

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



605

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

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

608 34519122) in healthy (a) and HCC (b) subjects. Brown dots at the bottom

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

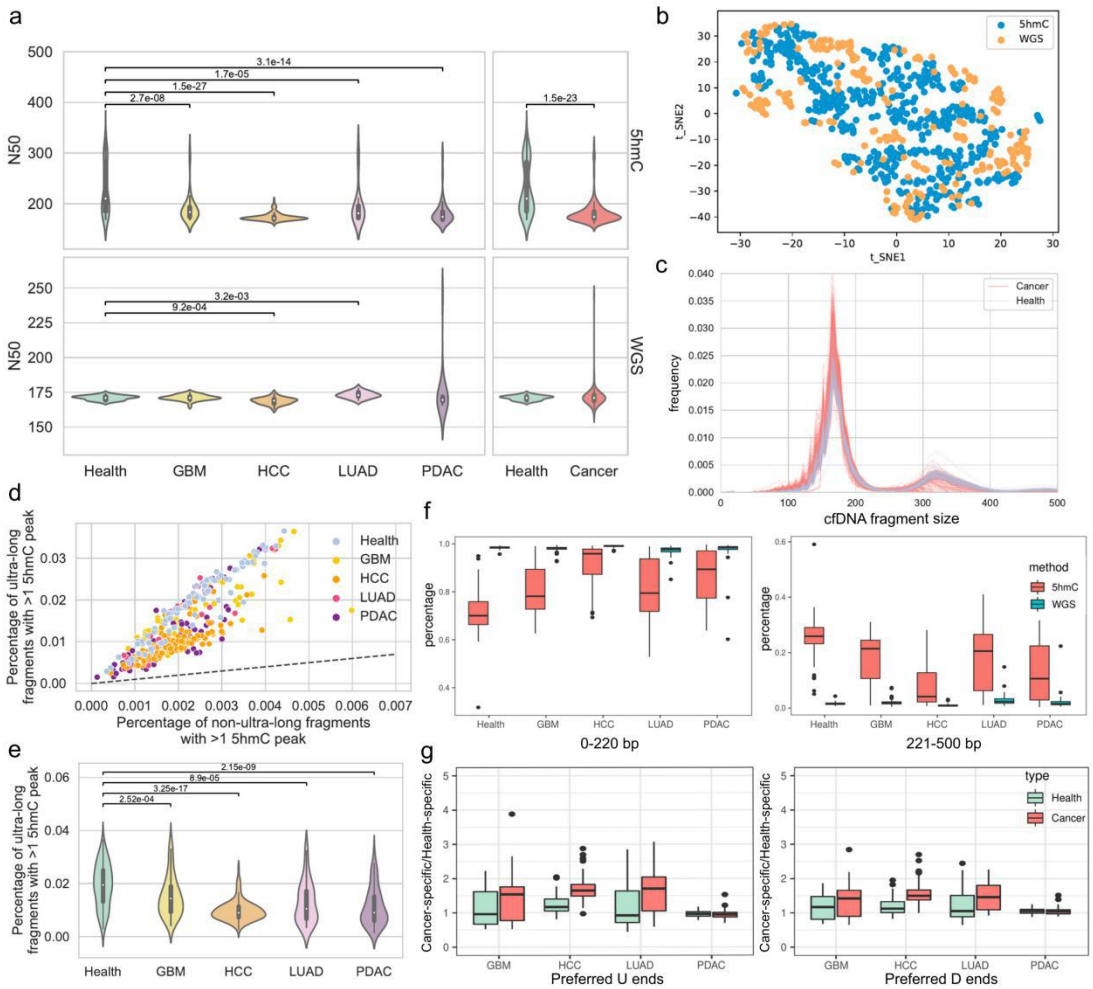


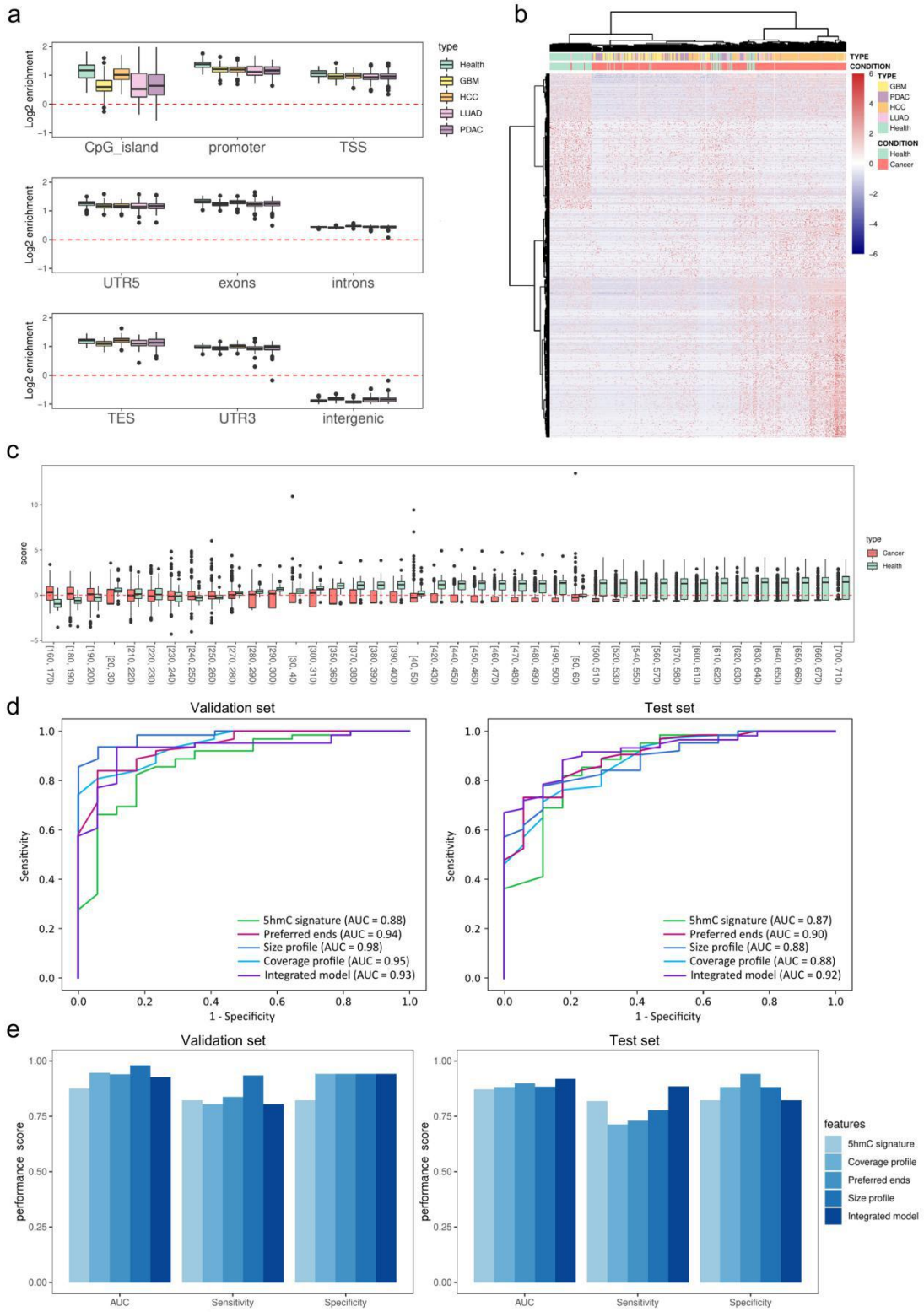
Figure 3 cfDNA fragmentomic information in 5hmC sequencing data. (a) Violin plots showing fragment size profile comparison between cancer samples and controls in terms of N50 (Mann–Whitney test). (b) tSNE plot showing 5hmC sequencing data and WGS data by fragments below 220 bp. (c) The cfDNA size profile 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

630 ranging from 0-220 bp and 221-500 bp among fragments ranging from 0-800
631 bp in WGS data and 5hmC sequencing data (Wilcoxon rank-sum test). (g)
632 Boxplots showing the ratios of the cancer-specific to health-specific preferred
633 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)

Coverage profile correlation to median healthy in control and four types of cancer in short, long, and ultra-long cfNDA fragments (Wilcoxon rank-sum test, p value<0.005).



643 **Figure 5 Performance of cancer detection combining fragmentomic**
644 **features and 5hmC signatures using 5hmC sequencing data.**
645 (a) Enrichment of 5hmC in genomic features in control and four types of
646 cancer (Wilcoxon rank-sum test, p value<0.001). (b) Heatmap of differentially
647 modified 5hmC peaks. (c) Selected size profile features in the size profile
648 model. (d) ROC curves for the validation set and test set in pan-cancer
649 diagnosis. (e) Performance evaluation of validation set and test set in pan-
650 cancer diagnosis.
651
652