De novo characterization of cell-free DNA fragmentation hotspots boosts the power for early detection and localization of multi-cancer

Xionghui Zhou[1], Yaping Liu[1-4 *]

1 Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229

2 Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229

3 Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229

4 Department of Electrical Engineering and Computing Sciences, University of Cincinnati College of Engineering and Applied Science, Cincinnati, OH 45229

* Email: lyping1986@gmail.com

Abstract

The global variation of cell-free DNA fragmentation patterns is a promising biomarker for cancer diagnosis. However, the characterization of its hotspots and aberrations in early-stage cancer at the fine-scale is still poorly understood. Here, we developed an approach to *de novo* characterize genome-wide cell-free DNA fragmentation hotspots by integrating both fragment coverage and size from whole-genome sequencing. These hotspots are highly enriched in regulatory elements, such as promoters, and hematopoietic-specific enhancers. Surprisingly, half of the high-confident hotspots are still largely protected by the nucleosome and located near repeats, named inaccessible hotspots, which suggests the unknown origin of cell-free DNA fragmentation. In early-stage cancer, we observed the increases of fragmentation level at these inaccessible hotspots from microsatellite repeats and the decreases of fragmentation level at accessible hotspots near promoter regions, mostly with the silenced biological processes from peripheral immune cells and enriched in CTCF insulators. We identified the fragmentation hotspots from 298 cancer samples across 8 different cancer types (92% in stage I to III), 103 benign samples, and 247 healthy samples. The fine-scale fragmentation level at most variable hotspots showed cancer-specific fragmentation patterns across multiple cancer types and non-cancer controls. Therefore, with the fine-scale fragmentation signals alone in a machine learning model, we achieved 42% to 93% sensitivity at 100% specificity in different early-stage cancer. In cancer positive cases, we further localized cancer to a small number of anatomic sites with a median of 85% accuracy. The results here highlight the significance to characterize the fine-scale cell-free DNA fragmentation hotspot as a novel molecular marker for the screening of early-stage cancer that requires both high sensitivity and ultra-high specificity.

Keywords

Cell-free DNA, Fragmentation hotspot, Whole-genome sequencing, Cancer screening, Cancer early detection, **C**ell f**R**ee dn**A** fra**G**mentation, CRAG

Introduction

Routine screening in the average-risk population provides the opportunity to detect cancer early and greatly reduce morbidity and mortality in cancer [1,2]. Early identification of lethal cancer versus non-lethal disease will minimize and better manage overdiagnosis [3]. Both medical needs require non-invasive biomarkers with ultra-high specificity (>99%) and high sensitivity at the same time, which, however, are still not available in the clinic.

Recent advances in circulating cell-free DNA (cfDNA) suggested a promising non-invasive approach for cancer diagnosis by using the tumor-specific genetic and epigenetic alterations, such as mutations, copy number variations, and DNA methylation [4–9]. The number of tumor-specific alterations identified by whole-genome sequencing (WGS), however, is small in patients with early-stage cancer. Moreover, Clonal Hematopoiesis of Indeterminate Potential (CHIP)-associated genetic variations are discovered in cfDNA, which are not specific to cancer but a normal phenotype of aging[10]. Sequence degradation caused by sodium bisulfite treatment for the measurement of DNA methylation will reduce the sensitivity of the test. These limitations bring the challenges of using genetic and epigenetic variations for the early diagnosis.

In contrast to the limited number of genetic alterations, the number of cfDNA fragments is large in the circulation. The cfDNA fragmentation patterns, such as the fragment coverages and sizes, are altered in cancer and their aberrations are not associated with CHIP [11,12]. Their derived patterns, such as nucleosome positions, patterns near transcription start sites, ended position of cfDNA, and large-scale fragmentation changes at mega-base level, offer extensive signals from the tumor, as well as possible alterations from immune cell deaths [11,13–15], which can significantly increase the sensitivity for cancer early detection. The *in vivo* cfDNA fragmentation process during the cell death, however, is complicated and still not well understood [16]. Moreover, sequence complexity, such as G+C% content and mappability, will often cause the variations of fragment coverages and sizes [17,18]. These challenges, especially the unknown sources of fragmentation variations, will bring the potential noises for cancer screening, which partially requires the ultra-high specificity.

To conquer these challenges, a genome-wide computational approach is needed to narrow down the regions of interest with large signal-to-noise ratios across different pathological conditions. Recent studies suggested that local nucleosome structure reduces the fragmentation process, which also indicated the potential existence of cfDNA fragmentation hotspots (lower coverage and smaller size) at the open chromatin regions [13,19]. Nucleosome positioning is still not well characterized at a variety of cell types across different pathological conditions, while the variations of DNA accessibility at open chromatin regions have been comprehensively profiled at many primary cell types across different physiological conditions, including cancer and immune cells with nuanced activated and rest status [20,21]. Therefore, instead of identifying “fragmentation coldspots” at nucleosome protected regions, we hypothesize that the characterization of fine-scale cfDNA fragmentation hotspots will not

99 only help us to better elucidate the unknown molecular mechanism of fragmentation patterns
100 but also boost the power of cfDNA fragmentation for the identification of nuanced
101 pathological conditions, such as early-stage cancer.
102 Here, by integrating fragment coverage and size information, we developed a probabilistic
103 based approach, named **C**ell f**R**ee dn**A** fra**G**mentation (CRAG), to *de novo* identify the
104 cfDNA fragmentation hotspots genome-wide from cfDNA paired-end WGS data. We
105 characterized the genome-wide distribution and novel unknown origin of these fragmentation
106 hotspots. Finally, our results suggested that fragmentation signals alone from these regions
107 can significantly boost the performance of screening and localization of early-stage cancer
108 with high sensitivity at ultra-high specificity.
109

110 ## Results

111 **CRAG: a probabilistic model to characterize the cell-free DNA fragmentation hotspots.**
112 To integrate fragment coverage and size information, we developed an integrated
113 fragmentation score (IFS) by weighting the coverage of each fragment based on the
114 probability of its size in the overall distribution of the fragment size across the whole
115 chromosome (Details in Methods). Since cfDNA in healthy individuals are mostly from
116 hematopoietic cells, to understand the relationship between IFS and open chromatin
117 regions, we generated the common open chromatin regions by the DNase-seq peaks shared
118 across the major hematopoietic cell types in peripheral blood [22]. The distribution of IFS
119 showed the expected depletion around the center of common open chromatin regions
120 (Supplementary Figure 1a). We also compared the distribution with window protection score
121 (WPS) and orientation-aware cfDNA fragmentation (OCF) [13,23], which were developed
122 previously to integrate the coverage and size information but for different purposes
123 (Supplementary Figure 1b-c). WPS around the common open chromatin regions showed a
124 higher signal than that from the global background similar to that from nucleosome protected
125 regions, which will bring the challenges to identify the fragmentation hotspots.
126 To *de novo* characterize the fragmentation hotspot, we utilized a 200bp sliding window
127 approach to scan the whole genome and built a background Poisson model for the local (+/-
128 5kb, +/-10kb) and global (whole chromosome) genomic regions (Details in Methods). After
129 filtering the low mappability regions, we further calculated the significance level in each
130 window and performed the multiple hypotheses correction. At the BH01 dataset, we
131 identified 277,109 cfDNA fragmentation hotspots. The whole CRAG workflow is shown in
132 Figure 1 and Supplementary Figure 2a. The IFS distribution, as well as the fragment
133 coverage and size distribution, showed the expected depletion at the center of hotspots,
134 which suggested that the model correctly captured the fragmentation hotspots
135 (Supplementary Figure 2b). Since traditional double-stranded WGS cannot fully capture the
136 small single-stranded fragments in open chromatin regions, we also plotted the distribution
137 of fragment size from a single-stranded cfDNA WGS dataset in a healthy individual (IH01

138 from [13]). The patterns of fragment length around the hotspots are still similar (Supplementary
139 Figure 2b).
140
141 **Cell-free DNA fragmentation hotspots are enriched in regulatory elements.**
142 We next sought to understand the genomic distribution of these *de novo* fragmentation
143 hotspots. The fragmentation hotspots covered the majority of the CpG island (CGI)
144 promoters and CTCF motifs, but showed almost no enrichment in non-CGI promoters, 5'
145 exon boundary, transcription terminated sites (TTS), and random genomic regions (Figure
146 2). DNase accessible signals from the major hematopoietic cell types in peripheral blood are
147 enriched in hotspots as expected (Figure 3a,d, Supplementary Figure 3). We also observed
148 the enrichment of active histone marks, such as H3K4me3 and H3K27ac, and the depletion
149 of repressive or gene body relevant histone marks, such as H3K27me3, H3K9me3, and
150 H3K36me3. The enhancer mark H3K4me1 from hematopoietic cell types but not other cell
151 types showed enrichment in hotspots (Figure 3b,d, Supplementary Figure 3). To further
152 understand the distribution of fragmentation hotspot in different chromatin states, we utilized
153 the 15-states chromHMM segmentation result across different cell types from NIH
154 Epigenome Roadmap Consortium (Figure 3c,d, Supplementary Figure 3) [22]. Similar to
155 histone modification results, the hotspots showed depletion in non-regulatory regions but
156 highest enrichment at active promoters (TssA and TssFlnk) and bivalent regions (TssBiv,
157 BivFlnk, and EnhBiv) across almost all the cell types. While in cell-type-specific chromatin
158 states, such as enhancer (Enh), the hotspots only showed the highest enrichment in
159 hematopoietic cell types and liver that often release cfDNA to the blood in healthy
160 individuals. The evolutionary conservation score (phastCons) in hotspots is also significantly
161 higher than matched random regions (two-sided Mann–Whitney U test, p value < $2.2e^{-16}$,
162 Supplementary Figure 4) [24].
163 To understand whether or not IFS in our model is the optimized way to integrate the
164 fragment coverage and size information, we performed similar hotspot calling and chromatin
165 states enrichment analysis by using WPS or OCF signals from the same cfDNA WGS
166 dataset, IFS signals but with the Kolmogorov–Smirnov test (K-S test), and IFS signals in
167 genomic DNA (gDNA) WGS obtained from 1000 genome project with merged similar
168 sequencing depth (Supplementary Figure 5) [25]. As expected, hotspots calling by WPS or
169 OCF from cfDNA did not show enrichment in promoter regions and depletion in all the non-
170 regulatory regions. Hotspot called by IFS from gDNA showed the enrichment in promoters
171 across all cell types due to the coverage dropout at high G+C% content genomic regions
172 near active promoters. However, the cell-type-specific enrichment of gDNA hotspots in
173 enhancer states from hematopoietic cell types is not observed, which further suggested that
174 the cell-type-specific enrichment of regulatory regions in hotspot from cfDNA WGS is not
175 simply due to the technical bias, such as G+C% content in next-generation sequencing.
176 We further explored the contributions of other possible technical factors, such as sequencing
177 depth, on the performance of our hotspot calling. To quantify the performance of CRAG and

178 the minimum requirement of the sequencing depth, we created the balanced fragmentation-
179 positive and fragmentation-negative regions that are randomly sampled from conserved
180 activated chromatin states (TssA) and silenced chromatin states (Quies) (Detailed in
181 Methods). We downsampled the effective fragments in BH01 from 1.2 billion to 10 million.
182 The performance is saturated at ~0.9 F1-score with 400 million fragments. Even with 200
183 million fragments, we can still achieve reasonable performance (~0.8 F1-score) together with
184 hematopoietic cell type specific enrichment in enhancer states (Supplementary Figure 6).
185 Taken together, CRAG can *de novo* accurately identify the cfDNA fragmentation hotspots
186 that are enriched in regulatory elements.
187
188 **Half of the cell-free DNA fragmentation hotspots are still protected by the local**
189 **nucleosomes.**
190 Previous studies indirectly suggested that cfDNA fragmentation hotspots are mostly co-
191 localized with open chromatin regions from hematopoietic cell types in healthy individuals [13].
192 To test this hypothesis, we collected 523 public available open chromatin region datasets
193 measured by DNase-seq or ATAC-seq across the liver and different hematopoietic cell types
194 at rest or active status from NIH Epigenome Roadmap, ENCODE, BLUEPRINT, and other
195 publications [21,22,26–28]. To characterize the saturation point of the overlap with collected open
196 chromatin regions, we subsampled the number of datasets and measured their overlaps with
197 the fragmentation hotspots. Surprisingly, in the saturation point, half of the common
198 fragmentation hotspots are not overlapped with existing open chromatin regions (Figure 4a,
199 Supplementary Figure 7).
200 We next asked whether or not the lack of overlapping between open chromatin regions and
201 fragmentation hotspots is due to the technical artifact in DNase-seq/ATAC-seq, such as the
202 threshold choice of peak calling. In such a case, we should still be able to observe the
203 residual accessibility level at the inaccessible hotspots that are not overlapped with any
204 existing open chromatin regions. However, the normalized accessibility signals from the
205 major hematopoietic cell types did not show any enrichment at the center of these
206 inaccessible hotspots (Figure 4b). Even by using the GpC methyltransferase accessibility
207 measured from NOMe-seq that are not confounded by the coverage variations, the
208 accessibility still showed significant differences between accessible and inaccessible
209 hotspots (Supplementary Figure 8). Moreover, the neutrophil is one of the major cell types to
210 contribute 40-60% cfDNA fragments. The direct measurement of DNA accessibility in
211 neutrophil, however, is still not available. To mitigate this concern, we plotted the histone
212 modification data from neutrophils around these inaccessible hotspots. The results did not
213 suggest any potential open chromatin regions (Supplementary Figure 9). To avoid the
214 possible fragmentation bias caused by k-mer differences, we normalized the IFS signal by k-
215 mer composition (n=2) and still found similar fragmentation patterns before and after the
216 correction (Supplementary Figure 10).
217 The distribution of cell-free DNA methylation levels from healthy individuals at the same

218 locations [29] suggested the intrinsic functional differences between accessible and
219 inaccessible hotspots (Figure 4c). We further performed the motif enrichment analysis. The
220 results showed that inaccessible hotspots are highly enriched with pioneer transcription
221 factors, such as OCT (POU, Homeobox), which usually bind the nucleosome occupied
222 regions (Figure 4d) [30], while accessible hotspots showed the highest enrichment in CTCF as
223 expected.
224 These results suggested that the existence of the inaccessible hotspot is not due to the
225 known technical artifact and shows the potential regulatory function.
226 We next sought to answer the question of where these inaccessible hotspots located in the
227 genome. Interestingly, 96% of the inaccessible hotspots are overlapped with the repeats,
228 which is significantly higher than that of random regions with matched chromosome and
229 length (71% overlapped, Hypergeometric test, $p<2.2e^{-16}$) (Figure 4e). Even by using the
230 most significant inaccessible hotspots (top 1% by p-value), the results are still largely the
231 same (Supplementary Figure 11). We further analyzed the enrichment of these inaccessible
232 hotspots at different types of repeat classes [31]. Highest enrichment is observed at the 5' end
233 of microsatellite repeats (i.e., simple repeats) and low complexity regions (Figure 4e,
234 Supplementary Figure 12). Interestingly, we also noticed the high enrichment of inaccessible
235 hotspots right after the 3' end of transposable elements, which are not due to the mappability
236 and G+C% bias (Supplementary Figure 13a-e). Moreover, the differences of DNA
237 methylation at 3'end of Alu elements with or without overlapping the inaccessible hotspots
238 indicated the potential functional association between fragmentation hotspots and
239 transposable elements (Supplementary Figure 13f).
240
241 **The variations of fine-scale fragmentation signals from the cell-free DNA**
242 **fragmentation hotspots boost the detection power for early-stage cancer.**
243 Further, we would like to characterize the aberrations of fine-scale fragmentation level at the
244 cfDNA fragmentation hotspots in early-stage cancer. To obtain enough fragment number for
245 the hotspot calling as that estimated above (Supplementary Figure 6b), we pooled the
246 publicly available low-coverage cfDNA WGS (~1X/sample) from 90 patients with early-stage
247 hepatocellular carcinoma (HCC, 85 of them are Barcelona Clinic Liver Cancer stage A, 5 of
248 them are stage B) and 32 healthy individuals from the same study, respectively (Details in
249 Methods)[32]. The volcano plot of the p-value (two-sample t-test) and z-score difference of IFS
250 between HCC and healthy across all the fragmentation hotspots showed more fractions of
251 hypo-fragmented hotspots in early-stage HCC (Figure 5a). Further, the unsupervised
252 hierarchical clustering of the top 10,000 most variable hotspots showed a clear
253 fragmentation dynamic between HCC and healthy (Figure 5b, Supplementary Figure 14-15).
254 Therefore, we utilized the IFS from the cfDNA fragmentation hotspots to classify the HCC
255 and healthy individuals by a linear Support Vector Machine (SVM) approach (Details in
256 Methods). By 10-fold cross-validation, we observed a much higher classification
257 performance (93% sensitivity at 100% specificity) than that by using copy number variations

258 (CNVs) with the same machine learning infrastructure and same data split (44% sensitivity at
259 100% specificity) and the previous publication by mitochondria DNA (mtDNA) [32] (53%
260 sensitivity at 100% specificity) (Figure 5c, Supplementary Table 2-3, Supplementary Figure
261 16).
262 We next asked why the IFS signal in fragmentation hotspots can boost the classification
263 performance. Based on the z-score differences between HCC and healthy as well as the
264 feature importance to the classification of the model, we split the hotspots that are
265 significantly contributed to the classification into two groups: Class I (Hypo-fragmented in
266 cancer) and Class II (Hyper-fragmented in cancer) (Figure 5d, Supplementary Table 4). The
267 Class I hotspots are mostly accessible hotspots and significantly enriched in promoter
268 regions, which suggested the potential silencing of genes with the decrease of
269 fragmentation. While Class II hotspots are mostly inaccessible hotspots and overlapped with
270 microsatellite repeats, which suggested the potential increases of fragmentation at
271 microsatellite repeats in early-stage cancer (Figure 5e). We further checked the enrichment
272 of motif and gene ontology biological processes at these two groups of hotspots. The results
273 further suggested the differences of motif enrichment between two groups and the potential
274 silencing of biological processes for the activation of peripheral immune cells that contributed
275 most to the cfDNA composition, such as neutrophil and myeloid cells (Figure 5f-g,
276 Supplementary Table 5).
277 Overdiagnosis is one of the major concerns for the screening of early-stage cancer. We next
278 sought to explore whether or not the IFS signal from cfDNA fragmentation hotspot could also
279 characterize the differences between early-stage HCC and non-malignant liver diseases. We
280 identified the hotspots on the additional cfDNA WGS datasets from 67 patients with chronic
281 HBV infection and 36 patients with HBV-associated liver cirrhosis in the same study [32]. PCA
282 analysis of IFS signals across all the hotspots suggested a clear separation between early-
283 stage HCC and non-malignant liver diseases, as well as healthy controls (Supplementary
284 Figure 17a). To exclude the possible batch effect, we also performed PCA on IFS from
285 matched random genomic regions in the same sample but did not observe a clear
286 separation between groups of samples (Supplementary Figure 17b). Another possible
287 artifact is our pooling strategy for the hotspot calling in low-coverage WGS data. The hotspot
288 calling on the pooled group may enrich the regions with similar depletions in the genome
289 without any biological meaningful indication. To avoid such an artifact, we randomly grouped
290 the samples and called the hotspots from these random groups with the matched group
291 sizes. The PCA results did not show any separations (Supplementary Figure 17c). We
292 further selected the top 30,000 most variable peaks, performed the unsupervised
293 hierarchical clustering, and observed the clear dynamics of the fragmentation patterns
294 among early-stage HCC, HBV, Cirrhosis, and healthy controls (Supplementary Figure 18-
295 19). Finally, by 10-fold cross validation, the linear SVM model showed a higher classification
296 performance (83% sensitivity at 100% specificity) than other methods, such as CNVs and
297 mtDNA (Supplementary Figure 20, Supplementary Table 6-7).

298

299 **Screening and localization of multiple early-stage cancer types by the cell-free DNA**
300 **fragmentation hotspots.**

301 One of the biggest challenges for the screening of early-stage cancer in the average-risk
302 population is to obtain high sensitivity at ultra-high specificity (>99%) across different cancer
303 types, which is not available in clinics yet. To further validate our method in a more
304 comprehensive early-stage cancer dataset, we collected publicly available low-coverage
305 cfDNA WGS data (~1X/sample) from 208 patients across 7 different kinds of cancer (88% in
306 stage I-III, colon, breast, lung, gastric, bile duct, ovary, and pancreatic cancer) and matched
307 215 healthy controls in the same study [11]. We applied a similar strategy as that in the HCC
308 study above for the hotspots calling. Across 7 different types of cancer and healthy
309 conditions, the z-score of IFS signals in the most variable fragmentation hotspots showed
310 clear cancer-specific fragmentation patterns in both t-SNE visualization and unsupervised
311 hierarchical clustering (Figure 6a-b, Supplementary Figure 21, Details in Methods). The
312 fragmentation patterns alone at these hotspots can separate the cancer types very well. By
313 10-fold cross validation, the linear SVM model showed a consistent high classification
314 performance across different stages for its high sensitivity at ultra-high specificity (64%
315 sensitivity to 82% sensitivity at 100% specificity), which is significantly and consistently
316 higher in different stages than previous reported results by using large-scale fragmentation
317 pattern, copy number variations, and mtDNA from the same dataset [11] (Figure 6c, Table 1).
318 For example, at 100% specificity, we achieved 93% sensitivity (95% CI: 85%-100%) in
319 gastric cancer, 88% sensitivity (95% CI: 76%-100%) in colorectal cancer and 81%
320 sensitivity (95% CI: 76%-91%) in breast cancer, which of these are poorly detected at ultra-
321 high specificity level by other liquid biopsy studies [5,8,9,11,33] . (Supplementary Figure 22-23,
322 Table 1, Supplementary Table 8). In the other types of cancer, the performance is largely
323 comparable to previous results [11]. We also tested the performance before GC bias
324 correction, the results are largely the same (Supplementary Figure 22). In addition, by the
325 same machine learning framework, the results of fragmentation hotspots are consistently
326 better (Supplementary Figure 22).

327 Another big challenge for the screening of early-stage cancer is to identify the cancer types
328 for the choices of most appropriate follow-up diagnosis and treatment. Here, we asked
329 whether or not we can identify the tissues-of-origin of cancer samples by using the fine-scale
330 fragmentation level alone. In the cancer positive samples identified above by machine
331 learning algorithm, without any clinical information about the patients, we further localized
332 the sources of cancer to one or two anatomic sites in a median of 85% of these patients
333 across five different cancer types, and 82.5% accuracy across six different cancer types.
334 Furthermore, we were able to localize the source of the positive test to a single organ in a
335 median of 65% of these patients across five different cancer types, and 56% accuracy
336 across six different cancer types, which is similar to previous reports using the combination
337 of mutation and protein [5] or DNA methylation [8], but consistently higher than that by using

338 large-scale fragmentation patterns from the same dataset [11] (Fig. 6d, Supplementary Table
339 9, Supplementary Figure 24) (Details in Methods). The prediction accuracy is varied with
340 tumor type, from 70% (95%CI: 44%-96%) in ovarian cancer to 98% (95%CI: 94%-100%) in
341 breast cancer (Fig. 6d and Supplementary Table 9), but significantly higher than random
342 choices by the sample frequency in each cancer type (Supplementary Figure 25).
343

344 Discussions

345 In summary, we have developed a computational approach, named CRAG, to *de novo*
346 identify the cfDNA fragmentation hotspots by integrating both fragment size and coverage
347 information. The fragmentation patterns of cfDNA have long been associated with local
348 nucleosome protection, which indicates the high overlap between fragmentation hotspots
349 and open chromatin regions. Extreme G+C% bias and low mappability often leads to the
350 drop out of fragment coverages and sometimes fragment sizes [17,18]. Therefore, genomic
351 regions with a higher fragmentation rate do not always mean open chromatin regions. After
352 excluding these potential technical bias, our genome-wide analysis here found that half of
353 the fragmentation hotspot regions are still protected by nucleosomes and mostly located at
354 repeats, especially microsatellite and 3'end of potential functional transposable elements,
355 which suggested the unknown origin of the cfDNA fragmentation process. At these
356 inaccessible hotspots, we found the significant enrichment of pioneer transcription factor
357 binding sites and increases of fragmentation level at the microsatellite repeats in early stage-
358 cancer, which further indicated its functional importance for the classification of early-stage
359 cancer. More importantly, the silencing of accessible hotspots is highly enriched at
360 promoters and the biological processes for the activation of peripheral immune cells. A lot of
361 recent efforts on cancer early detection, however, focused on how to enrich the circulating
362 tumor DNA signals from tumor cells [5,9], which ignored the critical role of peripheral immune
363 aberrations on cancer initiations [34]. In addition, the CTCF motif is highly enriched at the
364 silenced accessible hotspots, which indicated the potential three-dimensional chromatin
365 organization changes in the initiation of early-stage cancer [35]. Taken together, our results
366 suggested that the fine-scale cfDNA fragmentation hotspot should be treated as a novel
367 molecular marker rather than the representation of open chromatin regions.
368 Previous efforts had been made to characterize the nucleosome-free regions by using the
369 depletion of coverages from MNase-seq/ChIP-seq assay [36]. The measurement of cfDNA
370 fragmentation here, however, involves information from both fragment coverages and sizes.
371 CRAG can be further improved by better integration of the fragment coverages and sizes, or
372 even with more dimensions, such as the fragment orientation and endpoint. In addition,
373 G+C% bias is known to affect the peak calling result in ChIP-seq/ATAC-seq [37]. A better
374 statistical model with the incorporation of GC normalization on both of the fragment coverage
375 and sizes will potentially improve the performance of our method. PCR-free library
376 preparation for WGS will also mitigate the concerns of GC bias and other sequencing

377 artifacts [38].
378 For the detection and localization of early-stage cancer, there are still several limitations in
379 our study. First, due to the limited availability of public cfDNA WGS datasets from early-
380 stage cancer patients, the classification performance here is evaluated by multi-fold cross-
381 validation on a relatively small sample-size cohort in each cancer type, similar as other
382 cfDNA WGS studies [11]. Multiple independent large-scale prospective cohorts with similar
383 cancer types will be a better way to assess the power of our approach for the diagnosis of
384 early-stage cancer. Second, we pooled the low-coverage WGS samples from the same
385 condition for the hotspot calling. If the number of samples is small in a condition, due to the
386 random drop out of the fragment coverage and a large number of genomic windows in the
387 genome, the number of false discovered hotspots without any biological interpretations will
388 increase. Although our current strategy by filtering low mappability regions and correcting
389 GC bias is helpful, imputation methods or more stringent statistical models are still needed to
390 mitigate the potential missing data artifact. Third, the proportion of cancer types as well as
391 the ratio between cancer and healthy is not an unbiased representation of the average-risk
392 population in the US. The sensitivity and specificity here may not represent the true
393 performance in the large-scale screening. Fourth, the proof-of-concept study on HCC here
394 suggested the distinguished cfDNA fragmentation patterns between early-stage cancer and
395 non-malignant liver disease controls. More cfDNA studies on non-malignant cancer,
396 diseases and benign status are needed to minimize the overdiagnosis in the population-level
397 screening. Lastly, in some cancer types, our fine-scale study here showed complementary
398 classification performance compared with that in the previous large-scale fragmentation
399 study at the same dataset [11]. For example, our results on gastric, breast and colorectal
400 cancer outperformed previous large-scale fragmentation studies, while at bile duct and lung
401 cancer the result is reversed. Future combinations of the fragmentation patterns at multi-
402 scales, as well as information from other modalities or clinical meta-data, may further
403 improve the performance.
404 Our study here not only paves the road to further elucidate the molecular mechanism of
405 cfDNA fragmentation but also lays the practical foundation of a highly sensitive and ultra-
406 high specific blood test for multiple types of cancer on an existing matured high-throughput
407 platform in a cost-effective way.
408
409

410 # Methods

411 **Data collection.**
412 The details of public datasets used in this study are listed in Supplementary Table 1.
413

414 **Preprocess of whole-genome sequencing data.**
415 Adapter is trimmed by Trimmomatic (v0.36) in paired-end mode with following parameters:

416 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads MINLEN:36. After adapter
417 trimming, reads were aligned to the human genome (GRCh37, human_g1k_v37.fa) using
418 BWA-MEM 0.7.15 [39] with default parameters. PCR-duplicate fragments were removed by
419 samblaster (v0.1.24) [40]. Only high-quality autosomal reads (both ends uniquely mapped,
420 either ends with mapping quality score of 30 or greater, properly paired, and not a PCR
421 duplicate) were used for all of the downstream analysis.
422
423 **Preprocess of whole-genome bisulfite sequencing data.**
424 DNA methylation levels measured by WGBS in cfDNA from healthy individuals are obtained
425 from previous publications [41]. Single-end WGBS from cfDNA is processed by the following
426 internal pipeline. Based on FastQC result on the distribution of four nucleotides, the adapter
427 is trimmed by Trim Galore! (v0.6.0) with cutadapt (v2.1.0) and with parameters "--clip_R1 10"
428 and "--clip_R1 10 --three_prime_clip_R1 13" in different WGBS datasets according to the
429 distribution of base quality score along the sequencing cycle. After adapter trimming, reads
430 were aligned to the human genome (GRCh37, human_g1k_v37.fa) using Biscuit
431 (v0.3.10.20190402) with default parameters. PCR-duplicate reads were marked by samtools
432 (v1.9). Only high-quality reads (uniquely mapped, mapping quality score of 30 or greater,
433 and not a PCR duplicate) were used for all of the downstream analysis. Methylation level at
434 each CpG is called by Bis-SNP [42].
435
436 **Identification of cfDNA fragmentation hotspots by CRAG.**
437 The goal of CRAG is to identify the genomic regions with lower fragment coverage and
438 smaller fragment size, i.e., low IFS value. A 200bp sliding window with 20 bp step was
439 applied to scan each chromosome (Only autosome). In the i_{th} window, all the *n* fragments
440 centered within the window were utilized to calculate IFS:

441 $$C_i = n + \sum_{j=1}^{n} \frac{len_j}{S_{len}} S_{coverage} \quad (1)$$

442 len_j is the size of j_{th} fragment. S_{len} is the sum of fragment size across the whole
443 chromosome, $S_{coverage}$ is the sum of fragment number in the current chromosome. Based on
444 this equation, the sliding window with less fragments and shorter fragment size would have a
445 smaller IFS. The windows overlapped with dark regions or with average mappability scores
446 smaller than 0.9 were removed. Dark regions are defined by the merged DAC blacklist and
447 Duke Excluded from UCSC table browser. Mappability score was generated by the GEM
448 mappability program on the human reference genome (GRCh37, human_g1k_v37.fa,
449 45mer) [43]. Poisson distribution was applied to test whether the IFS in the current sliding
450 window was significantly smaller than the local background (5kb and 10 kb) and global
451 background (the whole chromosome).

452 $$P(C_i) = f(\frac{C_i}{\lambda}) = e^{-\lambda}\sum_{k=0}^{C_i}\frac{\lambda^k}{k!} \quad (2)$$

453 Only the windows with p value smaller than a cut-off (p value <= $1.0e^{-05}$) in both the local
454 background and global background (the whole chromosome) were kept for further analysis.
455 Then the p value from the comparison with the global background was used for multiple
456 hypothesis correction (Benjamini and Hochberg method). Windows with false discovery rate
457 (FDR) more than 0.01 were filtered. Finally, the significant windows with distance less than
458 200bp to each other were merged as the hotspots.
459 To remove the possible bias caused by G+C% content, in each sliding window, the IFS and
460 mean G+C% content was calculated. Locally weighted smoothing linear regression (loess,
461 span = 0.75) was used to fit the GC content and IFS. The mean IFS score in each
462 chromosome was added back to the residual value after the correction. The hotspots were
463 called based on the corrected IFS in all the windows.
464 For the comparison, OCF or WPS of the cfDNA fragmentation in each window was also
465 utilized to call hotspots. OCF and WPS value could be negative, which is not suitable for the
466 poisson model. Thus, we used a two-sample Kolmogorov-Smirnov test to test whether WPS
467 (OCF) in the current sliding window was significantly lower or higher than the local
468 background and the whole chromosome.
469
470 **Cancer early detection by cfDNA fragmentation hotspots.**
471 To test whether our CRAG could capture disease-specific hotspots, we used the IFS in the
472 hotspots called in disease plasma samples (hepatic carcinoma, HBV, HBV induced Cirrhosis,
473 breast cancer, bile duct, colorectal cancer, gastric cancer, ovarian cancer and pancreatic
474 cancer) and healthy plasma samples to classify these samples. Here, we took liver cancer vs.
475 healthy samples classification as an example. Ten-fold cross validation was used in this work.
476 In each fold, at the training data set, all the liver cancer samples and healthy samples were
477 merged, respectively (for each coordinate in the whole genome, the IFS was set as the sum
478 of the IFS across all the samples at that genomic position in the same pathological condition).
479 Then, hotspots for liver cancer and healthy individuals were called separately. We kept the
480 top 30,000 most stable hotspots (ranked by the sum of variances in case and control group)
481 as the feature for the classification. In particular, when we call hotspots in the data set from
482 Cristiano et.al. paper [11], since the sample size in the healthy condition is almost ten times
483 larger than single cancer type, we downsampled the healthy controls to the same size as the
484 cancer case in each classification (e.g. Breast cancer vs. Healthy). In addition, the hotspot
485 enrichment at enhancer chromatin states is better after the GC bias correction in Cristiano
486 et.al. data. Although IFS before and after GC bias correction have both been tested, IFS after
487 GC bias correction was shown in the main figure for the classification. Only genomic regions
488 at +/-100bp of the hotspot center were used to retrieve the IFS in each sample (the same
489 strategy was used in PCA and unsupervised clustering analysis). Then in each sample, the

490 IFS at the corresponding hotspot region was z-score transformed based on the mean and
491 standard deviation in each chromosome at that particular sample. Finally, a support vector
492 machine (SVM) classifier with linear kernel and default parameters (fitcsvm function at Matlab
493 2019b) was applied for the learning. At the testing dataset, z-score IFS score in each sample
494 is calculated at the hotspot regions which were already identified at the training set in that
495 particular fold. The average AUC and 95% Confidence Interval (95% CI) of the AUC was
496 calculated based on the classification results at testing dataset across the 10 folds.
497
498 **Tissues-of-origin prediction by cfDNA fragmentation hotspots.**
499 Only samples diagnosed as cancer at 100% specificity by the method above are kept for the
500 tissues-of-origin analysis. The saturation analysis of the fragment number needed for hotspot
501 calling suggested that 400 million fragments are required to achieve the best performance
502 (Supplementary Figure 6). Thus, pathological conditions with less than 400 million fragments
503 in total after merging the samples diagnosed as cancer were not used in the tissues-of-origin
504 analysis (e.g. lung cancer). Bile duct cancer is at the boundary condition with 380 million
505 fragments. Therefore, we performed the analysis with/without bile duct cancer. In the 10-fold
506 cross validation, similar as that for cancer early detection, hotspots for each cancer type in the
507 training set were identified. z-score transformed IFS after GC bias correction in each sample
508 was obtained as the feature. Since the sample size of breast cancer is much larger than that
509 in the other cancer types, after the hotspot calling, we downsampled the breast cancer
510 samples to the median sample size across all the cancer types. The centroid in each cancer
511 type was then calculated by the z-score transformed IFS across all the hotspots in the training
512 set. In the testing dataset, each sample was assigned to the top two candidate cancers based
513 on their distance to the centroids in each cancer type identified at the training set. Distance
514 was calculated by corr function with 'Type' of 'Spearman' at Matlab 2019b. To further narrow
515 down the best candidate cancer type, several decision tree models (fitctree function at Matlab
516 2019b) are learned to identify the better candidate by the top 100,000 most stable hotspots in
517 each possible pair of cancer types at the training set. Finally, between the top two candidates,
518 we applied the corresponding decision tree model to further characterize the best candidate.
519
520 **Saturation analysis of the fragment number needed for cfDNA fragmentation hotspots**
521 **calling.**
522 To benchmark the performance of CRAG on the identification of the known fragmentation
523 hotspots, a group of fragmentation-positive regions (known hotspots) and fragmentation-
524 negative regions were generated. For fragmentation-positive regions, we chose the CGI TSS
525 that are overlapped with conserved TssA chromHMM states (15-state chromHMM) shared
526 across the cell types in NIH Epigenome Roadmap. Regions that are -50bp to +150bp around
527 these TSS were defined as the fragmentation-positive regions. For fragmentation-negative
528 regions, we chose the same number of random genomic regions with the same size, G+C%
529 content, mappability score from conserved Quies chromHMM states shared across the cell

530 types.
531 The high-quality fragments in BH01 were downsampled and then utilized to call hotspots by
532 CRAG. Based on their overlap with the benchmark regions, TP (true positive), FP (false
533 positive), TN (ture negative), FN (false negative) was calculated. F-score for the hotspots was
534 calculated as follow:

535 $$Fscore = 2 * \frac{\Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \quad (3)$$

536 in which, Precision and Recall were calculated using equation (4) and equation (5)
537 respectively:

538 $$\Pr ecision = \frac{TP}{TP + FP} \quad (4)$$

539 $$\operatorname{Re} call = \frac{TP}{TP + FN} \quad (5)$$

540
541
542 **Enrichment of cfDNA fragmentation hotspots in chromHMM states and other**
543 **regulatory elements.**
544 The number of hotspots overlapped with each type of regulatory element was calculated. To
545 calculate its enrichment, after filtering out the dark regions and low mappability regions,
546 random genomic regions were generated with matched chromosomes and sizes. Fisher
547 exact test (two-tail) was performed to calculate the significance of enrichment.
548
549 **PCA analysis of cfDNA fragmentation hotspots across different diseases.**
550 The cfDNA fragmentation hotspots were called at each pathological condition similarly as
551 that in above “Cancer early detection” part. Principal Component Analysis (PCA) was
552 performed on the z-score transformed IFS across all the fragmentation hotspots (pca
553 function at Matlab 2019b).
554
555 **Unsupervised hierarchical clustering analysis of cfDNA fragmentation hotspots.**
556 The cfDNA fragmentation hotspots were called at each pathological condition similarly as
557 that in above “Cancer early detection” part. Top N most variable hotspots were kept for the
558 clustering (ranked by the variation across all the samples). Spearman's rank correlation was
559 used to evaluate the distance among the samples and Weighted average distance
560 (WPGMA, with 'weighted' as parameter in clustergram function at Matlab 2019b) was
561 applied together with linkage method. In samples from Cristiano et.al. paper[11], the similar
562 unsupervised hierarchical clustering strategy is applied. The number of healthy samples
563 here was downsampled to the median sample size (n=27) as the other cancer types. In
564 addition, one-way ANOVA (p value < = 0.01). was used to select the hotspots that showed
565 group specific fragmentation patterns. Top 5,000 hotspots in each group were finally used

566 (ranked by the z-score difference between the samples within the group and outside the
567 group).
568

569 **T-SNE visualization on multi-cancer and healthy.**

570 t-SNE (tsne function at Matlab 2019b) was utilized for the dimensionality reduction and
571 visualization of the fragmentation dynamics in the hotspots across multiple cancer and
572 healthy conditions. Hotspots with a p-value <= 0.01 (one-way ANOVA) were used for the
573 analysis. Distance similarity was calculated by Spearman correlation together with other
574 default parameters (tsne function at Matlab 2019b).

575 **Gene Ontology analysis of cfDNA fragmentation hotspots**
576 Gene Ontology analysis of the cfDNA fragmentation hotspots was performed by GREAT
577 (v4.04) [44]. In this work, the GO Biological Process with FDR q-value of the binomial test ≤
578 0.01 were selected. If more than 10 GO Biological Processes were significant, top ten of
579 them were shown in the main figure.
580

581 **Motif analysis of cfDNA fragmentation hotspots**
582 Motif analysis of cfDNA fragmentation hotspots was performed using HOMER (v4.11) using
583 command 'findMotifsGenome.pl hotspots_file hg19 output_file -size given' [45]. Only the motifs
584 with FDR q-value less than 0.01 were kept. When there were more than 10 motifs enriched,
585 the top 10 of all the enriched motifs were shown in the figures.
586

587 ## Acknowledgements

588 This work was supported by the CCHMC start-up grant and Trustee Award to YL. The
589 authors greatly acknowledge Dr. Yuk Ming Dennis Lo and his circulating nucleic acids
590 research group in the Chinese University of Hong Kong, Dr. Jay Shendure and his research
591 group in the University of Washington, and Drs. Robert B. Scharpf, Victor E. Velculescu and
592 their research group in the Johns Hopkins University School of Medicine for their cfDNA
593 data. This work is supported by the computational resources from Biomedical Informatics
594 (BMI) high performance computing cluster in CCHMC. This work also used the Extreme
595 Science and Engineering Discovery Environment (XSEDE), which is supported by National
596 Science Foundation grant number ACI-1548562. This work used the XSEDE at the
597 Pittsburgh Supercomputing Center (PSC) through allocation MCB190124P and
598 MCB190006P.
599

600 **Availability of data and materials**
601 CRAG is implemented in matlab. The source code is freely available on GitHub
602 (https://github.com/epifluidlab/CRAG.git) under the MIT license for academic researchers.

The source code, readme files, hotspots location, intermediate analysis scripts, and intermediate results are available at Zenodo.org (DOI: 10.5281/zenodo.3928546).

Authors' contributions

YL conceived the study. YL and XZ designed the methodological framework. XZ implemented the methods. XZ and YL performed the data analysis. YL and XZ wrote the manuscript together. All authors read and approved the final manuscript.

Competing interests

Not applicable.

Ethics approval

Although only using publicly available dataset, this research study was approved by Cincinnati Children's Hospital Medical Center Institutional Review Board (#2019-0601) in accordance with the Declaration of Helsinki.

643 # Figures

644

645 **Figure 1**. The schematic of CRAG approach

646

647 **Figure 2**. The enrichment of cfDNA fragmentation hotspots at (a). Transcription Starting
648 Sites (TSS) overlapped with and without CpG island. (b). Exon boundary. (c). Transcription
649 Termination Sites (TTS). (d). CTCF transcription factor binding sites (no TSS within +/- 4 kb).
650 (e). Random genomics regions.

651

652 **Figure 3**. CfDNA fragmentation hotspots are enriched in gene-regulatory regions. The
653 enrichment of (a). DNase-seq and (b). Histone modification signals from hematopoietic cells
654 around the hotspots. (c). The enrichment of hotspots at different chromHMM states across
655 different cell types. (d). Genome browser track of cfDNA fragmentation hotspots.

656

657 **Figure 4**. Half of the cfDNA fragmentation hotspots are still protected by nucleosomes. (a).
658 Less than half of the cfDNA fragmentation hotspots are overlapped with existing 523 open
659 chromatin regions datasets from hematopoietic cells and liver. (b). DNase-seq signal from
660 major hematopoietic cell types around inaccessible and accessible cfDNA fragmentation
661 hotspots. (c). DNA methylation level at cell-free DNA around inaccessible and accessible
662 cfDNA fragmentation hotspots from healthy individuals (d). Motif enrichment around
663 inaccessible and accessible cfDNA fragmentation hotspots. (e). The distribution of the
664 inaccessible and accessible cfDNA fragmentation hotspots in different repeat elements.

665

666 **Figure 5**. The integrated fragmentation score (IFS) from cfDNA fragmentation hotspots can
667 boost the performance of cancer early detection. (a). Volcano plot of z-score differences and
668 p-value (two-way Mann–Whitney U test) for the aberration of IFS in cfDNA fragmentation
669 hotspots between early-stage HCC and healthy. (b). Unsupervised clustering on the Z-score
670 of IFS at the top 10,000 most variable cfDNA fragmentation hotspots called from HCC and
671 healthy samples. (c). Receiver operator characteristics (ROC) for the detection of early-
672 stage HCC by using IFS (after GC bias correction) from all the cfDNA fragmentation
673 hotspots (red), copy number variations (brown), and mitochondrial genome copy number
674 analysis (black). (d). Scatter plots of z-score differences and feature importance (coefficient
675 in linear SVM) split the cfDNA fragmentation hotspots into two groups: hypo-fragmented in
676 cancer (Class I) and hyper-fragmented in cancer (Class II). (e). The fraction of Class I and
677 Class II hotspots that are overlapped with existing open chromatin regions (accessible
678 hotspots) or microsatellite repeats, as well as their relative distance to the nearest TSS. (f)
679 The top 10 motif enrichment at Class I and Class II hotspots. (g). The top 10 enrichment of
680 Gene Ontology Biological Process at Class I and Class II hotspots.

681

682 **Figure 6**. The detection and location of multiple early-stage cancer. (a). t-SNE visualization
683 on the Z-score of IFS (after GC bias correction) at the most variable cfDNA fragmentation
684 hotspots (one-way ANOVA test with p value < 0.01) across multiple different early-stage
685 cancer types and healthy conditions. (b). Unsupervised clustering on Z-score of IFS (after
686 GC bias correction) at the top 40,000 most variable cfDNA fragmentation hotspots across
687 multiple different early-stage cancer types and healthy conditions. (c). The sensitivity across
688 different cancer stages at 100% specificity to distinguish cancer and healthy condition by
689 using IFS (after GC bias correction) at cfDNA fragmentation hotspots. Error bars represent
690 95% confidence intervals. (d). Percentages of patients correctly classified by one of the two
691 most likely types (sum of orange and blue bars) or the most likely type (blue bar). Error bars
692 represent 95% confidence intervals.
693
694 **Table 1**. CRAG performance for cancer early detection.
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720

721 **References**

722 1. Schiffman, J. D., Fisher, P. G. & Gibbs, P. Early detection of cancer: past, present, and
723 future. *Am Soc Clin Oncol Educ Book* 57–65 (2015).
724 2. Early detection: a long road ahead. *Nature reviews. Cancer* vol. 18 401 (2018).
725 3. Welch, H. G. & Black, W. C. Overdiagnosis in cancer. *J. Natl. Cancer Inst.* **102**, 605–
726 613 (2010).
727 4. Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA.
728 *Sci. Transl. Med.* **9**, (2017).
729 5. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a
730 multi-analyte blood test. *Science* **359**, 926–930 (2018).
731 6. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA
732 with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
733 7. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human
734 malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
735 8. Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using
736 methylation signatures in cell-free DNA. *Ann. Oncol.* (2020)
737 doi:10.1016/j.annonc.2020.02.011.
738 9. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free
739 DNA methylomes. *Nature* **563**, 579–583 (2018).
740 10. Abbosh, C., Swanton, C. & Birkbak, N. J. Clonal haematopoiesis: a source of biological
741 noise in cell-free DNA analyses. *Annals of oncology: official journal of the European*
742 *Society for Medical Oncology / ESMO* vol. 30 358–359 (2019).
743 11. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer.
744 *Nature* **570**, 385–389 (2019).
745 12. Chabon, J. J. *et al.* Integrating genomic features for non-invasive early lung cancer
746 detection. *Nature* **580**, 245–251 (2020).
747 13. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA
748 Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*
749 **164**, 57–68 (2016).
750 14. Ulz, P. *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA.
751 *Nat. Genet.* **48**, 1273–1278 (2016).
752 15. Jiang, P. *et al.* Preferred end coordinates and somatic variants as signatures of
753 circulating tumor DNA associated with hepatocellular carcinoma. *Proc. Natl. Acad. Sci.*
754 *U. S. A.* **115**, E10925–E10933 (2018).
755 16. Han, D. S. C. *et al.* The Biology of Cell-free DNA Fragmentation and the Roles of
756 DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* **106**, 202–214 (2020).
757 17. Cheung, M.-S., Down, T. A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput
758 sequencing data and its correction by BEADS. *Nucleic Acids Res.* **39**, e103 (2011).
759 18. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-

760 throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
761 19. Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random
762 fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC*
763 *Genomics* **16 Suppl 13**, S1 (2015).
764 20. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers.
765 *Science* **362**, (2018).
766 21. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse
767 human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).
768 22. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human
769 epigenomes. *Nature* **518**, 317–330 (2015).
770 23. Sun, K. *et al.* Orientation-aware plasma cell-free DNA fragmentation analysis in open
771 chromatin regions informs tissue of origin. *Genome Res.* **29**, 418–427 (2019).
772 24. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and
773 yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
774 25. 1000 Genomes Project Consortium *et al.* A global reference for human genetic
775 variation. *Nature* **526**, 68–74 (2015).
776 26. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the
777 human genome. *Nature* **489**, 57–74 (2012).
778 27. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The
779 International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and
780 Discovery. *Cell* **167**, 1145–1149 (2016).
781 28. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts
782 human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
783 29. Jensen, T. J. *et al.* Whole genome bisulfite sequencing of cell-free DNA and its cellular
784 contributors uncovers placenta hypomethylated domains. *Genome Biol.* **16**, 78 (2015).
785 30. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to
786 initiate reprogramming. *Cell* **161**, 555–568 (2015).
787 31. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements.
788 *Trends Genet.* **16**, 418–420 (2000).
789 32. Jiang, P. *et al.* Lengthening and shortening of plasma DNA in hepatocellular carcinoma
790 patients. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1317–25 (2015).
791 33. Wan, N. *et al.* Machine learning enables detection of early-stage colorectal cancer by
792 whole-genome sequencing of plasma cell-free DNA. *BMC Cancer* **19**, 832 (2019).
793 34. Gonzalez, H., Hagerling, C. & Werb, Z. Roles of the immune system in cancer: from
794 tumor initiation to metastatic progression. *Genes Dev.* **32**, 1267–1284 (2018).
795 35. Liu, E. M. *et al.* Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole
796 Genomes. *Cell Syst* **8**, 446–455.e8 (2019).
797 36. Mammana, A., Vingron, M. & Chung, H.-R. Inferring nucleosome positions with their
798 histone mark annotation from ChIP data. *Bioinformatics* **29**, 2547–2554 (2013).
799 37. Teng, M. & Irizarry, R. A. Accounting for GC-content bias reduces systematic errors and

800 batch effects in ChIP-seq data. *Genome Res.* **27**, 1930–1938 (2017).
801 38. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing
802 libraries. *Genome Biol.* **12**, R18 (2011).
803 39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
804 transform. *Bioinformatics* **25**, 1754–1760 (2009).
805 40. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant
806 read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
807 41. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for
808 noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci.*
809 *U. S. A.* **112**, E5503–12 (2015).
810 42. Liu, Y., Siegmund, K. D., Laird, P. W. & Berman, B. P. Bis-SNP: combined DNA
811 methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* **13**, R61 (2012).
812 43. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLoS One*
813 **7**, e30377 (2012).
814 44. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions.
815 *Nat. Biotechnol.* **28**, 495–501 (2010).
816 45. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime
817 cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**,
818 576–589 (2010).

Figure 1

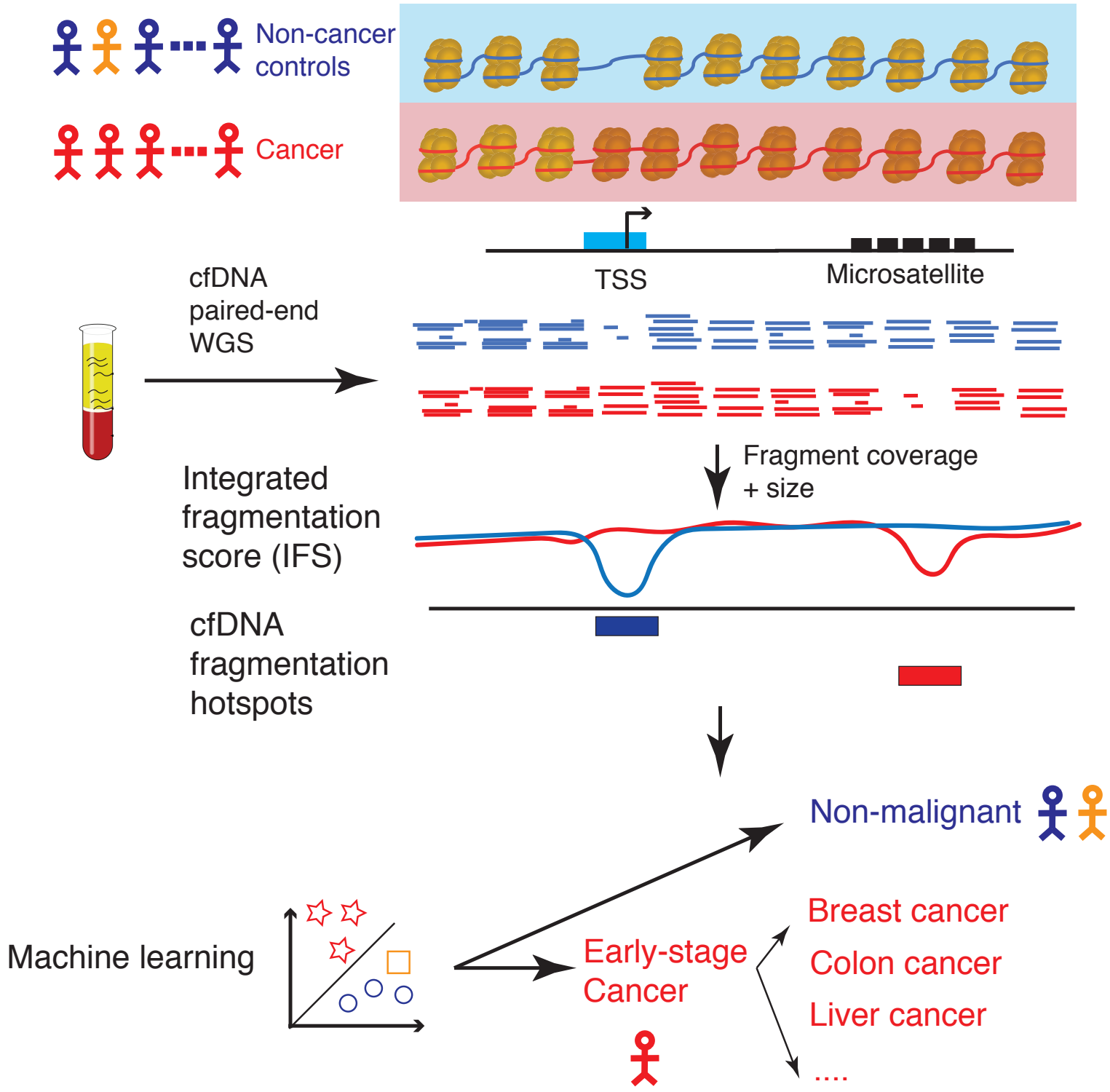

Non-cancer controls
Cancer
TSS
Microsatellite
cfDNA paired-end WGS
Fragment coverage + size
Integrated fragmentation score (IFS)
cfDNA fragmentation hotspots
Non-malignant
Machine learning
Early-stage Cancer
Breast cancer
Colon cancer
Liver cancer
....

Figure 2

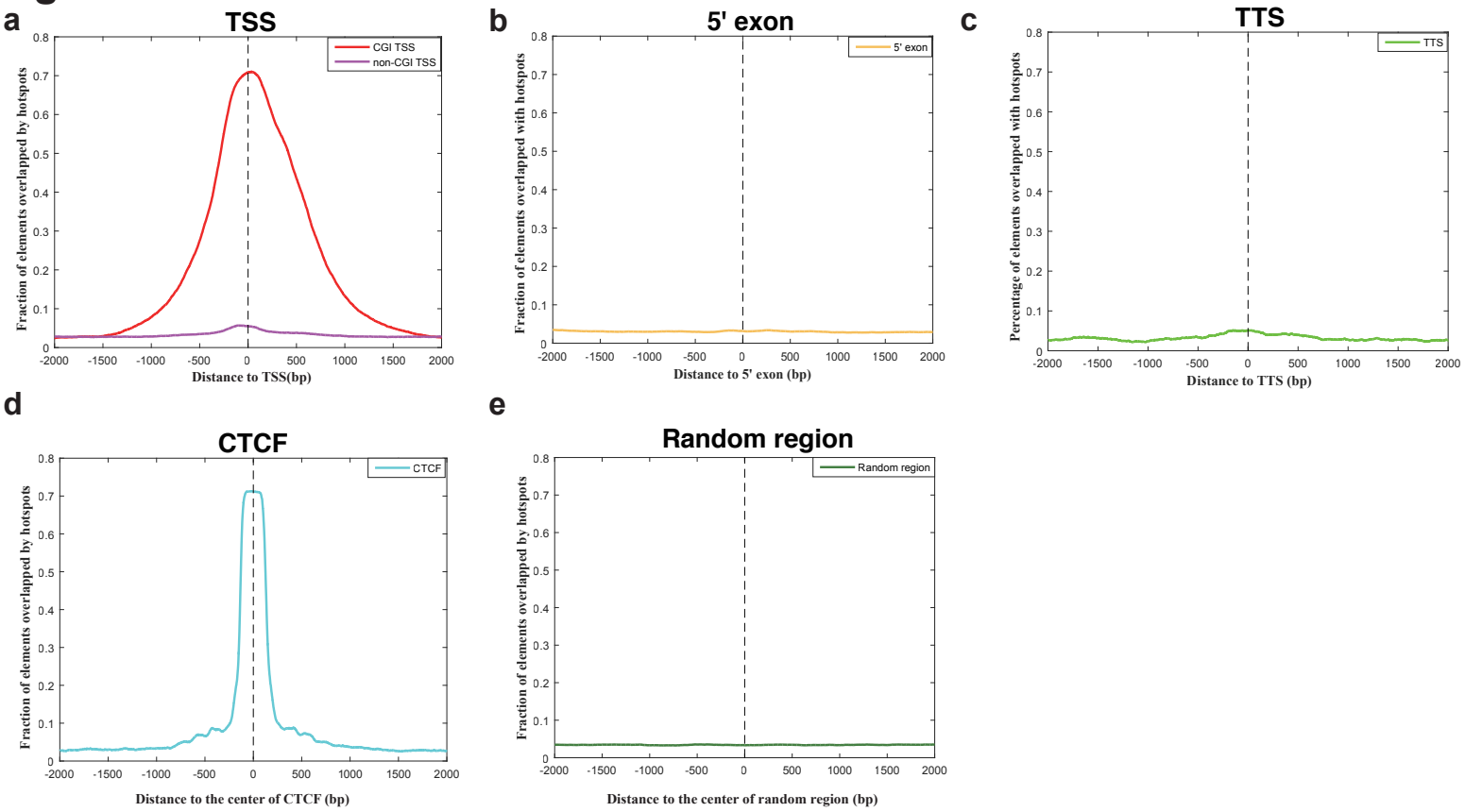
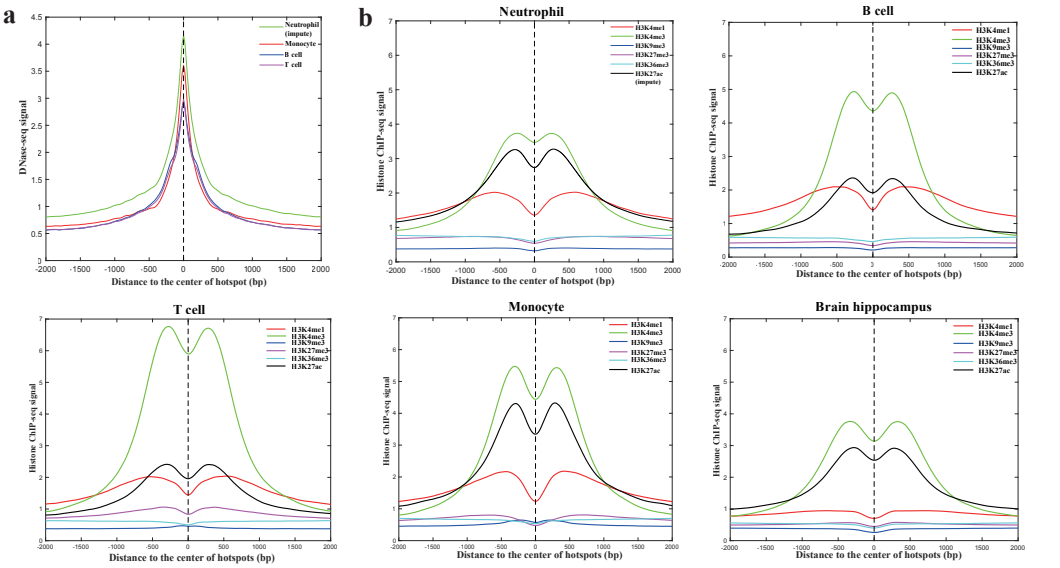
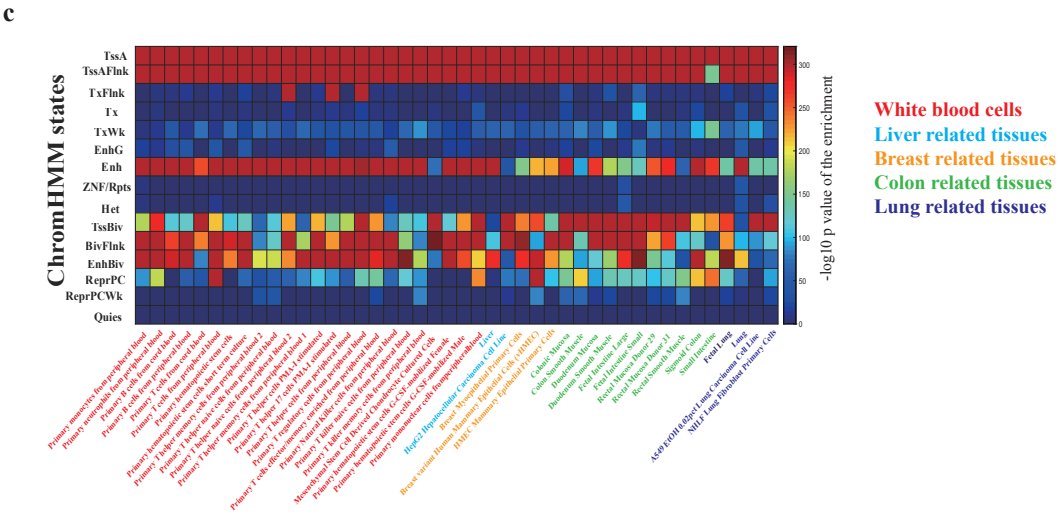

a
TSS
CGI TSS
non-CGI TSS
Fraction of elements overlapped by hotspots
Distance to TSS(bp)
b
5' exon
5' exon
Fraction of elements overlapped with hotspots
Distance to 5' exon (bp)
c
TTS
TTS
Percentage of elements overlapped with hotspots
Distance to TTS (bp)
d
CTCF
CTCF
Fraction of elements overlapped by hotspots
Distance to the center of CTCF (bp)
e
Random region
Random region
Fraction of elements overlapped by hotspots
Distance to the center of random region (bp)

Figure 3


a
DNase-seq signal
Distance to the center of hotspot (bp)
Neutrophil (impute)
Monocyte
B cell
T cell
b
Neutrophil
B cell
T cell
Monocyte
Brain hippocampus
Histone ChIP-seq signal
Distance to the center of hotspot (bp)
Distance to the center of hotspots (bp)
H3K4me1
H3K4me3
H3K9me3
H3K27me3
H3K36me3
H3K27ac
H3K27ac (impute)


c
ChromHMM states
TssA
TssAFlnk
TxFlnk
Tx
TxWk
EnhG
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
-log10 p value of the enrichment
White blood cells
Liver related tissues
Breast related tissues
Colon related tissues
Lung related tissues

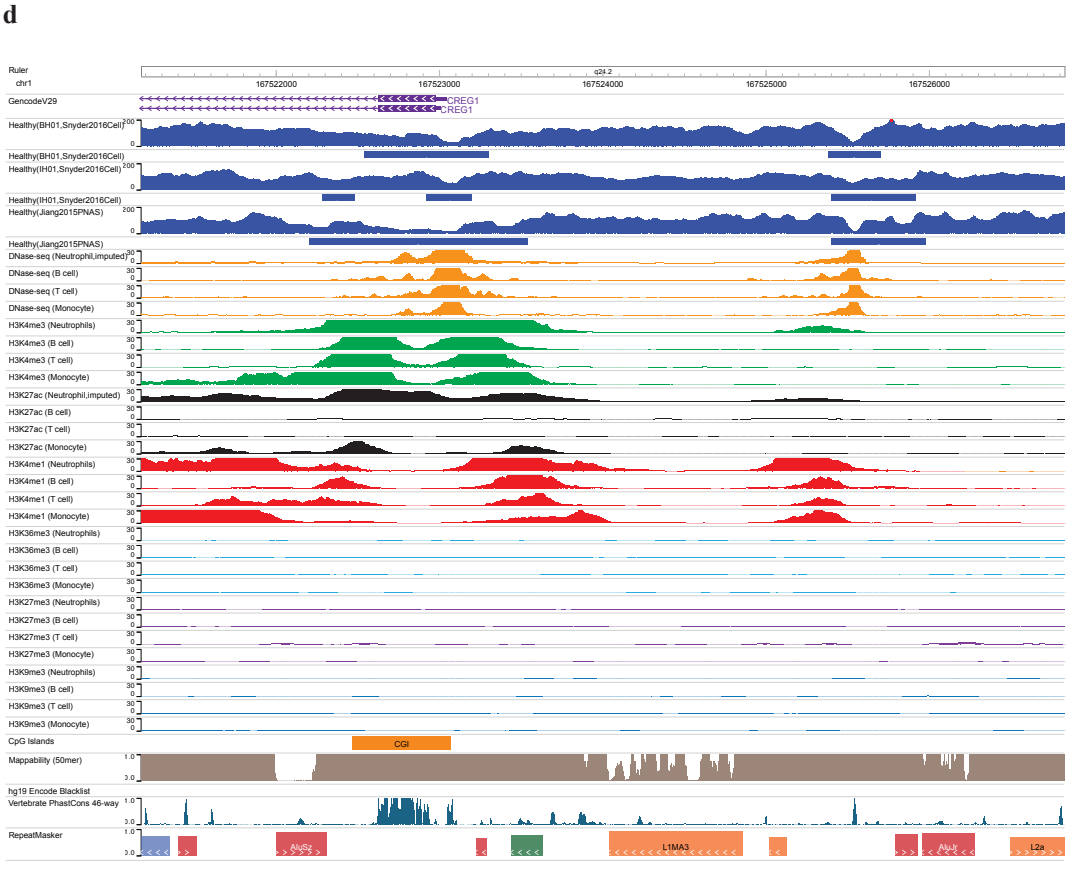

d
Ruler
chr1
GencodeV29
CREG1
Healthy(BH01,Snyder2016Cell)
Healthy(IH01,Snyder2016Cell)
Healthy(Jiang2015PNAS)
DNase-seq (Neutrophil,imputed)
DNase-seq (B cell)
DNase-seq (T cell)
DNase-seq (Monocyte)
H3K4me3 (Neutrophils)
H3K4me3 (B cell)
H3K4me3 (T cell)
H3K4me3 (Monocyte)
H3K27ac (Neutrophil,imputed)
H3K27ac (B cell)
H3K27ac (T cell)
H3K27ac (Monocyte)
H3K4me1 (Neutrophils)
H3K4me1 (B cell)
H3K4me1 (T cell)
H3K4me1 (Monocyte)
H3K36me3 (Neutrophils)
H3K36me3 (B cell)
H3K36me3 (T cell)
H3K36me3 (Monocyte)
H3K27me3 (Neutrophils)
H3K27me3 (B cell)
H3K27me3 (T cell)
H3K27me3 (Monocyte)
H3K9me3 (Neutrophils)
H3K9me3 (B cell)
H3K9me3 (T cell)
H3K9me3 (Monocyte)
CpG Islands
CGI
Mappability (50mer)
hg19 Encode Blacklist
Vertebrate PhastCons 46-way
RepeatMasker
AluSz
L1MA3
AluJr
L2a

Figure 4

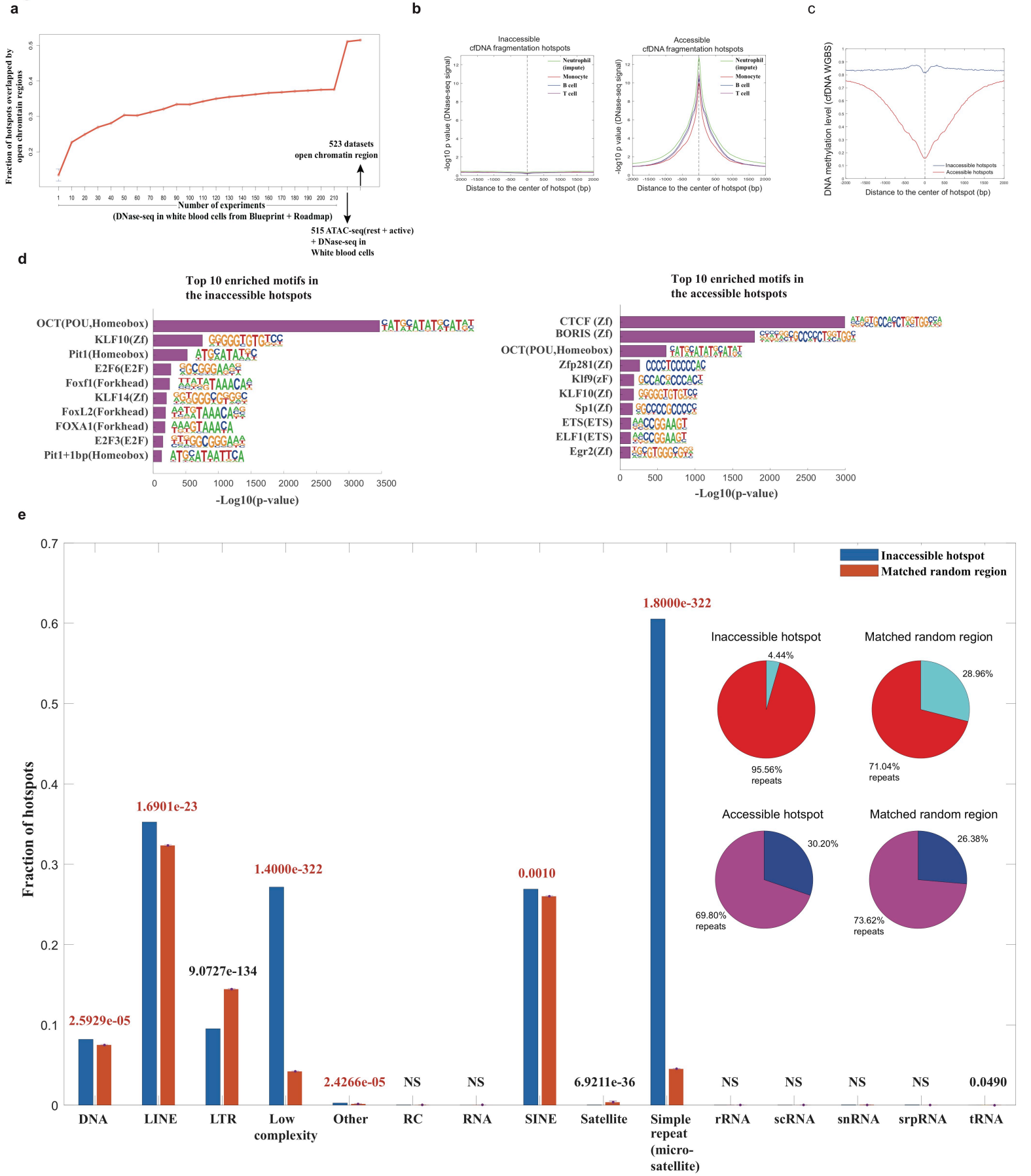

a
Fraction of hotspots overlapped by open chromtain regions
523 datasets
open chromatin region
Number of experiments
(DNase-seq in white blood cells from Blueprint + Roadmap)
515 ATAC-seq(rest + active)
+ DNase-seq in
White blood cells
b
Inaccessible
cfDNA fragmentation hotspots
Accessible
cfDNA fragmentation hotspots
Neutrophil (impute)
Monocyte
B cell
T cell
-log10 p value (DNase-seq signal)
Distance to the center of hotspot (bp)
c
DNA methylation level (cfDNA WGBS)
Inaccessible hotspots
Accessible hotspots
Distance to the center of hotspot (bp)
d
Top 10 enriched motifs in
the inaccessible hotspots
OCT(POU,Homeobox)
KLF10(Zf)
Pit1(Homeobox)
E2F6(E2F)
Foxf1(Forkhead)
KLF14(Zf)
FoxL2(Forkhead)
FOXA1(Forkhead)
E2F3(E2F)
Pit1+1bp(Homeobox)
-Log10(p-value)
Top 10 enriched motifs in
the accessible hotspots
CTCF (Zf)
BORIS (Zf)
OCT(POU,Homeobox)
Zfp281(Zf)
Klf9(zF)
KLF10(Zf)
Sp1(Zf)
ETS(ETS)
ELF1(ETS)
Egr2(Zf)
-Log10(p-value)
e
Inaccessible hotspot
Matched random region
Fraction of hotspots
2.5929e-05
1.6901e-23
9.0727e-134
1.4000e-322
2.4266e-05
NS
NS
0.0010
6.9211e-36
1.8000e-322
NS
NS
NS
NS
0.0490
DNA
LINE
LTR
Low complexity
Other
RC
RNA
SINE
Satellite
Simple repeat (micro-satellite)
rRNA
scRNA
snRNA
srpRNA
tRNA
Inaccessible hotspot
4.44%
95.56% repeats
Matched random region
28.96%
71.04% repeats
Accessible hotspot
30.20%
69.80% repeats
Matched random region
26.38%
73.62% repeats

Figure 5

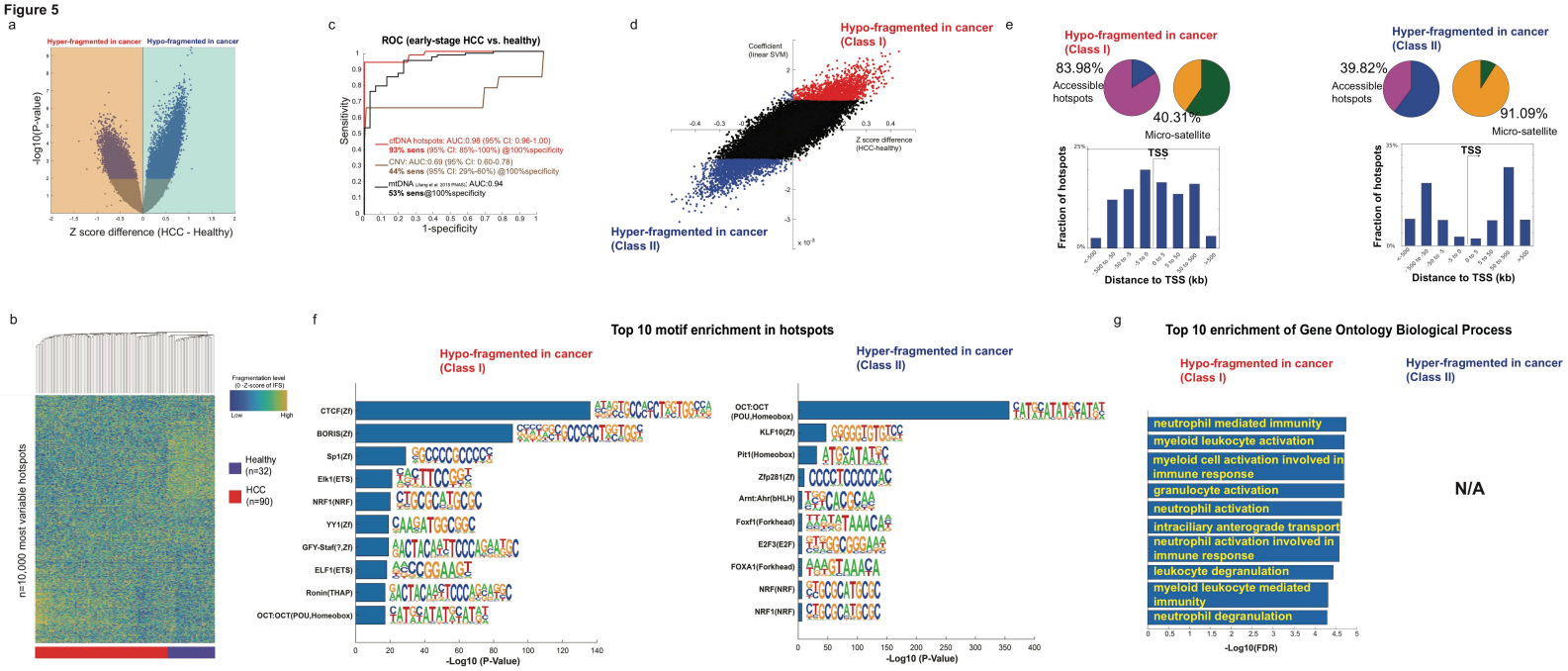

a
Hyper-fragmented in cancer
Hypo-fragmented in cancer
-log10(P-value)
Z score difference (HCC - Healthy)
c
ROC (early-stage HCC vs. healthy)
Sensitivity
1-specificity
cfDNA hotspots: AUC:0.98 (95% CI: 0.96-1.00)
93% sens (95% CI: 85%-100%) @100%specificity
CNV: AUC:0.69 (95% CI: 0.60-0.78)
44% sens (95% CI: 29%-60%) @100%specificity
mtDNA (Jiang et al. 2019 PNAS): AUC:0.94
53% sens@100%specificity
d
Hypo-fragmented in cancer (Class I)
Coefficient (linear SVM)
Z score difference (HCC-healthy)
Hyper-fragmented in cancer (Class II)
e
Hypo-fragmented in cancer (Class I)
83.98%
Accessible hotspots
40.31%
Micro-satellite
TSS
Fraction of hotspots
Distance to TSS (kb)
Hyper-fragmented in cancer (Class II)
39.82%
Accessible hotspots
91.09%
Micro-satellite
TSS
Fraction of hotspots
Distance to TSS (kb)
b
n=10,000 most variable hotspots
Fragmentation level (0 -Z-score of IFS)
Low
High
Healthy (n=32)
HCC (n=90)
f
Top 10 motif enrichment in hotspots
Hypo-fragmented in cancer (Class I)
CTCF(Zf)
BORIS(Zf)
Sp1(Zf)
Elk1(ETS)
NRF1(NRF)
YY1(Zf)
GFY-Staf(?,Zf)
ELF1(ETS)
Ronin(THAP)
OCT:OCT(POU,Homeobox)
-Log10 (P-Value)
Hyper-fragmented in cancer (Class II)
OCT:OCT (POU,Homeobox)
KLF10(Zf)
Plt1(Homeobox)
Zfp281(Zf)
Arnt:Ahr(bHLH)
Foxf1(Forkhead)
E2F3(E2F)
FOXA1(Forkhead)
NRF(NRF)
NRF1(NRF)
-Log10 (P-Value)
g
Top 10 enrichment of Gene Ontology Biological Process
Hypo-fragmented in cancer (Class I)
neutrophil mediated immunity
myeloid leukocyte activation
myeloid cell activation involved in immune response
granulocyte activation
neutrophil activation
intracilliary anterograde transport
neutrophil activation involved in immune response
leukocyte degranulation
myeloid leukocyte mediated immunity
neutrophil degranulation
-Log10(FDR)
Hyper-fragmented in cancer (Class II)
N/A

Figure 6

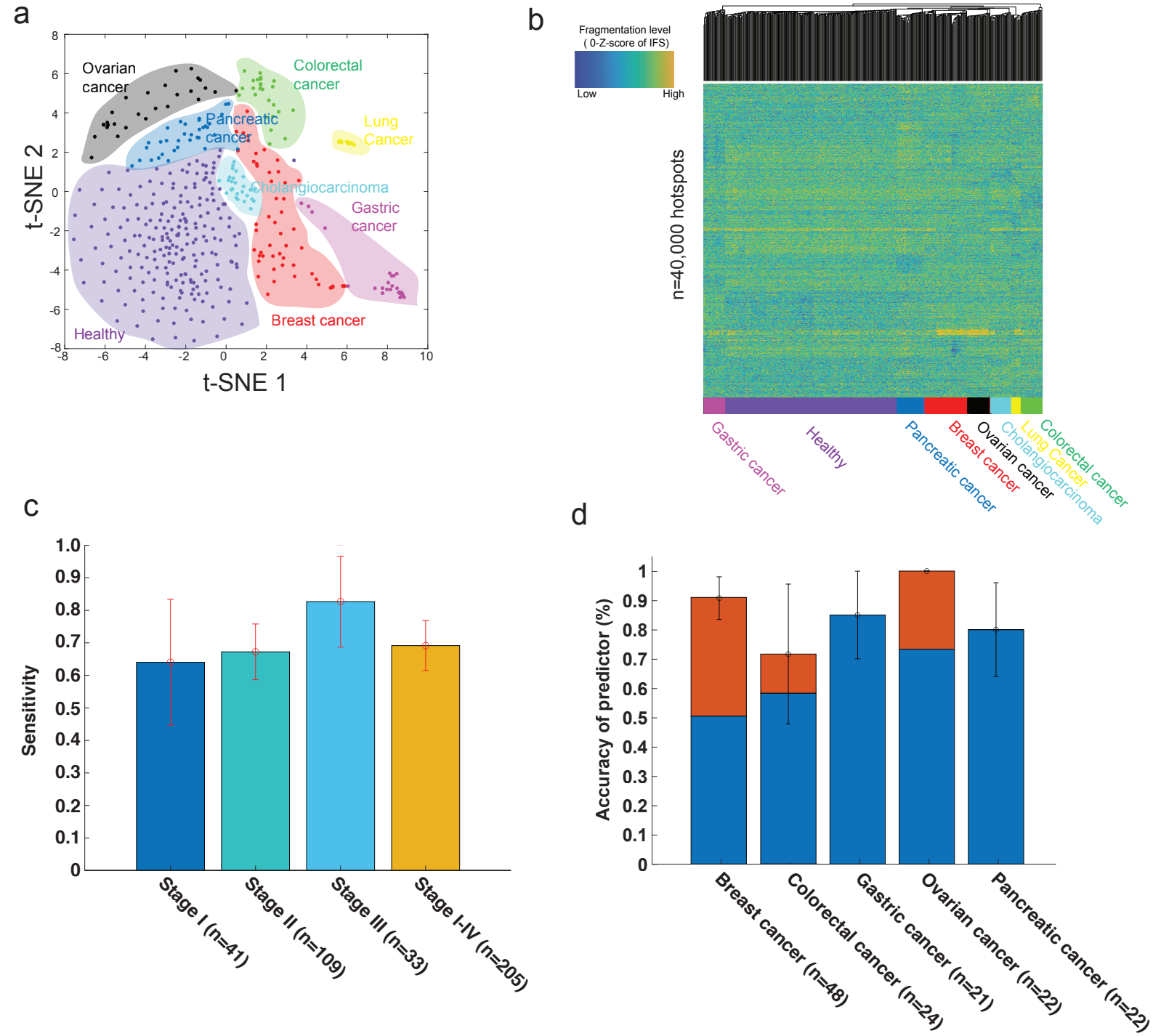

a
Ovarian cancer
Colorectal cancer
Pancreatic cancer
Lung Cancer
Cholangiocarcinoma
Gastric cancer
Healthy
Breast cancer
t-SNE 1
t-SNE 2
b
Fragmentation level
(0-Z-score of IFS)
Low
High
n=40,000 hotspots
Gastric cancer
Healthy
Pancreatic cancer
Breast cancer
Ovarian cancer
Cholangiocarcinoma
Lung Cancer
Colorectal cancer
c
Sensitivity
Stage I (n=41)
Stage II (n=109)
Stage III (n=33)
Stage I-IV (n=205)
d
Accuracy of predictor (%)
Breast cancer (n=48)
Colorectal cancer (n=24)
Gastric cancer (n=21)
Ovarian cancer (n=22)
Pancreatic cancer (n=22)

Table 1: CRAG performance for cancer early detection

			100% specificity			95% specificity			85% specificity		
		Individuals analysed	Individuals detected	Sensitivity (%)	95% CI (%)	Individuals detected	Sensitivity (%)	95% CI (%)	Individuals detected	Sensitivity (%)	95% CI (%)
Cancer Type	Breast	54	44	0.8133	0.7168 - 0.9098	48	0.8867	0.7892 - 0.9842	51	0.9433	0.8865 - 1.0000
	Bile duct	26	11	0.4167	0.1819 - 0.6515	12	0.4500	0.2320 - 0.6680	14	0.5333	0.3166 - 0.7500
	Colorectal	27	24	0.8833	0.7636 - 1.0000	25	0.9167	0.8051 - 1.0000	25	0.9167	0.8051 - 1.0000
	Gastric	27	25	0.9333	0.8462 - 1.0000	25	0.9333	0.8462 - 1.0000	26	0.9667	0.9013 - 1.0000
	Lung	12	6	0.5000	0.2078 - 0.7922	8	0.6500	0.3560 - 0.9400	8	0.6500	0.3560 - 0.9400
	Ovarian	28	15	0.5500	0.3519 - 0.7451	23	0.8167	0.6592 - 0.9741	24	0.8500	0.7263 - 0.9737
	Pancreatic	34	19	0.5667	0.4251 - 0.7082	21	0.6333	0.4790 - 0.7877	22	0.6583	0.4889 - 0.8278
Cancer Stage	I	41	28	0.6350	0.4428 - 0.8272	34	0.8250	0.7046 - 0.9454	34	0.8250	0.7046 - 0.9454
	II	109	72	0.6667	0.5816 - 0.7518	79	0.7297	0.6503 - 0.8092	85	0.7813	0.7018 - 0.8607
	III	33	25	0.8200	0.6813 - 0.9587	28	0.8850	0.7796 - 0.9904	30	0.9350	0.8698 - 1.0000
	I-IV	205	141	0.6858	0.6098 - 0.7617	159	0.7742	0.7229 - 0.8255	167	0.8128	0.7596 - 0.8659