

1 **Ultra-low coverage fragmentomic model of cell-free DNA for cancer detection**
2 **based on whole-exome regions.**

3
4 Apiwat Sangphukieo¹, Pitiporn Noisagul¹, Patcharawadee Thongkumkoon¹, and Parunya
5 Chaiyawat¹

6
7 ¹Center of Multidisciplinary Technology for Advanced Medicine (CMUTEAM), Faculty of
8 Medicine, Chiang Mai University, Chiang Mai, Thailand

9
10 **Corresponding Author:** Parunya Chaiyawat, parunya.chaiyawat@cmu.ac.th (P.C.)

11 Keywords; Cancer, DELFI, fragmentomic, DNA sequencing, cell-free DNA

12
13 **Abstract**

14 Cell-free DNA (cfDNA) has shown promise as a non-invasive biomarker for cancer screening
15 and monitoring. The current advanced machine learning (ML) model, known as DNA evaluation
16 of fragments for early interception (DELFI), utilizes the short and long fragmentation pattern of
17 cfDNA and has demonstrated exceptional performance. However, the application of cfDNA-
18 based model can be limited by the high cost of whole-genome sequencing (WGS). In this study,
19 we present a novel ML model for cancer detection that utilizes cfDNA profiles generated from
20 all protein-coding genes in the genome (exome) with only 0.08X of WGS coverage. Our model
21 was trained on a dataset of 721 cfDNA profiles, comprising 426 cancer patients and 295 healthy
22 individuals. Performance evaluation using a ten-fold cross-validation approach demonstrated that
23 the new ML model using whole-exome regions, called xDELFI, can achieve high accuracy in

24 cancer detection (Area under the ROC curve; AUC=0.896, 95%CI = 0.878 - 0.916), comparable
25 to the model using WGS (AUC=0.920, 95%CI = 0.901 – 0.936). Notably, we observed distinct
26 fragmentation patterns between exonic regions and the whole-genome, suggesting unique
27 genomic features within exonic regions. Furthermore, we demonstrate the potential benefits of
28 combining mutation detection in cfDNA with xDELFI, which enhance the model sensitivity. Our
29 proof-of-principle study indicates that the fragmentomic ML model based solely on whole-
30 exome regions retains its predictive capability. With the ultra-low sequencing coverage of the
31 new model, it could potentially improve the accessibility of cfDNA-based cancer diagnosis and
32 aid in early detection and treatment of cancer.

33 **Introduction**

34 Cell-free DNA (cfDNA) is a term for DNA fragments that are mostly released into body
35 fluids from various sources, such as apoptotic or necrotic cells, as well as from active secretion
36 [1]. The elevation of cfDNA in the blood can be an indicator of various health conditions,
37 especially cancer [2, 3]. Analysis of cfDNA in the context of a liquid biopsy is gaining
38 popularity in the field of oncology as it shows variety of benefits, for example, non-invasiveness
39 that require only blood draw, the ability of each sample to reflect all tumor lesions in the body,
40 and ability to real-time monitor cancer progression due to short half-life of cfDNA (~2 h) [4]. A
41 particular subset of cfDNA known as circulating tumor DNA (ctDNA), which contains genetic
42 alterations associated with cancer development, holds immense potential for diagnosing the
43 disease, tracking cancer progression, evaluating treatment response, and identifying potential
44 therapeutic targets [1]. However, the extremely low concentration of ctDNA in plasma, which is
45 lower than 0.01% of the total cfDNA concentration [5], presents a significant obstacle for
46 accurate detection and analysis.

47 Numerous studies demonstrated that the fragment length of cfDNA derived from normal
48 cells is about 166 bp on average, while the cfDNA derived from tumor cells is more fragmented
49 with the size between 90–150 bp [6-9]. Furthermore, there has been a notable observation of a
50 positive correlation between the proportion of short cfDNA fragments (<150 bp) and the tumor
51 DNA fraction present in the plasma [7]. In addition, numerous research studies have examined
52 the possibility of using shorter cfDNA fragments to improve the detection of copy number
53 variations (CNV) and single nucleotide variations (SNV) in the cancer patients [7, 10, 11]. In
54 recent years, there has been a growing interest in the study of cfDNA fragmentation patterns,
55 known as fragmentomics [12, 13], including various aspects of cfDNA fragmentation, such as

56 fragment sizes, abundance, integrity, end motifs, window protection score, and preferable end
57 coordinates [12, 14]. The advancement of next-generation sequencing technology has enabled
58 whole-genome sequencing (WGS) showing that cfDNA fragment size distribution patterns in
59 cancer patients are more variable than those in healthy individuals [15]. These differences in
60 cfDNA fragmentation patterns reflect changes in chromatin structures, as well as other genomic
61 and epigenomic abnormalities in cancer [16] providing a framework for developing diagnostic
62 tools [17]. Recently, a machine-learning (ML) approach was applied to learn the pattern of
63 cfDNA fragmentation from low-coverage WGS data, known as DNA evaluation of fragments
64 for early interception (DELFI), showing excellent performance in the classification of cancer-
65 carried individuals and healthy individuals [15]. DELFI evaluated the fragment size coverage of
66 short cfDNA fragments (100–150 bp) and long cfDNA fragments (151–220 bp) inside 100
67 kilobase (Kb) non-overlapping consecutive bins and integrated them into a 5-megabase (Mb)
68 non-overlapping consecutive window. The aberrations of the ratio between short cfDNA
69 fragments and long cfDNA fragments within each window is increased in cancer patients, but
70 not in healthy individuals. This cfDNA fragmentation size difference was utilized as a key
71 feature in gradient-boosting model and showed outstanding performance with an overall area
72 under the curve (AUC) of 0.94 [15]. The performance of DELFI model has been further
73 improved by reducing feature dimension, applying different feature extraction strategy, and
74 utilizing ensemble algorithm [13, 18]. Nevertheless, its practical application is limited by the
75 current WGS cost. While reducing the sequencing coverage to 0.1X WGS has been
76 demonstrated as an effective cost-saving measure for estimating tumor fraction [19], it also
77 introduced significant alterations in the fragmentation profiles [13, 15], which could potentially
78 impact the accuracy and reliability of DELFI score.

79 While most of the human genome is non-coding and its function remains largely
80 unknown, only 1% of the genome is comprised of protein-coding regions, known as exons.
81 Exons are the protein-coding regions of a gene that are transcribed into mRNA and ultimately
82 translated into proteins. The sequences of exons are typically highly conserved across different
83 species, reflecting the importance of these regions in protein function and evolution [20].
84 Mutations in exons can lead to altered protein function and are often associated with genetic
85 diseases. With the availability of large databases of known SNPs and known pathogenic variants,
86 whole-exome sequencing (WES), an alternative approach to whole-genome sequencing (WGS)
87 by targeting the exonic regions, have been extensively applied to identify casual mutations in
88 cancer patients with much lower cost than WGS [21]. Recently, researchers have been exploring
89 the use of cfDNA fragmentation profiles at exonic regions to infer gene expression, which can
90 apply to various clinical applications such as tumor detection, subtype classification, treatment
91 response assessment, and prognostic implications [22].

92 In this study, we present a novel approach for developing the DELFI model that can
93 classify cancer patients and healthy individuals based on exonic regions. Our new exome-based
94 cfDNA fragmentation model, called xDELFI, can efficiently distinguish between cancer patients
95 and healthy individuals and can classify the tissue of origin with reasonable accuracy.
96 Furthermore, our study showed that combining xDELFI with mutation information can further
97 enhance the prediction performance, which highlights the potential benefits of utilizing WES
98 data in mutation calling and xDELFI score prediction. The new model paves the way to create
99 more cost-effective methods for cancer diagnosis and monitoring.

100

101 **Methods**

102 **Data collection**

103 The paired-end cfDNA whole-genome sequencing samples from four different articles stored in
104 FinaleDB database were collected. In total of 426 samples from 16 cancer types and 295 healthy
105 individuals were processed [7, 15, 16, 23] (Supplementary Figure 1). The data was pre-processed
106 steps, which had been proceed in finaleDB, to have high quality data by 1) trimming all
107 sequences to 50 bp to minimize possible batch effects, 2) exclude un-properly mapped in pair, 3)
108 exclude non-primary alignments, 4) exclude reads with mapping quality < 30, and 4) remove
109 duplicated reads. Data quality control was proceeded by measuring correlation of the whole
110 genome sequence coverage observed in this study and those reported in the original article
111 (Supplementary Figure S3).

112

113 **DELFI and exomeDELFI score calculation**

114 DELFI score calculation is followed [15]. The original script of DELFI is available in
115 https://github.com/cancer-genomics/delfi_scripts. Briefly, all chromosomes have been split into
116 consecutive, non-overlapping 100 kb bins. The lowest coverage bins (top10%) and fragments
117 falling in the Duke blacklisted regions
118 (<http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/>) were
119 removed. Short (100 and 150 bp) and long (151 and 220 bp) cfDNA fragment coverage of each
120 100k bin were counted. For exomeDELFI, only fragments falling in the exome regions based on
121 Agilent - SureSelect All Exon V7 (bed file is available at <https://github.com/mobidic/BARMEN>)
122 were counted. Loess regression-based approach is applied to account for GC content bias for
123 each 100kb bin in the genome of each sample. The 100k bins were combined into 5Mb bins (504
124 bins). Therefore, the total number of bins/features from short and long fragment coverage is

125 1,008. Then, GC-adjusted short and long fragment coverage were centered and scaled for each
126 sample to have mean 0 and unit standard deviation. Feature selection, performed only on the
127 training data (in each cross-validation run), removed bins that were highly correlated (correlation
128 > 0.9) or had near zero variance. Gradient boosting machine (GBM) was implemented using the
129 caret package in R with parameters of n.trees=150, interaction.depth=3, shrinkage=0.1, and
130 n.minobsinside=10. The prediction error was evaluated by performing ten-fold cross validation
131 (CV) repeated ten times. The final DELFI score is an average probability of the cancer class
132 from ten repeated CV.

133

134 **xDELFI score calculation**

135 Firstly, we applied three fragment size thresholds, which are short (100-150), medium (150-220),
136 and long (>220), to count the number of fragments in each 100k bin along the genome, instead of
137 using short (100-150) and long (150-220) threshold in previous study (Figure 1B). In addition to
138 fragment length coverage, overall fragment coverage is counted by summation of all fragments
139 in each 100k bin. Secondly, loess regression-based approach is applied to account for GC content
140 bias for each 100kb bin. Thirdly, the 100k bins were combined into 5Mb bins (504 bins) for each
141 chromosome arm. Then, the fragment size distribution (FSD) of 5bp bin in range of 100 to 220bp
142 (24 bins) in each chromosome arm (39 arms) was calculated, and loess regression-based
143 approach is applied to account for GC content bias for each bin. Therefore, the total number of
144 features from fragment length coverage (504x3), overall coverage (504), and FSD (24x39) is
145 2,952. Feature selection method was applied only on the training data to remove highly
146 correlated features (correlation > 0.9) or uninformative features (zero variance). Finally, two
147 different machine learning models, GBM and support vector machine (SVM), were combined as

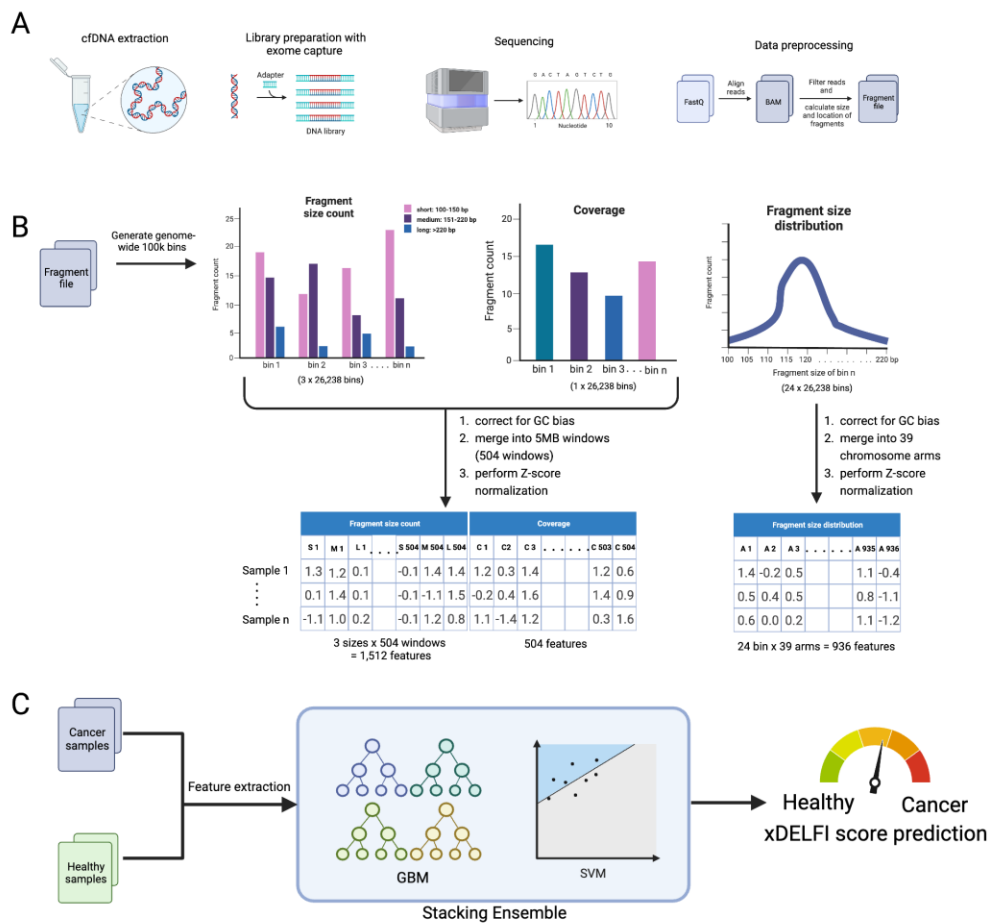
148 stacking ensemble by GBM to train xDELFI model. The stacking ensemble model was
 149 implemented using caretEnsemble package with GBM parameters of n.trees=70,
 150 interaction.depth=3, shrinkage=0.1, and n.minobsinside=10. Ten-fold CV repeated ten times was
 151 conducted to evaluate the prediction error and the average probability of the cancer class was
 152 used as the final xDELFI score. xDELFI script is freely available at Github:
 153 <https://github.com/asangphukieo/xDELFI>.

154

155

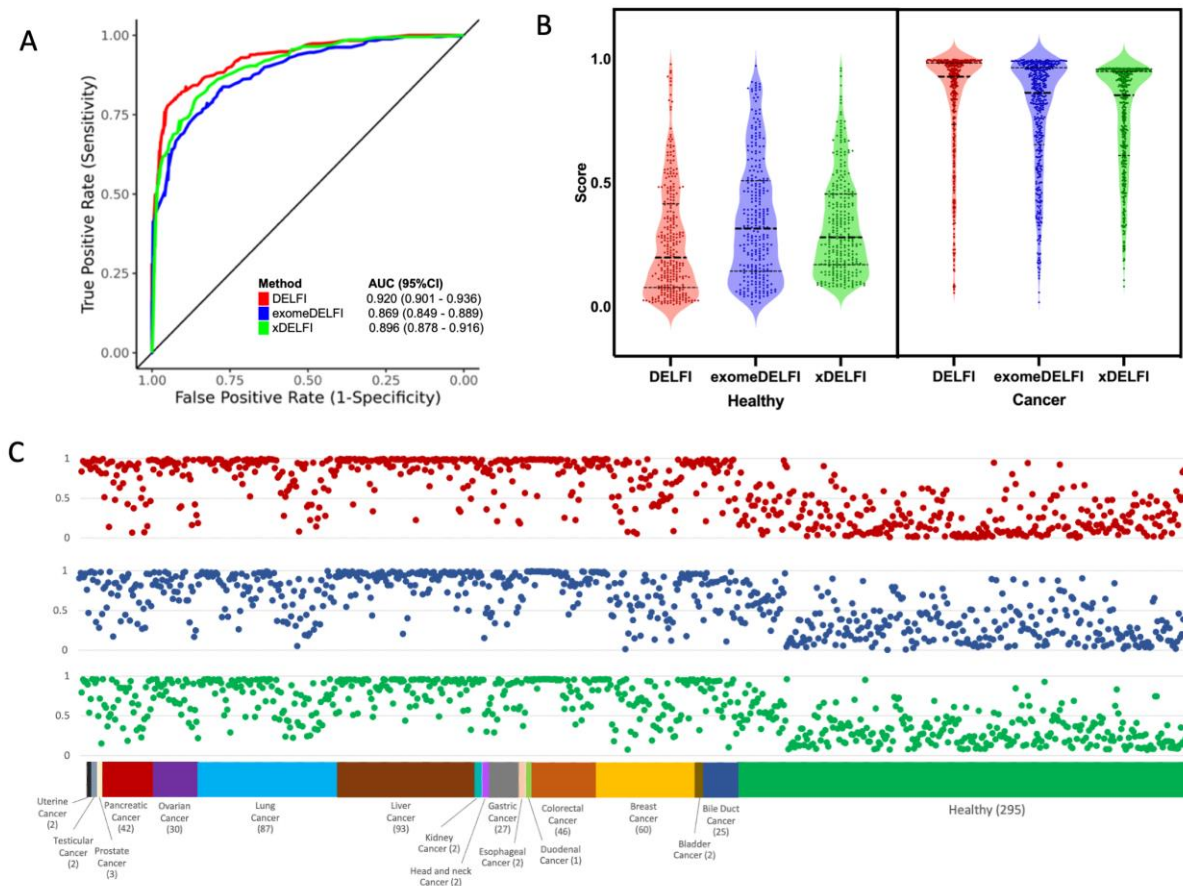
156 **Results**

157



158

159 **Figure 1** Schema of xDELFI calculation (A) general procedure of cell-free DNA process
160 consists of blood collection, cfDNA extraction, library preparation for whole-genome sequencing
161 or whole-exome sequencing, sequencing using next-generation sequencing platforms, and data
162 preprocessing. (B) xDELFI feature extraction contains three fragment size count (100-150bp,
163 151-220bp, and >220bp), overall fragment count in each 100k bin along the genome, and
164 fragment size distribution of 5bp bin in in each chromosome arm. Loess regression-based
165 approach is applied to correct for GC bias and z-score is applied for normalization. (C) Gradient
166 boosting machine (GBM) is combined with support vector machine (SVM) algorithms by
167 stacking ensemble method to learn normalized fragmentation pattern and generate xDELFI
168 score.
169
170



171
172
173 **Figure 2** Prediction performance of DELFI score, and exomeDELFI and xDELFI visualized by
174 (A) Receiver operating characteristic curve (ROC) and area under the curve (AUC). Violin plot
175 (B) shows the distribution of scores across the different DELFI methods. Scatter plot (C) of the
176 scores in different DELFI methods stratified by each cancer type. The Y-axis displays the DELFI
177 scores, while the X-axis represents the cancer types. Red dots represent DELFI score, blue dots
178 represent exomeDELFI and green dot represents xDELFI.
179
180 **DELFI score based on exonic regions show good performance for cancer**
181 **detection**

182 To develop the whole exome-based DELFI model, the paired-end cfDNA whole-genome
183 sequencing samples from four different articles stored in FinaleDB database were used. In total
184 of 426 samples from 16 cancer types and 295 healthy individuals were processed
185 (Supplementary Figure S1). Initially, we examine the reproducibility of the original DELFI score
186 using short and long fragment coverage as a model feature. Reconstructing DELFI model with
187 the same parameters employed in Cristiano et al., 2019 of the present dataset was conducted and
188 showed high prediction performance with AUC score of 0.920 (95%CI = 0.901 – 0.936). The
189 DELFI scores have a strong Pearson correlation coefficient (r) with those published in the
190 original publication ($r = 0.855$) (Supplementary Figure S2), demonstrating high reproducibility
191 of the DELFI score.

192 To observe the potential of using whole exome for DELFI score calculation, we extracted
193 the cfDNA fragments that localized in only exon region. With the same machine learning
194 procedure as in DELFI, we observed the good performance of DELFI score based on the whole
195 exome, hereafter called exomeDELFI, with AUC of 0.869 (95%CI = 0.849-0.889) (Figure 2A
196 and Figure 2B). Although the exomeDELFI performed significantly worse than DELFI model
197 (One sample t-test $P < 0.001$), the sequence coverage for constructing the model was
198 dramatically decreased from 2.8X whole-genome coverage to 0.08X on average (97% decrease).
199 In addition, the trend of the score is highly correlated with DELFI score ($r = 0.805$) (Figure 2C
200 and Supplementary Table S2). These results suggest that exome region can be used to calculate
201 DELFI score with high prediction performance, and it paves the way in using targeted regions of
202 whole-exome sequencing data as a source of DELFI score calculation.

203 To investigate whether the fragmentomic profile from the exome regions hold distinct
204 informative features compared to the profile from whole genome, we conducted a performance

205 comparison between the exomeDELFI model and the DELFI model constructed using an
206 equivalent proportion of sequencing reads. We sampled the cfDNA fragments along the genome
207 in each sample to be equal to 0.08x whole-genome coverage and used these fragments to
208 construct the DELFI model. Interestingly, the reduced DELFI model (AUC = 0.839; 95%CI
209 =0.816 – 0.862) performed significantly worse than the exomeDELFI model (One sample t-test
210 $P < 0.001$). The fragmentomic profile of short and long fragments along the genome in the
211 reduced DELFI model was highly correlated with those used in the original DELFI model ($r =$
212 0.654). On the other hand, fragmentomic profile of short and long fragments used in the
213 exomeDELFI model have no correlation with those used in the original DELFI model ($r = -$
214 0.030) (Table 1). These findings suggest that exomeDELFI relies on distinct genomic features
215 differing from the features utilized by the original DELFI model to differentiate between cancer
216 patients and healthy individuals.

217

218

219 **Table 1** Correlation between whole genome-based fragmentomic profile and whole exome-based
220 fragmentomic profile

221

Method	Source of fragmentomic profile	Whole genome coverage	AUC (95% CI)	Accuracy (95% CI)	r^*
DELFI	Whole genome	2.8X	0.920 (0.901 - 0.936)	0.840 (0.831- 0.848)	1

DELFI	Whole genome	0.08X	0.839 (0.816 - 0.862)	0.755 (0.744 - 0.764)	0.654
exomeDELFI	Exome	0.08X	0.869 (0.849 - 0.889)	0.787 (0.777 - 0.796)	-0.031

222 *Pearson correlation coefficient against whole-genome fragmentomic profile

223 **Improvement of exome-based DELFI score**

224 To improve the prediction performance of exome-based DELFI score, we developed a
225 new feature extraction strategies and redesigned the model structure (Figure 1). Firstly, we have
226 included large cfDNA fragments (>220 bp) in the calculation, as these fragments have been
227 found in the blood and may have been released from necrotic cells [24, 25]. Thus, we applied
228 three fragment size thresholds, which are short (100-150 bp), medium (150-220 bp), and long
229 (>220 bp), to count for number of fragments in each 100k bin along the genome, instead of using
230 short (100-150 bp) and long (150-220 bp) threshold in previous study. Secondly, we included
231 total number of fragments in each 100k bin in the model, as numerous studies have reported the
232 predictive power of cfDNA concentration as a biomarker in cancer diagnosis [3, 26]. Thirdly, we
233 calculated the fragment size distribution of 5 bp bins in each chromosome arm, based on its
234 efficient prediction power from previous study [27]. Finally, we have utilized a more advanced
235 algorithm, stacking ensemble, to learn the fragmentation pattern, as it has successfully improved
236 the prediction performance of the cfDNA model [27-29]. The concept of stacking ensemble is to
237 use unique advantage of different types of ML models, each of which can learn some part of the
238 problem, to generate base models. Another model is then used to learn from the output of these
239 base models for the same problem, leading to the improvement of overall performance. In our

240 model, two different ML algorithms, GBM and SVM, were combined as a stacking ensemble
241 model for the classification.

242 By applying these strategies, we observed the improvement of the exome-based DELFI
243 score with AUC of 0.896 (95%CI = 0.878 - 0.916) (Figure 2), hereafter called xDELFI. A
244 receiver operator characteristic (ROC) curve indicated the improvement of both sensitivity and
245 specificity (Figure 2 and Supplementary Table S3). By using threshold at 90% specificity,
246 xDELFI showed better sensitivity of 74% in comparison with exomeDELFI which had
247 sensitivity of 69%. We also observed the strong correlation between xDELFI score and the
248 whole genome-based DELFI score ($r = 0.785$) (Figure 2C and Supplementary Table S2)
249 indicating the consistent relationship between both scores.

250 In order to assess the importance of new features, the model was retrained using a leave-
251 one-feature-out approach. This involves iteratively removing one type of feature at a time before
252 training the model. The contribution of the long fragment feature (>220 bp) and the overall
253 fragment coverage to the prediction performance of xDELFI was the lowest (Supplementary
254 Table S4). On the other hand, the ensemble stacking algorithm had the greatest impact on the
255 prediction performance, followed by FSD.

256 The model prediction performance on stratifying cancer type revealed that the
257 improvement of xDELFI was observed in almost all cancer types in comparison with the
258 prediction performance of exomeDELFI (Supplementary Table S5-S7). Prediction Performance
259 on the three most abundant cancer types in the dataset was comparable between DELFI and
260 xDELFI model. At 95% specificity, DELFI sensitivity performance on liver cancer (n=93), lung
261 cancer (n=87), and breast cancer (n=60) are 88%, 76% and 57%, respectively, while xDELFI are
262 87%, 75%, and 53%, respectively. Similarly, the performance on the stratification of cancer

263 stage showed that the improvement of xDELFI was observed in all cancer stages (Supplementary
264 Figure S4 and Table S8). Especially at the most difficult stage I, the exomeDELFI achieved only
265 53% of sensitivity score at 90% specificity threshold, while xDELFI was able to achieve 60% of
266 the sensitivity score.

267 To further enhance the sensitivity of cancer detection, we conducted an evaluation of the
268 potential benefits of combining DELFI scores with mutation detection approach. To evaluate
269 this, we analyzed mutation data from a dataset comprising 125 cancer samples [30], and initially
270 found that the targeted mutation approach had a sensitivity of 0.656 for cancer detection. When
271 the exomeDELFI score was combined with the mutation data using the "or" condition, we
272 noticed a significant enhancement in sensitivity from 0.424 to 0.744 (Table 2 and Supplementary
273 Table S1). Additionally, the use the xDELFI score together with the mutation data improved the
274 sensitivity from 0.576 to 0.808. These findings are noteworthy because the standard sequencing
275 coverage of WES is sufficient to detect mutations [31]. Thus, it is possible to calculate the
276 xDELFI score and obtain the mutation profile from WES data, which can be used together to
277 predict cancer patients with higher accuracy than the DELFI approach alone.

278

279

280 **Table 2** Detection of 125 cancer patients using different DELFI models at 95%Specificity
281 threshold and targeted mutation cfDNA approach

Method	TP	FN	Sensitivity
Mutations	82	43	0.656
DELFI	93	32	0.744
exomeDELFI	53	72	0.424

xDELFI	72	53	0.576
DELFI + Mutations	108	17	0.864
exomeDELFI+ Mutations	93	32	0.744
xDELFI+ Mutations	101	24	0.808

282

283

284

285 **xDELFI can predict tissue of origin**

286 One potential application of cfDNA fragmentation profiles is the ability to indicate the
287 tissue of origin. We developed a multiclass machine learning model that can classify the tissue of
288 origin for eight types of cancer, including bile duct, breast, colorectal, gastric, liver, lung,
289 ovarian, and pancreatic cancers. It is important to note that we excluded cancer types with a
290 small sample size (<20) from the model. The performance of the multiclass xDELFI model was
291 comparable to that of the multiclass exome DELFI model, with a mean balanced accuracy of
292 0.676 (95% CI= 0.598 – 0.754) and 0.666 (95% CI= 0.596 – 0.736), respectively. There was no
293 significant difference between the two models, as determined by a one-sample t-test ($P = 0.067$)
294 (Table 3). However, xDELFI showed greater mean prediction sensitivity with a marginal
295 significant difference (One sample t-test $P = 0.054$) than the exomeDELFI model, although the
296 mean prediction specificity was not different. As expected, the DELFI model outperformed both
297 xDELFI and exome DELFI in all evaluation metrics (One sample t-test $P < 0.001$). The
298 performance of all models was higher than that of random class assignments (Binomial test $P <$
299 0.001). Assessment of the prediction performance of each class showed that all the models had
300 high specificity but lower sensitivity. Notably, liver, colorectal, and lung cancers had the highest
301 prediction accuracy in all models, likely due to their larger sample sizes, highlighting the

302 importance of sample size in multiclass models (Supplementary Table S9). These findings
303 suggest that the DELFI score based on whole-exome regions can also be used to predict the
304 tissue of origin.

305

306

307 **Table 3** Overall prediction performance of different cfDNA fragmentation models, DELFI,
308 exomeDELFI, and xDELFI on tissue of origin classification

309

Method	Mean balanced Accuracy (95%CI)	Mean accuracy (95%CI)	Accuracy of random assignment (95%CI)	P-Value (Acc > Random)	AUC (95%CI)	Mean Sensitivity (95%CI)	Mean Specificity (95%CI)
exomeDELFI	0.666 (0.596 – 0.736)	0.536 (0.420 - 0.652)	0.213 (0.201 - 0.226)	< 0.001	0.823 (0.745 – 0.902)	0.403 (0.278 – 0.527)	0.930 (0.912 – 0.947)
xDELFI	0.676 (0.598 – 0.754)	0.547 (0.429 - 0.666)	0.219 (0.206 - 0.232)	< 0.001	0.829 (0.762 – 0.896)	0.422 (0.283 – 0.561)	0.931 (0.913 – 0.949)
DELFI	0.741 (0.657 – 0.825)	0.638 (0.517 - 0.759)	0.202 (0.190 - 0.214)	< 0.001	0.886 (0.818 – 0.954)	0.536 (0.386 – 0.687)	0.946 (0.928 – 0.964)

310

311 Discussion

312 In our study, we have shown that it is feasible to develop a DELFI model using exome
313 regions. This approach generally offers a significant reduction in sequencing cost, while
314 maintaining a reasonable prediction performance. Interestingly, we observed a notable
315 distinction in the feature profile between the WGS-based DELFI and exome-based DELFI,
316 despite the prediction trend being similar. This finding suggests the potential for novel features
317 that can be incorporated into the existing DELFI model to improve prediction accuracy.

318 We demonstrated that implementing a new feature extraction scheme and utilizing more
319 advanced algorithms can significantly improve the accuracy of the exome-based DELFI model.
320 While short (100-150 bp) and long (151-220 bp) cfDNA fragment coverage within each window
321 were the key features utilized in the DELFI model, the model performance can be enhanced by
322 incorporating new fragmentomic features such as fragmentation size coverage (FSC),
323 fragmentation size distribution (FSD), and employing ensemble stacking algorithms [28]. With
324 these strategies, we can also improve exome-based DELFI performance in xDELFI. Further
325 analysis using leave-one-feature-out approach to determine feature importance revealed that the
326 long fragment feature (>220 bp) had the lowest contribution to the prediction performance of
327 xDELFI. In contrast, the ensemble stacking algorithm was found to have the most contribution to
328 the prediction performance, followed by FSD and overall coverage. These findings suggest that
329 exome-based DELFI performance can be improved without incurring additional sequencing
330 costs. According to suggestion in Liu's 2021 review [32], new feature extraction strategies
331 should be developed to further enhance DELFI performance, for example, three-dimensional
332 chromatin organization [33], and cfDNA- accessibility score near the transcription factor-
333 binding sites [34]. Additionally, several studies have demonstrated a correlation between
334 cfDNA-fragmentation patterns at the transcription start site (TSS) and gene expression [14, 35,
335 36]. Thus, incorporating WES plus untranslated region (UTR) could potentially provide unique
336 features and improve the prediction performance of DELFI.

337 The DELFI, exomeDELFI, and xDELFI approaches produce scores that are highly
338 correlated with each other. However, the fragmentation patterns observed in whole genome and
339 whole exome are different suggesting that the fragmentation pattern is not uniformly distributed
340 across all regions. The exonic regions, which are regions of DNA that encode proteins, holds

341 specific fragmentation pattern different from the pattern of non-exonic regions, and can influence
342 the prediction performance of the model. These findings are consistent with previous research
343 showing that exonic and intronic regions have different expression and chromatin patterns [37].

344 The performance of the DELFI score has been shown to be enhanced when combined
345 with mutation detection results [15]. However, implementing two sequencing services, one for
346 DELFI score and the other for targeted mutation detection, can be expensive and may not be
347 feasible in practical scenarios [38]. Thus, calculating xDELFI score from WES technology is an
348 alternative approach with comparable accuracy to DELFI score at a more affordable cost. WES
349 can provide mutation information that can be utilized in together with the xDELFI score to
350 improve prediction sensitivity up to 81%. However, it is important to conduct a systemic study to
351 validate the actual performance of this approach with actual WES data, which may subject to
352 significant bias caused by exome capture probs [39]. Therefore, a new method for correcting
353 prob-bias is required. In summary, our findings suggest that the use of exome regions is a viable
354 alternative for developing the DELFI score, given its reasonable accuracy and affordable cost.

355

356

357

358 **Data Availability Statement**

359 All code and data generated or analyzed during this study are deposited in Github:

360 <https://github.com/asangphukieo/xDELFI>.

361 **Conflict of Interest Statement**

362 The authors state no conflict of interest.

363 **Acknowledgments**

364 This work was supported by Faculty of Medicine Research Fund, Chiang Mai University under
365 award number 031-2566. This work was supported by the Google Cloud Research Credits
366 program with the award GCP19980904.

367 **Author Contributions Statement**

368 A.S. and P.C. conceived of the presented idea. A.S. developed the theory, designed the model
369 and the computational framework, and analysed the data. P.N. and P.T. verified the analytical
370 methods. P.C. helped supervise the project. A.S. wrote the manuscript in consultation with P.C.,
371 P.T. and P.N. All authors discussed the results and contributed to the final manuscript.

372

373 **Reference**

374

- 375 1. Cisneros-Villanueva, M., et al., *Cell-free DNA analysis in current cancer clinical trials: a review*. Br J Cancer, 2022. **126**(3): p. 391-400.
- 376 2. Dang, D.K. and B.H. Park, *Circulating tumor DNA: current challenges for clinical utility*. J Clin Invest, 2022. **132**(12).
- 377 3. Dziadziuszko, R., et al., *Circulating Cell-free DNA as a Prognostic Biomarker in Patients with Advanced ALK+ Non-small Cell Lung Cancer in the Global Phase III ALEX Trial*. Clin Cancer Res, 2022. **28**(9): p. 1800-1808.
- 378 4. Yao, W., et al., *Evaluation and comparison of in vitro degradation kinetics of DNA in serum, urine and saliva: A qualitative study*. Gene, 2016. **590**(1): p. 142-8.
- 379 5. Fiala, C. and E.P. Diamandis, *Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection*. BMC Med, 2018. **16**(1): p. 166.
- 380 6. Lo, Y.M., et al., *Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus*. Sci Transl Med, 2010. **2**(61): p. 61ra91.
- 381 7. Jiang, P., et al., *Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients*. Proc Natl Acad Sci U S A, 2015. **112**(11): p. E1317-25.
- 382 8. Mouliere, F., et al., *Enhanced detection of circulating tumor DNA by fragment size analysis*. Sci Transl Med, 2018. **10**(466).
- 383 9. Udomruk, S., et al., *Characterization of Cell-Free DNA Size Distribution in Osteosarcoma Patients*. Clin Cancer Res, 2023. **29**(11): p. 2085-2094.
- 384 10. Underhill, H.R., et al., *Fragment Length of Circulating Tumor DNA*. PLoS Genet, 2016. **12**(7): p. e1006162.
- 385 11. Mouliere, F., et al., *Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients*. EMBO Mol Med, 2018. **10**(12).

- 398 12. Lo, Y.M.D., et al., *Epigenetics, fragmentomics, and topology of cell-free DNA in liquid*
399 *biopsies*. *Science*, 2021. **372**(6538).
- 400 13. Mathios, D., et al., *Detection and characterization of lung cancer using cell-free DNA*
401 *fragmentomes*. *Nat Commun*, 2021. **12**(1): p. 5060.
- 402 14. Ivanov, M., et al., *Non-random fragmentation patterns in circulating cell-free DNA*
403 *reflect epigenetic regulation*. *BMC Genomics*, 2015. **16 Suppl 13**(Suppl 13): p. S1.
- 404 15. Cristiano, S., et al., *Genome-wide cell-free DNA fragmentation in patients with cancer*.
405 *Nature*, 2019. **570**(7761): p. 385-389.
- 406 16. Snyder, M.W., et al., *Cell-free DNA Comprises an In Vivo Nucleosome Footprint that*
407 *Informs Its Tissues-Of-Origin*. *Cell*, 2016. **164**(1-2): p. 57-68.
- 408 17. Ding, S.C. and Y.M.D. Lo, *Cell-Free DNA Fragmentomics in Liquid Biopsy*. *Diagnostics*
409 (Basel), 2022. **12**(4).
- 410 18. Zhang, X., et al., *Ultrasensitive and affordable assay for early detection of primary liver*
411 *cancer using plasma cell-free DNA fragmentomics*. *Hepatology*, 2022. **76**(2): p. 317-329.
- 412 19. Adalsteinsson, V.A., et al., *Scalable whole-exome sequencing of cell-free DNA reveals*
413 *high concordance with metastatic tumors*. *Nat Commun*, 2017. **8**(1): p. 1324.
- 414 20. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in*
415 *Saccharomyces cerevisiae*. *Nature*, 2000. **403**(6770): p. 623-7.
- 416 21. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer*. *Nucleic*
417 *Acids Res*, 2019. **47**(D1): p. D941-D947.
- 418 22. Esfahani, M.S., et al., *Inferring gene expression from cell-free DNA fragmentation*
419 *profiles*. *Nat Biotechnol*, 2022. **40**(4): p. 585-597.
- 420 23. Sun, K., et al., *Orientation-aware plasma cell-free DNA fragmentation analysis in open*
421 *chromatin regions informs tissue of origin*. *Genome Res*, 2019. **29**(3): p. 418-427.
- 422 24. Hahn, T., K.S. Drese, and C.K. O'Sullivan, *Microsystem for isolation of fetal DNA from*
423 *maternal plasma by preparative size separation*. *Clin Chem*, 2009. **55**(12): p. 2144-52.
- 424 25. Mouliere, F., et al., *High fragmentation characterizes tumour-derived circulating DNA*.
425 *PLoS One*, 2011. **6**(9): p. e23418.
- 426 26. Yu, D., et al., *Diagnostic Value of Concentration of Circulating Cell-Free DNA in Breast*
427 *Cancer: A Meta-Analysis*. *Front Oncol*, 2019. **9**: p. 95.
- 428 27. Ma, X., et al., *Multi-dimensional fragmentomic assay for ultrasensitive early detection of*
429 *colorectal advanced adenoma and adenocarcinoma*. *J Hematol Oncol*, 2021. **14**(1): p.
430 175.
- 431 28. Wang, S., et al., *Multi-Dimensional Cell-free DNA Fragmentomic Assay for Detection of*
432 *Early-Stage Lung Cancer*. *Am J Respir Crit Care Med*, 2022.
- 433 29. Bao, H., et al., *Letter to the Editor: An ultra-sensitive assay using cell-free DNA*
434 *fragmentomics for multi-cancer early detection*. *Mol Cancer*, 2022. **21**(1): p. 129.
- 435 30. Phallen, J., et al., *Direct detection of early-stage cancers using circulating tumor DNA*.
436 *Sci Transl Med*, 2017. **9**(403).
- 437 31. Manier, S., et al., *Whole-exome sequencing of cell-free DNA and circulating tumor cells*
438 *in multiple myeloma*. *Nat Commun*, 2018. **9**(1): p. 1691.
- 439 32. Liu, Y., *At the dawn: cell-free DNA fragmentomics and gene regulation*. *Br J Cancer*,
440 2022. **126**(3): p. 379-390.
- 441 33. Liu, Y., et al., *Abstract 5177: Spatial co-fragmentation pattern of cell-free DNA*
442 *recapitulates in vivo chromatin organization and identifies tissue-of-origin*. *Cancer*
443 *Research*, 2019. **79**(13_Supplement): p. 5177-5177.

- 444 34. Ulz, P., et al., *Inference of transcription factor binding from cell-free DNA enables tumor*
445 *subtype prediction and early detection*. Nat Commun, 2019. **10**(1): p. 4666.
- 446 35. Ulz, P., et al., *Inferring expressed genes by whole-genome sequencing of plasma DNA*.
447 Nat Genet, 2016. **48**(10): p. 1273-8.
- 448 36. Han, B.W., et al., *Noninvasive inferring expressed genes and in vivo monitoring of the*
449 *physiology and pathology of pregnancy using cell-free DNA*. Am J Obstet Gynecol, 2021.
450 **224**(3): p. 300 e1-300 e9.
- 451 37. Wilhelm, B.T., et al., *Differential patterns of intronic and exonic DNA regions with*
452 *respect to RNA polymerase II occupancy, nucleosome density and H3K36me3 marking in*
453 *fission yeast*. Genome Biol, 2011. **12**(8): p. R82.
- 454 38. Vanderpoel, J., et al., *Total cost of testing for genomic alterations associated with next-*
455 *generation sequencing versus polymerase chain reaction testing strategies among*
456 *patients with metastatic non-small cell lung cancer*. J Med Econ, 2022. **25**(1): p. 457-468.
- 457 39. Krumm, N., et al., *Copy number variation detection and genotyping from exome*
458 *sequence data*. Genome Res, 2012. **22**(8): p. 1525-32.
- 459

460

461 **Supplementary data**

462 **Figure S1** Proportion of cancer patients in each type and healthy individuals in the dataset

463 **Figure S2** Correlation between sequence coverage from original article and from this study

464 **Figure S3** Correlation between DELFI score reported in 2019 and DELFI score reported in this
465 study

466 **Figure S4** Violin plot of DELFI, exomeDELFI, and xDELFI score in each cancer stage

467 **Table S1** DELFI score, exomeDELFI score and xDELFI score of 721 samples

468 **Table S2** Pearson's correlation coefficient of different DELFI scores

469 **Table S3** Prediction performance of different cfDNA fragmentation models, DELFI,
470 exomeDELFI, and xDELFI

471 **Table S4** Leave-one-out prediction performance on feature type of xDELFI model at 95%
472 specificity threshold

473 **Table S5** DELFI prediction performance of 721 samples on the stratification of cancer types

474 **Table S6** ExomeDELFI prediction performance of 721 samples on the stratification of cancer
475 types

476 **Table S7** xDELFI prediction performance of 721 samples on the stratification of cancer types

477 **Table S8** Prediction performance of different cfDNA fragmentation models, DELFI,
478 exomeDELFI, and xDELFI on the stratification of cancer stage

479 **Table S9** Prediction performance of each class of different cfDNA fragmentation models,
480 DELFI, exomeDELFI, and xDELFI on tissue of origin classification

481

482

483