Depletion-assisted multiplexed cell-free RNA sequencing reveals distinct human and microbial signatures in plasma versus extracellular vesicles

Hongke Wang[1,2,†], Qing Zhan[1,2,†], Meng Ning[3,†], Hongjie Guo[4,†], Qian Wang[5], Jiuliang Zhao[5], Pengfei Bao[1,2,6], Shaozhen Xing[1,2], Shanwen Chen[4], Shuai Zuo[4], Mengtao Li[5,*], Pengyuan Wang[4,*], Zhi John Lu[1,2,*]

[1] MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China

[2] Institute for Precision Medicine, Tsinghua University, Beijing 100084, China

[3] Tianjin Third Central Hospital, Jintang Road, Hedong District, Tianjin, China

[4] Translational Cancer Research Center, Division of General Surgery, Peking University First Hospital, Beijing 100034, China

[5] Department of Rheumatology and Clinical Immunology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, National Clinical Research Center for Dermatologic and Immunologic Diseases (NCRC-DID), MST State Key Laboratory of Complex Severe and Rare Diseases, MOE Key Laboratory of Rheumatology and Clinical Immunology, Beijing 100730, China

[6] Peking University–Tsinghua University–National Institute of Biological Sciences Joint Graduate Program, School of Life Sciences, Tsinghua University, Beijing, China

[*] To whom correspondence should be addressed: Zhi John Lu, Email: zhilu@tsinghua.edu.cn; Pengyuan Wang, Email: pengyuan_wang@bjmu.edu.cn; Mengtao Li, Email: mengtao.li@cstar.org.cn.

[†] These authors contributed equally to this work.

Abstract

Cell-free long RNAs in human plasma and extracellular vesicles (EVs) have shown promise as biomarkers in liquid biopsy, despite their fragmented nature. To investigate these fragmented cell-free RNAs (cfRNAs), we developed a cost-effective cfRNA sequencing method called DETECTOR-seq (depletion-assisted multiplexed cell-free total RNA sequencing). DETECTOR-seq utilized a meticulously tailored set of customized guide RNAs to remove large amounts of unwanted RNAs (i.e., fragmented ribosomal and mitochondrial RNAs) in human plasma. Early barcoding strategy was implemented to reduce costs and minimize plasma requirements. Using DETECTOR-seq, we conducted a comprehensive analysis of cell-free transcriptomes in both whole human plasma and EVs. Our analysis revealed discernible distributions of RNA types in plasma and EVs. Plasma exhibited pronounced enrichment in structured circular RNAs, tRNAs, Y RNAs, and viral RNAs, while EVs showed enrichment in mRNAs and srpRNAs. Functional pathway analysis highlighted RNA splicing-related ribonucleoproteins (RNPs) and antimicrobial humoral response genes in plasma, while EVs demonstrated enrichment in transcriptional activity, cell migration, and antigen receptor-mediated immune signals. Our study indicates the comparable potential of cfRNAs from whole plasma and EVs in distinguishing cancer patients (i.e., colorectal and lung cancer) from healthy donors. And microbial cfRNAs in plasma showed potential in classifying specific cancer types. Our comprehensive analysis of total and EV cfRNAs in paired plasma samples provides valuable insights for determining the need for EV purification in cfRNA-based studies. We

46 envision the cost-effectiveness and efficiency of DETECTOR-seq will empower
47 transcriptome-wide investigations in the fields of extracellular vesicles and liquid biopsy.

48 **Keywords:** Cell-free RNA, Extracellular vesicles, Cancer classification, Liquid biopsy,

49

50 Introduction

51 In recent years, liquid biopsy has emerged as a non-invasive approach for assessing
52 circulating biomarkers in various body fluids, enabling the monitoring of physiologic and
53 disease states [1]. Cell-free RNAs (cfRNAs), given their virtue of being highly dynamic,
54 hold great potential to reflect the pathophysiological processes, thus offering unique
55 opportunities for disease monitoring. Previous reports have suggested that cfRNAs are
56 packaged into various extracellular complexes, such as extracellular vesicles (EVs,
57 including micro-vesicles and exosomes) and non-vesicular ribonucleoproteins (RNPs) [2].
58 Due to the protection of EV, RNA binding proteins, and/or their self-structures, cfRNAs are
59 capable of being stably present in human bloodstream [3]. While previous studies have
60 predominantly focused on total cfRNAs [4-6] or EV [7-9] cfRNAs in plasma, the
61 transcriptional differences between these two entities remain poorly understood.

62 Efforts to characterize cfRNAs initially centered around small RNAs like microRNAs
63 (miRNAs) because of the nature of RNA degradation and fragmentation in biofluids.
64 However, miRNAs represent only a small proportion of the human transcriptome [10].
65 Consequently, investigations have expanded to encompass a broader range of cfRNA
66 species, including messenger RNAs (mRNAs), long non-coding RNAs (lncRNAs), and

67 circular RNAs (circRNAs) [4-7, 11]. These cell-free long RNA species (>50 nt) have
68 relatively low concentrations in human blood due to the presence of RNases, and they are
69 typically fragmented (~50–200 nucleotides) with incomplete RNA ends [12]. Conventional
70 small RNA-seq approaches, which rely on ligating sequencing adapters based on the 5'
71 phosphate (5' P) and 3' hydroxyl (3' OH) ends of RNA, are inadequate for analyzing these
72 fragmented cfRNAs [13].

73 Recently, several sequencing approaches have been developed to profile cell-free
74 long RNA fragments in plasma or EVs. Phospho-RNA-seq integrates T4 polynucleotide
75 kinase into ligation-based TruSeq small RNA-seq, enabling the recovery of mRNA and
76 lncRNA fragments lacking 5' P and/or 3' OH ends. However, the libraries generated by
77 Phospho-RNA-seq contain high proportions of ribosomal RNAs (rRNAs) and Y RNAs,
78 limiting the capacity to detect other informative RNA species [12]. Another method,
79 SILVER-seq, captures both small and long cfRNAs from extremely low-input serum
80 samples [14]. However, substantial DNA contamination seemed to be an issue of SILVER-
81 seq [15]. Recently, SMARTer stranded total RNA-seq (hereafter called SMARTer-seq) has
82 been employed in several cfRNA studies [4-7], utilizing a proprietary ZapR and R-probes
83 to deplete unwanted ribosomal sequences [16, 17]. However, as a commercial kit,
84 SMARTer-seq is not specifically optimized for cfRNA library preparation from plasma and
85 is cost-inefficient. Overall, the current cfRNA sequencing approaches were hindered by
86 unwanted RNAs, DNA contamination, and high cost.

87 Current targeted depletion strategies for unwanted RNAs, such as RiboMinus kits

88 (Thermo Fisher Scientific), Ribo-Zero technology (Illumina), and RNase H-mediated
89 digestion of RNA:DNA hybrids [16], primarily operate at the RNA level and require
90 relatively intact RNA molecules. Consequently, these methods are unsuitable for low-input
91 and fragmented cfRNA samples. In contrast, the Cas9-mediated targeted DNA cleavage
92 technique, also known as DASH (depletion of abundant sequences by hybridization) [18],
93 provides the capability to selectively cleave cDNA molecules derived from rRNAs during
94 the double-stranded DNA stage after library amplification. Notably, this method only
95 requires the design of a set of specific single-stranded guide RNAs (sgRNAs) to direct
96 Cas9 cleavage of undesirable sequences. Therefore, CRISPR-Cas9 presents a highly
97 advantageous approach for the targeted removal of over-represented sequences in the
98 libraries of low-input and fragmented cfRNA samples derived from plasma and EVs.

99 In this study, we present an optimized cfRNA sequencing method, DETECTOR-seq
100 (depletion-assisted multiplexed cell-free total RNA sequencing), which utilizes early
101 barcoding and CRISPR-Cas9 to reduce costs and deplete highly abundant, fragmented
102 rRNAs and mitochondrial RNAs (mtRNAs) in human plasma. Subsequently, we used
103 DETECTOR-seq to investigate 113 plasma cfRNA samples (including 61 plasma total RNA
104 and 52 EV RNA libraries), derived from healthy donors, lung and colorectal cancer patients.
105 To the best of our knowledge, this study is the first to compare paired total and EV-selected
106 transcriptomes in the same plasma samples, suggesting their distinct signatures and
107 different utilities in cancer liquid biopsy.

108 **Results**

109 **Development of DETECTOR-seq to profile cell-free transcriptome**

110 The sequencing of cfRNAs in plasma and extracellular vesicles (EVs) usually meets the
111 following obstacles. First, consistent with previous reports [10], we observed that plasma
112 cfRNAs were degraded with a fragment length of <200 nucleotides **(Figure 1A)**. These
113 fragmented cfRNAs are hard to be detected by RNA-seq protocols based on ligation
114 techniques requiring intact RNA ends. Second, ribosomal RNAs (rRNAs) and
115 mitochondrial RNAs (mtRNAs) accounted for ~92% of all clean reads (reads after
116 removing adapters and filtering low-quality reads), while messenger RNAs (mRNAs) and
117 long non-coding RNAs (lncRNAs) collectively made up only a small fraction (~4%) of cell-
118 free transcriptome **(Figure 1B)**. It is worth noting that microbe-derived RNAs can also be
119 detected in human plasma with a relatively small fraction (~0.4%) **(Figure 1B)**. The high
120 fractions of rRNAs and mtRNAs hamper the detection of other informative RNA species.
121 And they are fragmented into pieces in plasma, making them hard to be removed **(Figures**
122 **1C, D)**. Third, cfRNAs are usually in the range of hundred picograms to several nanograms
123 per ml of human plasma [14], which can be easily lost and contaminated during purification
124 and amplification. Furthermore, low cfRNA input usually requires 20-24 PCR amplification
125 cycles for library preparation, which produces a high duplication ratio of raw reads.
126 Meanwhile, DNA contamination ignorable in conventional RNA-seq is often over-amplified,
127 causing a big issue in cfRNA-seq [15].

128 To improve the efficiency and reliability of cfRNA detection, we developed
129 DETECTOR-seq (depletion-assisted multiplexed cell-free total RNA sequencing) to profile

130 cell-free transcriptome in human plasma **(Figures 1E, F)**. DETECTOR-seq captures
131 fragmented cfRNAs with unbiased random priming and template-switching. Then, we
132 adapted and modified a previously described method termed CRISPR/Cas9–based
133 Depletion of Abundant Species by Hybridization (DASH) [18] to remove the abundant
134 sequences derived from ribosomal and mitochondrial RNAs in the complementary DNA
135 (cDNA) library. In this step, guide RNAs (sgRNAs) in the CRISPR-Cas9 are specifically
136 optimized for human plasma cfRNAs **(Supplementary Figures 1,2)**, covering the
137 fragmented rRNA and mtRNA sequences **(Figures 1D, E)**. The sgRNAs are in vitro
138 transcribed using T7 RNA polymerase, then bind with Cas9 nuclease to form
139 ribonucleoprotein (RNP) complex and induce site-specific cleavage with the endonuclease
140 activity of Cas9 **(Figure 1E)**, thus preventing further amplification of cDNAs derived from
141 rRNAs and mtRNAs in the final sequencing library. Meanwhile, DETECTOR-seq utilizes
142 early barcoding during reverse transcription. The multiplexed library will cope with low
143 content of plasma cfRNAs, and reduce experimental time and cost as well. It is also worth
144 mentioning that unique molecular identifiers (UMIs) are added to every sequence in the
145 reverse transcription step, hence DETECTOR-seq is capable of removing PCR duplicates
146 to avoid RNA quantification bias **(Figure 1F)**. In addition, we also optimized cfRNA
147 extraction **(Supplementary Figure 3)** and residual DNA digestion **(Supplementary**
148 **Figure 4)** protocols.
149

150 **Analytical validation demonstrating superior performance of DETECTOR-seq**

151 To examine whether DETECTOR-seq can deplete the unwanted rRNA and mtRNA
152 sequences effectively and specifically, we split a single plasma sample into two equal
153 aliquots for experimental conditions of untreated versus depleted, with six biological
154 replicates. In the untreated samples, reads mapped to rRNAs and mtRNAs collectively
155 represented ~94% of all mapped reads. After CRISPR-Cas9 treatment, these unwanted
156 sequences were decreased to only ~15% of mapped reads, only about one-sixth of the
157 untreated ones **(Figure 2A)**. By comparing untreated and depleted aliquots, we observed
158 evident decreases in the normalized coverage of rRNAs and mtRNAs **(Figure 2B)**.
159 Meanwhile, the expression levels of detected genes other than rRNAs and mtRNAs
160 between the untreated and depleted aliquots were well correlated, indicating minimal off-
161 target effect (Pearson correlation, R = 0.92, *P*-value < $2.2\times^{-16}$; **Figure 2C**). By comparing
162 the cfRNA expression profiles obtained from DETECTOR-seq and SMARTer-seq, we
163 found that the expression levels of detected genes using these two methods were also
164 well correlated (Pearson correlation, R = 0.90, *P*-value < $2.2\times^{-16}$; **Figure 2D**). In summary,
165 the above results demonstrate the efficient and specific depletion of unwanted sequences
166 in DETECTOR-seq.

167 To further evaluate the performance of DETECTOR-seq, we prepared cfRNA libraries
168 in a 3-plex, 4-plex, or 5-plex manner determined by RNA concentrations. The total read
169 numbers of different barcoded samples in one multiplexed pool were relatively uniform,
170 varying less than 1.5-fold in the 3- and 4-plex samples and less than 3-fold in the 5-plex
171 samples **(Supplementary Figure 5A)**. In addition, the UMI strategy in DETECTOR-seq

172 retained significantly more reads than the non-UMI approach after duplicated reads were
173 removed **(Supplementary Figure 5B)**. And a sharp edge of reads' distribution across
174 exon-intron splice junctions suggested that the majority of DNA contamination was
175 effectively removed **(Supplementary Figure 5C)**. To evaluate the impact of plasma input
176 volume on the number of detected genes, we sequenced cfRNAs with 200, 400, 600, 800,
177 and 1000 μL of plasma aliquots from the same individual with five biological replicates.
178 Around 4000 genes were detected with the minimum (i.e., 200 μL) volume. The detected
179 gene number linearly increased until a plateau between 800 and 1000 μL, suggesting the
180 detected genes would be saturated after 1 mL of plasma **(Supplementary Figure 5D)**.
181 While highly correlated cfRNA expression levels were observed within technical triplicates
182 (R1-R3), the correlations were slightly decreased between biological triplicates (N1-N3)
183 **(Supplementary Figure 5E)**. Furthermore, based on ERCC RNA Spike-In Mix, we found
184 a high correlation between expected and observed levels of transcript abundance
185 (Pearson correlation, R = 0.91, *P*-value < $2.2\times^{-16}$; **Supplementary Figure 5F**). These
186 results not only demonstrate DETECTOR-seq's high accuracy and reproducibility but also
187 suggest its capability of capturing subtle differences in cfRNA profiles between different
188 individuals.

189 Then, we randomly subsampled a dataset (n=24) of DETECTOR-seq for saturation
190 analyses of detected UMIs (transcripts) and genes. Although the detected UMIs kept
191 increasing when more reads in 1ml plasma were sequenced **(Supplementary Figure 5G)**,
192 the detected gene numbers were quickly saturated at approximately 5 million genome-

193 aligned reads **(Supplementary Figure 5H)**. These results indicate that DETECTOR-seq
194 achieves saturation of cfRNA detection at a low sequencing depth.
195
196 **Better contamination control and cost-effectiveness of DETECTOR-seq than other**
197 **cfRNA-seq methods**
198 We benchmarked the performance of DETECTOR-seq compared to three other cfRNA-
199 seq methods, including Phospho-RNA-seq [12], SILVER-seq [14], and SMARTer-seq [19].
200 The number of samples used in the comparison was listed in **Supplementary Table 5**.
201 Within the total genome-aligned reads, DETECTOR-seq and SMARTer-seq had
202 comparable ratios of exonic reads (~70%), while those of SILVER-seq and Phospho-RNA-
203 seq were under 40% **(Figure 3A)**. The lower ratio of exonic reads for SILVER-seq was
204 presumably due to severe DNA contamination according to a previous report [15]. We also
205 visualized the read coverage across exon boundary sites flanked upstream and
206 downstream by 50 bp, where DETECTOR-seq and SMARTer-seq showed more evident
207 decreases of read coverage from exon to intron/intergenic region than SILVER-seq and
208 Phospho-RNA-seq **(Figure 3B)**. As far as we know, all of the four cell-free RNA-seq
209 methods should preserve the strand specificity of RNAs. Thus, the enrichment of exons'
210 sense over antisense reads of DETECTOR-seq and SMARTer-seq further confirmed their
211 reads' quality **(Figure 3C)**. The above results demonstrate that DETECTOR-seq and
212 SMARTer-seq have better DNA contamination control than SILVER-seq. It was worth
213 noting that Phospho-RNA-seq was developed from a small RNA-seq method, and the read

214 coverage across exon boundary sites and the enrichment of exons' sense over antisense
215 reads may be affected by the read distribution of small RNAs.
216 In addition, we showed that DETECTOR-seq displayed a higher ratio of reads
217 mapped to human genome (~71%) than those of SMARTer-seq (~48%) because
218 DETECTOR-seq removed mitochondrial RNAs more efficiently than SMARTer-seq
219 **(Figure 3D)**. Furthermore, because of its early barcoding and multiplexing strategy,
220 DETECTOR-seq can produce more raw reads and genome-aligned reads than the other
221 cfRNA-seq approaches **(Figure 3E, Supplementary Figure 6)**. Cost details were
222 explained in Supplementary Tables 6 and 7. Overall, by summarizing and comparing key
223 characteristics of these approaches **(Figure 3F)**, we collectively demonstrate that
224 DETECTOR-seq has better contamination control and more efficient cost than the other
225 cfRNA-seq methods.
226
227 **Distinct human and microbial RNA signatures in plasma versus extracellular**
228 **vesicles**
229 Subsequently, we employed DETECTOR-seq to conduct pairwise investigations of total
230 cfRNAs and EV cfRNAs in human plasma **(Figure 4A)**. A proportion of cfRNAs are
231 enclosed inside EVs such as MVs and exosomes [20]. Meanwhile, it is also reported that
232 a significant proportion of cfRNAs are not within EVs but associated with proteins to form
233 non-vesicular RNPs [21]. Although both plasma total cfRNAs [4-6] and EV cfRNAs [7, 9]
234 have been used in liquid biopsy studies, a pairwise comparison of their distinct signatures

235 and utilities has not been conducted yet.

236 In this study, a total of 139 plasma cfRNA samples were sequenced, which included
237 samples obtained from healthy donors as well as patients with lung cancer and colorectal
238 cancer (**Supplementary Figure 7, Supplementary Table 10**). EVs were purified using a
239 membrane affinity column, concentrating particles predominantly within the size range of
240 50 to 200 nm, with a peak around 110 nm. Morphological examination using transmission
241 electron microscopy confirmed the presence of the characteristic cup-shaped structure
242 commonly associated with EVs **(Figure 4B)**. After conducting quality control (QC)
243 procedures on the RNA samples and sequencing data, a total of 113 datasets passed the
244 QC criteria (**Supplementary Figures 7-9**). Out of these 113 datasets, 61 were derived
245 from total cfRNA-seq of plasma, while 52 were obtained from EV cfRNA-seq of plasma.
246 Among them, 44 datasets were paired, meaning they originated from the same plasma
247 samples. In the following description, total cfRNA-seq of plasma and EV cfRNA-seq of
248 plasma will be abbreviated to *Plasma* cfRNA and *EV* cfRNA, respectively.

249 From a general view, there was a high degree of similarity between *Plasma* and *EV*
250 cfRNAs, with ~90% of aligned reads mapping to human genome and ~10% mapping to
251 microbe genomes **(Figure 4C)**. For human cfRNAs, mRNA, lncRNA, and circRNA were
252 the major RNA types. For microbial cfRNAs, the most abundant phylum was
253 *Proteobacteria*, followed by *Firmicutes* and *Actinobacteria*. The human and microbial RNA
254 compositions resembled previous reports [19, 22].

255 In addition, distinctive signatures were revealed for the first time by our pairwise

256 comparison between *Plasma* (n = 44) and *EV* (n = 44) cfRNAs (all samples were paired).
257 We first observed that *Plasma* cfRNAs had more short fragments (20~100 nt), while *EV*
258 cfRNAs had more long fragments (>100 nt) **(Supplementary Figure 10)**. We also
259 observed that structured tRNAs, Y RNAs, and circRNAs were enriched in *Plasma* cfRNAs,
260 while mRNAs and signal recognition particle RNAs (srpRNAs) were enriched in *EV*
261 cfRNAs **(Figure 4D)**. These findings align with a previous study that reported a significant
262 enrichment of tRNA and Y RNA fragments in extracellular RNPs [2]. Moreover, we also
263 found that the relative abundance of circRNAs was slightly higher in *Plasma* cfRNAs than
264 *EV* cfRNAs (median 13.6% vs. 8.8%, *P*-value < 0.0001, Wilcoxon rank sum test; **Figure**
265 **4D**, **Supplementary Figure 11)**, perhaps due to its circle-like structure resisting
266 degradation outside of EVs. We totally identified 13 circRNAs differentially enriched in
267 *Plasma* versus *EV* cfRNAs. Only one of them, hsa_circ_0048555, was enriched in EVs
268 **(Supplementary Figure 12)**. Reads mapped to the back-spliced junction were used to
269 calculate the enrichment.

270 A recent study provided a framework to infer cell types of origin of the cell-free
271 transcriptome [23]. We utilized this method and found a high similarity of the cell types of
272 origin between *Plasma* and *EV* transcriptomes **(Figure 4E)**. Platelets and erythrocytes
273 were inferred as the major origins for both *Plasma* and *EV* cfRNAs, which was in
274 agreement with the previous study [23]. Intriguingly, we found non-blood cells contributed
275 more to *EV* cfRNAs than to *Plasma* cfRNAs (*P*-value < 0.01, Wilcoxon rank sum test;
276 **Figure 4E**). Therefore, the diversities of cell types of origin (measured by Simpson's index)

277 of *EV* cfRNAs were slightly higher than those of *Plasma* cfRNAs (0.75 vs. 0.70, *P*-value <
278 0.01, Wilcoxon rank sum test; **Figure 4E**).

279 A noteworthy discovery has been made regarding the presence of RNAs originating
280 from transposable elements and other repetitive elements in the cell-free transcriptome
281 [24]. In our current investigation, we provide evidence demonstrating a significant
282 enrichment of cfRNAs derived from transposable elements (TEs) in *Plasma* cfRNAs
283 compared to *EV* cfRNAs. These transposable elements include short interspersed
284 elements (SINEs), long interspersed elements (LINEs), long interspersed elements with
285 long terminal repeats (LTR), and DNA transposons **(Figure 4F)**.

286 We also identified distinct microbe genera in *Plasma* and *EV* cfRNAs
287 **(Supplementary Figure 13)**. While there was no significant difference in the ratio of
288 microbial reads between *Plasma* and *EV* cfRNAs, we did observe a significant increase in
289 cfRNAs mapped to viral genomes in *Plasma* cfRNAs **(Figure 4G)**. Meanwhile, viruses
290 such as *Senecavirus*, *Cheravirus*, *Orthopoxvirus*, *Tenuivirus*, and *Rhadinovirus* were
291 enriched in *Plasma* cfRNAs, while *Intestinimonas*, *Mordavella*, and *Jonquetella* were
292 enriched in *EV* cfRNAs **(Supplementary Figure 13)**. In summary, the above comparison
293 results have revealed distinct molecular characteristics between *Plasma* and *EV* cfRNAs
294 in terms of fragment size, RNA species, cell types of origin, TE RNAs and microbe genera.

295

296 **Functional pathways and sequence motifs of selective *Plasma* and *EV* cfRNAs**

297 To find selective functions and motifs of cfRNAs in EVs, we identified 545 selectively

298 distributed RNAs showing significantly differential abundance between *Plasma* and *EV*
299 transcriptomes **(**|Fold-change| >1 and FDR < 0.1; **Figure 5A, Supplementary Figure 14)**.
300 Among them, 271 cfRNAs were enriched in *Plasma*, while 274 cfRNAs were enriched in
301 *EVs*. We investigated the functional roles and biological pathways of these selective
302 cfRNAs **(Figure 5B, Supplementary Figure 14)**. Based on KEGG pathway enrichment
303 analysis, we found that the selective RNAs elevated in *Plasma* were significantly enriched
304 in terms associated with RNA splicing, RNP (e.g., mRNA 5' splice site recognition, U1
305 snRNP, spliceosomal snRNP complex and Sm-like protein family complex), antimicrobial
306 and innate immune responses. Meanwhile, the selective RNAs that were enriched in *EVs*
307 were primarily associated with DNA binding transcription factor activity, focal adhesion,
308 cell-substrate junction, and T cell receptor signaling immune pathway. Notably, we also
309 found different immune pathways enriched in the selective cfRNAs of *Plasma* versus *EVs*
310 **(Figure 5B, Supplementary Figure 15)**. The organ or tissue-specific immune response
311 and antimicrobial humoral response immune signaling pathways are enriched in *Plasma*
312 cfRNAs, while defense response to other organisms, Fc receptor signaling pathway, T cell
313 receptor signaling pathway are enriched in *EV* cfRNAs **(Supplementary Figure 15)**.
314 We further investigated sequence motifs and their associated RNA binding proteins
315 (RBPs) for the selective cfRNAs **(Figure 5C, Supplementary Figure 16)**. And we found
316 that the selective cfRNAs enriched in *Plasma* contained binding motifs/sites for ABCF1, a
317 protein that plays a role in innate immune response [25]; SFPQ, a splicing factor; LARP4,
318 a La RNP; TROVE2, a Y RNA binding protein; and DKC1, a snoRNP. Meanwhile, the

319 selective cfRNAs enriched in *EVs* contained binding motifs/sites for PUM1, a protein that
320 participates in human innate immune response [26]; BCLAF1, a transcription factor;
321 HNRNPU, a transcription suppressor; PCBP1, a previously reported immune checkpoint
322 [27]; APOBEC3C, an RNA editing enzyme. These enriched motifs and their associated
323 RBPs were consistent with the biological functions of the selective cfRNAs revealed above.
324

325 **Specific cancer-related signals in *Plasma* and *EV* cfRNAs**

326 Next, we compared the potential of *Plasma* cfRNAs and *EV* cfRNAs to discriminate
327 between cancer patients and healthy individuals in a proof-of-concept cohort. We
328 sequenced cfRNAs in the plasma samples of lung cancer (LC, Plasma n = 19, EV n = 19,
329 18 of them paired) and colorectal cancer (CRC, Plasma n = 23, EV n = 19, 19 of them
330 paired) patients **(Supplementary Figure 7)**. To maximize the sample size, we merged CRC
331 and LC together as a combined cancer group. Based on differential expression analysis
332 between this combined cancer group (Plasma n = 42, EV n = 38, 37 of them paired) and
333 normal controls (NC, Plasma n = 19, EV n = 14, 7 of them paired) using the criteria of
334 $|\log_2$fold-change$|>1$ and FDR<0.05, we defined a set of cancer-relevant cfRNAs in both
335 *Plasma* and *EVs* **(Supplementary Figure 17)**. Interestingly, when we intersected the
336 cancer-relevant cfRNAs and selective cfRNAs mentioned above, we found that cancer-
337 relevant cfRNAs accounted for 59.8% (162/271) of the selectively enriched cfRNAs in
338 *Plasma*, whereas they only represented 6.9% (19/274) of the selectively enriched cfRNAs
339 in *EVs*. Therefore, cancer-relevant cfRNAs appear to be more enriched in *Plasma*'s

340 selective cfRNA fraction. **(Figure 6A).** We also found that enriched functions of these
341 cancer-relevant *Plasma* cfRNAs were termed as RNA splicing, snRNP signals, etc **(Figure**
342 **6B)**, which were consistent with the enriched pathways of *Plasma* cfRNAs revealed in
343 **Figure 5B**.

344 Based on these selectively distributed cancer-relevant cfRNAs, we endeavored to
345 discriminate cancer patients from NCs. Although the selective cfRNAs in *Plasma*
346 performed slightly better than those in *EVs* (average AUROC: 0.909 versus 0.877, **Figure**
347 **6C, Supplementary Figure 18)**, comparable performances were observed between
348 *Plasma* and *EV* cfRNAs when a large number of non-selective cfRNAs were included as
349 well (average AUROC: 0.936 versus 0.953, **Figure 6D, Supplementary Figure 19**).
350 Collectively, these results imply that the purification of EV can reveal distinct cancer
351 signals, but it has a very subtle effect on the accuracy of detection of cancer patients from
352 healthy controls.

353 We further assessed the potential of cfRNAs (human-derived only) in *Plasma* and *EV*
354 for classifying CRC from LC. Initially, neither cfRNAs in *Plasma* or *EV* exhibited strong
355 classification potential (average AUROC: 0.628 versus 0.659, **Figure 6E, Supplementary**
356 **Figure 20)**. A recent study revealed that microbe-derived cfRNAs in human plasma reflect
357 cancer-type-specific information [19]. Based on the RNA abundance of the contamination-
358 filtered microbe genera, we found the microbial cfRNAs improved the classification of
359 cancer types for both *Plasma* and *EV* cfRNAs (average AUC: 0.898 versus 0.772, **Figure**
360 **6E, Supplementary Figure 20)**.

361 Notably, the microbial cfRNAs in *Plasma* performed better than those in *EV*.
362 Consistently, we also found more cancer-type-specific features in *Plasma* cfRNAs than in
363 *EV* cfRNAs **(Figure 6F)**. We identified the microbial features recurrently showing
364 differential abundance between CRC and LC in all of the 20 bootstrap samplings. The
365 abundance of top recurrent microbe genera, along with fold-change and false discovery
366 rates were illustrated **(Figure 6G)**. For instance, we observed a higher relative abundance
367 of *Methanothrix* in CRC compared to LC using *EV* cfRNA-seq data. This is consistent with
368 a previous study reporting that *Methanothrix soehngenii* was enriched in gut microbiome
369 of CRC patients [28]. Meanwhile, many cancer-relevant viral RNAs in *Plasma* classified
370 cancer types, consistent with the observation of more viral RNAs detected in *Plasma* than
371 in *EVs* **(Figure 4F)**. For instance, *Plasma* cfRNA-seq data revealed a higher abundance
372 of *Alpha-polyomavirus* and *Beta-polyomavirus*. Supportively, some polyomaviruses were
373 also reported to be detectable in gastrointestinal tract and respiratory aspirates [29]. These
374 findings suggest that microbe-derived cfRNAs in *Plasma*, at least in this small cohort with
375 limited sample size, present promising but yet poorly investigated signatures for specific
376 cancer types. Further validation in larger cohorts is required to establish the clinical utility
377 and significance of these preliminary findings.
378

379 ## Conclusion and Discussion

380 **Conclusions**. In summary, this study introduced a depletion-assisted cost-effective cfRNA
381 profiling approach, termed DETECTOR-seq, which overcomes challenges associated with

382 low quantity and low quality of fragmented cfRNAs, over-represented rRNAs and mtRNAs,
383 DNA contamination, and high costs. Using DETECTOR-seq, we recapitulated molecular
384 characteristics of *Plasma* and *EV* cfRNAs and identified their distinct human and microbial
385 signatures, thus illustrating the gain and loss of certain cfRNA signals due to EV
386 purification. Our work provides a practical reference for researchers engaged in plasma
387 and EV cfRNA-based liquid biopsy **(Table 1)**. Moreover, we envision that DETECTOR-seq
388 would be a useful tool to facilitate further studies in the fields of extracellular RNA biology
389 and plasma or EV cfRNA-based liquid biopsy, paving the way for advancements in both
390 fundamental research and translational medicine.
391
392 **Technologies utilized and optimized in DETECTOR-seq**. Plasma cell-free
393 transcriptome remains challenging to study owing to the low quantity and quality of
394 fragmented cfRNAs [11]. Over-represented rRNA and mtRNA species [12], DNA
395 contamination [15], and high cost are still the major issues of cfRNA sequencing. Multiple
396 technologies were included in DETECTOR-seq to address these issues **(Figure 4F)**. First,
397 DETECTOR-seq captures fragmented cfRNAs with random priming and template-
398 switching strategies, which have been proven to be highly efficient in single-cell RNA-seq
399 [30]. Second, the early barcoding protocol of DETECTOR-seq enables us to prepare
400 cfRNA libraries in a multiplexed manner, thus reducing the volume of required plasma and
401 experimental costs. In fact, DETECTOR-seq is capable of detecting cfRNAs with a low
402 input volume of 0.2 to 1 mL plasma with a 2- to 6-fold cost saving compared to existing

403 approaches. Third, with UMIs tagging to cDNAs of RNA fragments, DETECTOR-seq can
404 accurately quantify the low-quantity cfRNAs. Fourth, by optimizing the procedures of RNA
405 extraction and residual DNA digestion (**Supplementary Figures 3-4**), DETECTOR-seq
406 avoids the potential contamination of genomic DNAs. Fifth, DETECTOR-seq uses
407 CRISPR-Cas9 technology to deplete rRNA and mtRNA sequences. A CRISPR-based
408 depletion strategy, DASH (Depletion of Abundant Sequences by Hybridization) [18] has
409 been utilized in other fields, such as ATAC-seq [31], small RNA-seq [32], bacterial RNA-
410 seq [33], Ribo-seq [34] and single-cell total RNA-seq [35]. Here, we applied this CRISPR-
411 based method to cfRNA sequencing and designed a specific set of sgRNAs for human
412 plasma **(Supplementary Figures 1,2)**. Of note, our sgRNAs target almost the entire
413 length of human rRNAs and mtRNAs, enabling the use of our guides to deplete rRNAs
414 and mtRNAs from any intact or fragmented RNA samples, regardless of the specimen
415 type. This underscores the versatility of our approach beyond plasma samples.

416 ***Plasma* vs. *EV* in cancer detection and cancer type classification**. Researchers
417 have used both *Plasma* cfRNA-seq [4-6] and *EV* cfRNA-seq [7, 20, 36-38] to identify
418 disease biomarkers. By pairwise comparison between *Plasma* and *EV* cfRNA, we found
419 that both of them can distinguish cancer patients from controls with comparable
420 performance. However, cancer types can be better classified with microbe-derived
421 features in *Plasma* cfRNAs than those in *EV* cfRNAs.

422 **Distinct signatures in *Plasma* vs. *EV* cfRNAs.** This study has brought new insights
423 into distinct cfRNA signatures in *Plasma* versus *EVs*. *Plasma* contains miscellaneous

424 cfRNAs released from alive or apoptotic cells, while RNAs in *EV* cargos are considered to
425 be secreted actively by cells for functional roles in intercellular communications [39]. This
426 study revealed distinct biological pathways, enriched motifs, and RBP-binding sites in
427 *Plasma* vs. *EV* cfRNAs. We also found that short RNA fragments (20 to 100 nt) associated
428 with RNPs were enriched in *Plasma* cfRNAs, indicating higher degradation extent of non-
429 vesicular RNAs than those of EV RNAs.

430 **Limitations of this study**. While analyzing paired plasma samples can increase
431 statistical power, it is important to note that the conclusions regarding the comparison of
432 cfRNAs in *Plasma* and *EVs* for cancer differentiation in this study are still preliminary due
433 to the small sample size. These results serve as a proof-of-concept exploration of
434 DETECTOR-seq's potential for uncovering intriguing insights in real-world clinical samples.
435 Larger-scale cohorts are required to validate these findings and establish their clinical
436 utility. Furthermore, although DETECTOR-seq offers several advantages compared to
437 other approaches, there is room for further improvement. For instance, the efficiency of
438 random priming in DETECTOR-seq is influenced by the fragment length of RNAs, which
439 can introduce bias in the library preparation. And DETECTOR-seq involves several
440 purification steps to eliminate by-products such as empty library constructs, adapter
441 dimers, and excessive primers. These purification procedures tend to retain longer RNA
442 fragments, resulting in the discarding of RNA fragments shorter than 50 nucleotides, along
443 with the by-products. To obtain a complete spectrum of cfRNAs, including both small and
444 long fragments, DETECTOR-seq could be modified by incorporating alternative strategies

445 such as poly(A) tailing [40, 41].

446 # Materials and Methods

447 **Cohort design**

448 Seventy-five participants, including patients with colorectal cancer (n=24), lung cancer
449 (n=20), and healthy controls (n=31), were enrolled in this study. Samples were obtained
450 between November 2018 to January 2022. Individuals with colorectal, lung cancer and
451 healthy controls were recruited from Peking University First Hospital. The characteristics
452 of participants in this study were summarized **(Supplementary Table 9)**.

453

454 **Sample collection**

455 Peripheral whole blood samples were collected in EDTA-coated vacutainer tubes for each
456 participant. Blood samples of patients with cancer were collected before any treatment of
457 surgery, chemotherapy, or neoadjuvant chemotherapy. Within 2 hours after blood
458 collection, blood samples were centrifuged at 1,900g for 30 min at room temperature.
459 Plasma was separated and then centrifuged at 16,000g for another 10 min at 4 °C to
460 remove cellular debris. All plasma samples were aliquoted and stored at -80 °C until
461 analysis.

462

463 **Plasma and extracellular vesicles RNA extraction**

464 Cell-free RNAs (cfRNAs) were extracted from 1 ml of plasma using QIAzol Lysis Reagent
465 (Qiagen, 79306) according to the manufacturer's instructions. The upper, aqueous phase
466 containing cfRNAs was mixed with 1 volume of ethanol (95-100%) and then added to the

467 Zymo-Spin column (Zymo, R1016) for RNA binding. Samples were subsequently washed,
468 eluted, and treated with DNase I (TaKaRa, 2270A) for 20 min at 37 °C. Following residual
469 DNA digestion, cfRNAs were then purified and concentrated into 6 μl using an RNA Clean
470 and Concentrator-5 kit (Zymo, R1016). Plasma extracellular vesicles (EVs) were purified
471 by a membrane-affinity approach using an exoRNeasy Midi Kit (Qiagen, 77144) following
472 the manufacturer's instructions [42]. EVs were eluted with 400 μl of elution buffer and
473 characterized by transmission electron microscopy (TEM) and nanoparticle tracking
474 analysis (NTA). For RNA isolation, EVs were lysed on the exoRNeasy column using
475 QIAzol Lysis Reagent, and EV RNAs were extracted and purified using Zymo-Spin column
476 as mentioned above.
477
478 **Optimization of cell-free RNA extraction and residual DNA digestion.**
479 RNA extraction is one of the most critical steps for low-input RNA-seq. To this end, we
480 compared three cfRNA extraction approaches, including QPCB (QIAzol lysis, phenol-
481 chloroform extraction, and column binding), Norgen (Plasma/Serum Circulating and
482 Exosomal RNA Purification Kit), and QPIP (QIAzol lysis, phenol-chloroform extraction, and
483 isopropanol precipitation). QPCB was considered the best approach for cfRNA extraction.
484 **(Supplementary Figure 3)**.
485 In previous reports, DNA contamination has been emphasized as a hinder to the
486 cfRNA study [15]. Therefore, we examined two major residual DNA digestion approaches:
487 On-column vs. In-buffer (On-column: residual cell-free DNA was digested on the spin-

488 column during RNA extraction; In-buffer: DNA was digested in the aqueous buffer after
489 RNA extraction). We observed a significantly higher human genome mapping ratio and
490 exonic read ratio with In-buffer DNA digestion than On-column approach (P-value < 0.0001,
491 Wilcoxon rank sum test; **Supplementary Figure 4**), suggesting In-buffer DNA digestion
492 was more effective to a certain extent. DETECTOR-seq was carried out in the following
493 assays with RNAs extracted using QPCB and residual DNA digested with an In-buffer
494 approach unless specified.
495
496 **Reverse transcription**
497 Cell-free RNAs were captured using random primers with a unique sample barcode and
498 then reverse transcribed with SMARTScribe reverse transcriptase (Clontech, 639538) and
499 template-switching oligos tagging 8-nt UMI sequences. Sample barcodes were designed
500 in R using the DNABarcodes package [43]. We generated barcodes with a length of 4
501 nucleotides and a minimum Hamming distance of 3 and filtered self-complementary
502 sequences, triplets, and sequences that have an unbalanced ratio of bases G or C versus
503 A or T. PEG 8000 (Beyotime, R0056-2ml) was used as molecular crowding reagent to
504 further improve the efficiency of reverse transcription reaction [44]. The 20-μl reaction
505 mixture was incubated at 42 °C for 90 min with a heat inactivation step at 70 °C for 10 min.
506 Primers for the reverse transcription of DETECTOR-seq were shown in **Supplementary**
507 **Table 3**.
508

509 **Quantitative PCR analysis**

510 The total abundance level of *Plasma* or *EV* cfRNAs was assessed by amplifying a
511 fragment from the human gene of *ACTB* spanning the exon-exon junction (ACTB-ee). The
512 level of residual DNA contamination was measured by amplifying a short fragment from
513 *ACTB* within intron regions (ACTB-i). We measured the microbiome contamination by the
514 threshold cycle (Ct) value difference of the ACTB-ee and the bacterial 16S ribosomal RNA
515 V4 fragment (16S-V4). The 2.5 μl of cDNA template was amplified in a final volume of 20
516 μl using the FastFire qPCR PreMix (SYBR Green) (TIANGEN, FP207). Samples with low
517 RNA content (Ct of ACTB-ee > 32) or high DNA contamination (Ct of ACTB-i < 35), or high
518 bacterial contamination (ΔCt (ACTB-ee – 16S-V4) > 5) were excluded for further analysis.
519 We summarized the quality control primers for *Plasma* and *EV* cfRNA samples
520 **(Supplementary Table 1).**

521

522 **Design of guide RNAs (gRNAs)**

523 To remove highly abundant sequences (rRNAs and mtRNAs) in the cfRNA library of
524 human plasma, we designed 302 and 315 high-quality single guide RNAs (sgRNAs)
525 specifically targeting the ribosomal and mitochondrial RNA sequences **(Supplementary**
526 **Figure 1)**. The sgRNAs were selected and filtered by DASHit [45] and Benchling
527 (https://www.benchling.com/crispr) based on the following criteria: 1) off-target score
528 (specificity score) and on-target score (sensitivity score); 2) poorly structured sgRNAs
529 were excluded; 3) cover approximately every 50 bp over the target sequences. First,

530 candidate guides in rRNA (4324) and mtRNA (1907) were outputted by scanning all the
531 protospacer adjacent motif (PAM) sequences. Meanwhile, the transcriptome of hg38
532 (excluded rRNA and mitochondrial RNA) was also scanned for all CRISPR sites as an off-
533 target list. Second, we excluded off-target guides and filtered poorly structured guides,
534 including a). G/C frequency too high (> 15/20) or too low (< 5/20); b). Homopolymer: more
535 than 5 consecutive repeated nucleotides; c). Dinucleotide repeats: the same two
536 nucleotides alternate for > 3 repeats; d). Hairpin: complementary subsequences near the
537 start and end of a binding site, causing a hairpin. And we got 1191 rRNA guides and 1425
538 mtRNA guides as qualified guides. Next, we downloaded guides designed by Benchling
539 to score our guides and remove redundant guides (overlapped with each other), and kept
540 guides with higher on-target and off-target scores. Thus, we got a pool of filtered guides.
541 Finally, we manually added some guides to cover the non-guide regions and SNP sites
542 and obtained a final sgRNA pool containing 302 guides targeting rRNA sequences and
543 315 guides targeting mtRNA sequences. The DNA templates of final sgRNAs were
544 synthesized through a one-step PCR using two paired primers to achieve the addition of
545 T7 promoter and guide RNA scaffold sequences (Primers for preparation of sgRNA DNA
546 templates were shown in **Supplementary Table 2**). The final sgRNAs were in vitro
547 transcribed using T7 RNA polymerase (NEB, E2050) and stored at -80 °C.
548

549 **DETECTOR-seq library preparation**

550 The 17.5 μl of remaining samples with similar cDNA content (ΔCt of ACTB-ee < 1) were

551 pooled for 3-, 4- or 5-plex library preparation. The pooled cDNAs were pre-amplified using
552 SeqAmp DNA Polymerase (Clontech, 638509) with the following PCR setup: initial
553 denaturation at 94 °C for 1 min, denaturation at 98 °C for 15 s, annealing at 55 °C for 30
554 s, elongation at 68 °C for 30 s, and final elongation at 68 °C for 10 min. Pre-amplification
555 was repeated for 6 cycles, and the DNA was cleaned and size-selected using Hieff NGS
556 DNA selection Beads (Yeasen, 12601ES56) at a ratio of 1:0.8 of DNA to beads twice. The
557 DNA was eluted with the 20 μl CRISPR-Cas9 reaction mix consisting of 300 ng sgRNAs
558 for rRNA sequences, 40 ng sgRNAs for mtRNA sequences, 1× NEBuffer3.1 and 1 μM
559 Cas9 Nuclease, *S. pyogenes* (NEB, M0386). The 20-μl reaction mixture was incubated at
560 37 °C for 60 min and heat-inactivated at 65°C for 5 min. Following the depletion of DNA
561 fragments derived from rRNAs and mtRNAs, the remaining DNA samples with complete
562 library structure were amplified for 16-18 cycles depending on the initial input. The final
563 clean-up was conducted at a ratio of 1:1 of DNA to beads. The library concentration was
564 measured using a Qubit dsDNA HS Assay Kit (Thermo Fisher, Q32854), and the size
565 distribution of the library was assessed using an Agilent 2100 Bioanalyzer with a High-
566 Sensitivity DNA analysis kit (Agilent, 5067-4626). DETECTOR-seq libraries were
567 sequenced on an Illumina HiSeqX platform to a depth of 10 million 150 bp paired-end
568 reads per sample. Primers for PCR amplification of DETECTOR-seq were shown in
569 **Supplementary Table 4**.
570
571 **Sequencing data processing**

572 Raw sequencing data were demultiplexed using sabre (https://github.com/najoshi/sabre)
573 according to sample barcodes. UMI sequences were extracted using UMI-Tools [46].
574 Adapters were removed by cutadapt, and read pairs with an average quality score below
575 30 in either read were removed. The remaining read pairs were then mapped to ERCC's
576 spike-in sequences, NCBI's UniVec sequences, human rRNA sequences, human mtRNA
577 sequences, human genome (hg38), and circular RNA sequentially using STAR (version
578 2.5.3a_modified) [47]. The UMI-Tools package was used to remove duplicated reads
579 caused by PCR amplification. Finally, a gene count matrix was generated using
580 featureCounts v1.6.2 [48] with the GENCODE v38 annotation. Unmapped reads were
581 classified using kraken2 [49] to obtain microbe (including bacterial, archaeal, and viral)
582 genus abundance. Potential contaminations in genera were filtered before downstream
583 analysis as in previously published work [19]. We summarized all the datasets for the
584 development, validation, and application of DETECTOR-seq in **Supplementary Table 8**.
585

586 **Quality control (Sample filtering and Gene filtering)**

587 To filter low-quality DETECTOR-seq datasets, the following quality control criteria were
588 used **(Supplementary Table 7)**: (1) raw reads > 4 M; (2) clean reads (reads remained
589 after trimming low-quality and adapter sequences) > 3.8 M; (3) ribosomal RNA reads <
590 20%; (4) mitochondrial RNA reads < 20%; (5) genome-aligned reads > 2 M; (6) de-
591 duplicated RNA reads > 0.1 M. Next, we retained genes with TPM > 1 in at least 50% of
592 samples or filtered genes by filterbyExpr of edgeR package [50].

593

594 **Differential expression and functional enrichment analysis**

595 The count matrix of gene expression or microbe genus abundance was normalized using
596 the trimmed mean of M-values (TMM) method in the edgeR package [50]. Differential
597 expression analysis was conducted using a quasi-likelihood method with FDR<0.1 to
598 identify RNAs showing a selective distribution in paired *Plasma* and *EV* samples and to
599 identify differentially expressed genes (DEGs) ($|\log_2$fold-change$|>1$ and FDR<0.05)
600 between cancer patients and normal controls. GSEA analysis of GO and KEGG pathway
601 was carried out using clusterProfiler [51].

602

603 **Enrichment analysis of RBP binding motifs/sites**

604 After identifying selective RNAs showing significantly differential abundance between
605 *Plasma* and *EV* transcriptomes, we conducted an enrichment analysis of RBP (RNA
606 binding protein) binding motifs/sites using MEME SEA [52]. We first created a gene-wise
607 "RBP binding hotspot" sequence set by expanding annotated exon junction
608 sites upstream and downstream by 20 nt and combined with 5' UTR and 3' UTR regions
609 (GENCODE v38), as these regions were reported to be frequently bound by RBPs [53].
610 Background sequences were extracted from 500 random subsets of cfRNAs whose
611 abundance showed no significant difference between *Plasma* and *EV* (FDR>0.1).
612 Database files of RBP binding motifs/sites for enrichment analysis were annotated from
613 our previous research [53]. Finally, top enriched RBPs (ranked by *E*-value) were annotated

614 and summarized, and sequence logo images were created from POSTAR3 database [53]
615 using WebLogo [54].
616
617 **Deconvolution of cell types of origin**
618 We applied Nu-SVR to deconvolve the fractions of cell-type-specific RNAs based on
619 Tabula Sapiens version 1.0 (TSP), a multiple-donor whole-body cell atlas spanning 24
620 tissues and organs as previously reported [23].
621
622 **Cancer classification**
623 We normalized and scaled gene expression and genus abundance for evaluating the
624 cancer-differentiating capacity of human and microbial features in *Plasma* and *EV* cfRNAs.
625 All of the 61 *Plasma* and 52 *EV* DETECTOR-seq datasets passed QC were used, thus
626 including as many cases as possible in the training and test sets. Most of the cancer
627 samples were paired between *Plasma* and *EV* (CRC samples: *Plasma* 23, *EV* 19, 19 of
628 them were paired; LC samples: *Plasma* 19, *EV* 19, 18 of them were paired; NC samples:
629 *Plasma* 19, *EV* 14, 7 of them were paired). The data were trained and tested with
630 bootstrapping sampling, which was randomly repeated 20 times. For human RNAs, a
631 quasi-likelihood method was used for the differential expression analysis in each
632 bootstrapping procedure. In Figure 7C, differentially expressed features with $|\log_2$fold-
633 change$|>1$ and FDR<0.05 overlapped with RNAs that were enriched in *Plasma* or *EV*
634 (defined in Figure 6A) were further used to fit a random forest classifier. In Figure 7D, we

635 selected the top 200 features ranked by FDR in each bootstrapping procedure. For
636 microbial RNAs, we selected all of the microbe genera with $|\log_2 \text{fold-change}|>1$ and
637 FDR<0.1 in each bootstrapping procedure. For the combination of human RNAs and
638 microbial RNAs, we combined human gene expression and genus abundance and
639 selected the top 200 features ranked by FDR. The area under the receiver operating
640 characteristic curves (AUROC) was calculated from the final probability using the pROC
641 [55] package in R.
642

643 **Cost estimation**

644 The cost for cell-free RNA library preparation of DETECTOR-seq was determined using
645 the sum of the price for each component used in our protocol. The price of SMARTer
646 Stranded Total RNA-Seq Kit v2-Pico Input Mammalian (TaKaRa, 634413) was searched
647 on the official website of TaKaRa for the estimation of SMARTer-seq. The cost of Phospho-
648 RNA-seq was estimated using T4 polynucleotide kinase (NEB, M0201S) and TruSeq small
649 RNA kit (Illumina, RS-200-0012). In the case of SILVER-seq, there was no publicly
650 available step-by-step protocol, thus the cost of SILVER-seq was estimated by the Ovation
651 SoLo RNA-Seq Kit (NuGEN, 0500-96). In all cases, the prices listed in **Supplementary**
652 **Tables 6 and 7** included sales tax. Because the costs of SMARTer-seq, Phospho-RNA-
653 seq, and SILVER-seq were estimated using commercial kits (including additional selling
654 costs and profits), for a fair comparison, we determined the cost of DETECTOR-seq as
655 twice the calculated price.

Declarations

Ethics approval and consent to participate

This study was approved by the institutional review board of Peking University First Hospital (2018-15). Informed consent was obtained from all patients.

Consent for publication

Not applicable.

Availability of data and materials

Data generated with DETECTOR-seq are available at the Gene Expression Omnibus under accession number GSE216561. For benchmarking, we used the following datasets: GSE126049 (Phospho-RNA-seq), GSE131512 (SILVER-seq), and GSE174302 (SMARTer-seq). *For editors and reviewers: the data can be downloaded from the GEO with a secure token: cnwpoiwufhunbqd.*

Declaration of interests

A patent application on the described technology has been filed by HKW and ZJL. Other authors declare no conflict of interest.

Funding and Acknowledgments

This work is supported by National Natural Science Foundation of China (32170671, 8237061277 and 81902384), the Capital's Funds for Health Improvement and Research (CFH 2022-2-4075), the National Key Research and Development Plan of China (2022ZD0117700), Tsinghua University Guoqiang Institute Grant (2021GQG1020), Tsinghua University Initiative Scientific Research Program of Precision Medicine (2022ZLA003), Tsinghua University Spring Breeze Fund (2021Z99CFY022). This study was also supported by Bayer Micro-funding, Bio-Computing Platform of Tsinghua University Branch of China National Center for Protein Sciences.

Authors' contributions

HKW, QZ, and ZJL conceived and designed the project; HKW developed DETECTOR-seq and generated the datasets; SZ and PYW collected the clinical samples; QZ and HKW conducted the analyses; all authors wrote and approved the manuscript.

690 # References

691

692 1. Heitzer E, Haque IS, Roberts CES, Speicher MR: **Current and future perspectives of liquid**
693 **biopsies in genomics-driven oncology.** *Nature Reviews Genetics* 2019, **20:**71-88.
694 2. Wei ZY, Batagov AO, Schinelli S, Wang JT, Wang Y, El Fatimy R, Rabinovsky R, Balaj L, Chen
695 CC, Hochberg F, et al: **Coding and noncoding landscape of extracellular RNA released by**
696 **human glioma stem cells.** *Nature Communications* 2017, **8:**1-15.
697 3. Gruner HN, McManus MT: **Examining the evidence for extracellular RNA function in**
698 **mammals.** *Nature Reviews Genetics* 2021, **22:**448-458.
699 4. Moufarrej MN, Vorperian SK, Wong RJ, Campos AA, Quaintance CC, Sit RV, Tan M, Detweiler
700 AM, Mekonen H, Neff NF, et al: **Early prediction of preeclampsia in pregnancy with cell-**
701 **free RNA.** *Nature* 2022, **602:**689-694.
702 5. Rasmussen M, Reddy M, Nolan R, Camunas-Soler J, Khodursky A, Scheller NM, Cantonwine
703 DE, Engelbrechtsen L, Mi JD, Dutta A, et al: **RNA profiles reveal signatures of future health**
704 **and disease in pregnancy.** *Nature* 2022, **601:**422-427.
705 6. Ngo TTM, Moufarrej MN, Rasmussen MLH, Camunas-Soler J, Pan WY, Okamoto J, Neff NF,
706 Liu KL, Wong RJ, Downes K, et al: **Noninvasive blood tests for fetal development predict**
707 **gestational age and preterm delivery.** *Science* 2018, **360:**1133-1136.
708 7. Yu S, Li Y, Liao Z, Wang Z, Wang Z, Li Y, Qian L, Zhao J, Zong H, Kang B, et al: **Plasma**
709 **extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection**
710 **of pancreatic ductal adenocarcinoma.** *Gut* 2020, **69:**540-550.
711 8. Li YC, Zhao JL, Yu SL, Wang Z, He XG, Su YH, Guo TAN, Sheng HY, Chen J, Zheng QP, et al:
712 **Extracellular Vesicles Long RNA Sequencing Reveals Abundant mRNA, circRNA, and**
713 **lncRNA in Human Blood as Potential Biomarkers for Cancer Diagnosis.** *Clinical Chemistry*
714 2019, **65:**798-808.
715 9. Ji J, Chen R, Zhao L, Xu YL, Cao Z, Xu H, Chen X, Shi XL, Zhu YS, Lyu J, et al: **Circulating**
716 **exosomal mRNA profiling identifies novel signatures for the detection of prostate cancer.**
717 *Molecular Cancer* 2021, **20:**58.
718 10. Larson MH, Pan WY, Kim HJ, Mauntz RE, Stuart SM, Pimentel M, Zhou YQ, Knudsgaard P,
719 Demas V, Aravanis AM, Jamshidi A: **A comprehensive characterization of the cell-free**
720 **transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection.**
721 *Nature Communications* 2021, **12:**1-11.
722 11. Cabus L, Lagarde J, Curado J, Lizano E, Perez-Boza J: **Current challenges and best practices**
723 **for cell-free long RNA biomarker discovery.** *Biomarker Research* 2022, **10:**62.
724 12. Giraldez MD, Spengler RM, Etheridge A, Goicochea AJ, Tuck M, Choi SW, Galas DJ, Tewari M:
725 **Phospho-RNA-seq: a modified small RNA-seq method that reveals circulating mRNA**
726 **and lncRNA fragments as potential biomarkers in human plasma.** *EMBO J* 2019,
727 **38:**e101695.
728 13. Akat KM, Lee YA, Hurley A, Morozov P, Max KE, Brown M, Bogardus K, Sopeyin A, Hildner K,
729 Diacovo TG, et al: **Detection of circulating extracellular mRNAs by modified small-RNA-**

730 **sequencing analysis.** *JCI Insight* 2019, **5:**e127317.
731 14. Zhou Z, Wu Q, Yan Z, Zheng H, Chen CJ, Liu Y, Qi Z, Calandrelli R, Chen Z, Chien S, et al:
732 **Extracellular RNA in a single droplet of human serum reflects physiologic and disease**
733 **states.** *Proc Natl Acad Sci U S A* 2019, **116:**19200-19208.
734 15. Verwilt J, Trypsteen W, Van Paemel R, De Preter K, Giraldez MD, Mestdagh P, Vandesompele
735 J: **When DNA gets in the way: A cautionary note for DNA contamination in extracellular**
736 **RNA-seq studies.** *Proc Natl Acad Sci U S A* 2020, **117:**18934-18936.
737 16. Stark R, Grzelak M, Hadfield J: **RNA sequencing: the teenage years.** *Nature Reviews Genetics*
738 2019, **20:**631-656.
739 17. Farmer AA, Betts C, Bolduc N: **Methods of depleting a target molecule from an initial**
740 **collection of nucleic acids, and compositions and kits for practicing the same.** 2018.
741 18. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL: **Depletion**
742 **of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-**
743 **abundance species in sequencing libraries and molecular counting applications.** *Genome*
744 *Biol* 2016, **17:**1-13.
745 19. Chen S, Jin Y, Wang S, Xing S, Wu Y, Tao Y, Ma Y, Zuo S, Liu X, Hu Y, et al: **Cancer type**
746 **classification using plasma cell-free RNAs derived from human and microbes.** *Elife* 2022,
747 **11:**e75181.
748 20. Moller A, Lobb RJ: **The evolving translational potential of small extracellular vesicles in**
749 **cancer.** *Nat Rev Cancer* 2020, **20:**697-709.
750 21. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, Mitchell PS, Bennett CF,
751 Pogosova-Agadjanyan EL, Stirewalt DL, et al: **Argonaute2 complexes carry a population of**
752 **circulating microRNAs independent of vesicles in human plasma.** *Proc Natl Acad Sci U S*
753 *A* 2011, **108:**5003-5008.
754 22. Pan WY, Ngo TTM, Camunas-Soler J, Song CX, Kowarsky M, Blumenfeld YJ, Wong RJ, Shaw
755 GM, Stevenson DK, Quake SR: **Simultaneously Monitoring Immune Response and**
756 **Microbial Infections during Pregnancy through Plasma cfRNA Sequencing.** *Clinical*
757 *Chemistry* 2017, **63:**1695-1704.
758 23. Vorperian SK, Moufarrej MN, Tabula Sapiens C, Quake SR: **Cell types of origin of the cell-**
759 **free transcriptome.** *Nat Biotechnol* 2022, **40:**855-861.
760 24. Reggiardo RE, Maroli SV, Peddu V, Davidson AE, Hill A, LaMontagne E, Aaraj YA, Jain M, Chan
761 SY, Kim DH: **Profiling of repetitive RNA sequences in the blood plasma of patients with**
762 **cancer.** *Nat Biomed Eng* 2023.
763 25. Lee MN, Roy M, Ong SE, Mertins P, Villani AC, Li WB, Dotiwala F, Sen J, Doench JG, Orzalli MH,
764 et al: **Identification of regulators of the innate immune response to cytosolic DNA and**
765 **retroviral infection by an integrative approach.** *Nature Immunology* 2013, **14:**179-185.
766 26. Liu YH, Qu LL, Liu YY, Roizman B, Zhou GG: **PUM1 is a biphasic negative regulator of innate**
767 **immunity genes by suppressing LGP2.** *Proceedings of the National Academy of Sciences of*
768 *the United States of America* 2017, **114:**6902-6911.
769 27. Ansa-Addo EA, Huang HC, Riesenberg B, Iamaswat S, Borucki D, Nelson MH, Nam JH, Chung
770 D, Paulos CM, Liu B, et al: **RNA binding protein PCBP1 is an intracellular immune**

checkpoint for shaping T cell responses in cancer immunity. *Science Advances* 2020, **6:**3865.

28. Coker OO, Wu WKK, Wong SH, Sung JJY, Yu J: **Altered Gut Archaea Composition and Interaction With Bacteria Are Associated With Colorectal Cancer.** *Gastroenterology* 2020, **159:**1459-1470.

29. Moens U, Calvignac-Spencer S, Lauber C, Ramgvist T, Feltkamp MCW, Daugherty MD, Verschoor EJ, Ehlers B, Consortium IR: **ICTV Virus Taxonomy Profile: Polyomaviridae.** *Journal of General Virology* 2017, **98:**1159-1160.

30. Verboom K, Everaert C, Bolduc N, Livak KJ, Yigit N, Rombaut D, Anckaert J, Lee S, Veno MT, Kjems J, et al: **SMARTer single cell total RNA sequencing.** *Nucleic Acids Research* 2019, **47:**e93.

31. Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W, et al: **The landscape of accessible chromatin in mammalian preimplantation embryos.** *Nature* 2016, **534:**652-657.

32. Hardigan AA, Roberts BS, Moore DE, Ramaker RC, Jones AL, Myers RM: **CRISPR/Cas9-targeted removal of unwanted sequences from small-RNA sequencing libraries.** *Nucleic Acids Res* 2019, **47:**e84.

33. Prezza G, Heckel T, Dietrich S, Homberger C, Westermann AJ, Vogel J: **Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads.** *RNA* 2020, **26:**1069-1078.

34. Wilkins OG, Ule J: **Ribocutter: Cas9-mediated rRNA depletion from multiplexed Ribo-seq libraries.** *bioRxiv* 2021**:**2021.2007.2014.451473.

35. Loi DSC, Yu L, Wu AR: **Effective ribosomal RNA depletion for single-cell total RNA-seq by scDASH.** *PeerJ* 2021, **9:**e10717.

36. Su YH, Li YC, Guo R, Zhao JJ, Chi WR, Lai HY, Wang J, Wang Z, Li L, Sang YT, et al: **Plasma extracellular vesicle long RNA profiles in the diagnosis and prediction of treatment response for breast cancer.** *Npj Breast Cancer* 2021, **7:**1-10.

37. Toden S, Zhuang JL, Acosta AD, Karns AP, Salathia NS, Brewer JB, Wilcock DM, Aballi J, Nerenberg M, Quake SR, Ibarra A: **Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing.** *Science Advances* 2020, **6:**1654.

38. He YD, Tao W, He T, Wang BY, Tang XM, Zhang LM, Wu ZQ, Deng WM, Zhang LX, Shao CK, et al: **A urine extracellular vesicle circRNA classifier for detection of high-grade prostate cancer in patients with prostate-specific antigen 2-10 ng/mL at initial biopsy.** *Molecular Cancer* 2021, **20:**1-6.

39. Nabet BY, Qiu Y, Shabason JE, Wu TJ, Yoon T, Kim BC, Benci JL, DeMichele AM, Tchou J, Marcotrigiano J, Minn AJ: **Exosome RNA Unshielding Couples Stromal Activation to Pattern Recognition Receptor Signaling in Cancer.** *Cell* 2017, **170:**352-366.

40. Salmen F, De Jonghe J, Kaminski TS, Alemany A, Parada GE, Verity-Legg J, Yanagida A, Kohler TN, Battich N, van den Brekel F, et al: **High-throughput total RNA sequencing in single cells using VASA-seq.** *Nature Biotechnology* 2022.

41. Isakova A, Neff N, Quake SR: **Single-cell quantification of a broad RNA spectrum reveals**

812 **unique noncoding patterns associated with cell types and states.** *Proceedings of the*
813 *National Academy of Sciences of the United States of America* 2021, **118**.
814 42. Enderle D, Spiel A, Coticchia CM, Berghoff E, Mueller R, Schlumpberger M, Sprenger-Haussels
815 M, Shaffer JM, Lader E, Skog J, Noerholm M: **Characterization of RNA from Exosomes and**
816 **Other Extracellular Vesicles Isolated by a Novel Spin Column-Based Method.** *Plos One*
817 2015, **10:**e0136133.
818 43. Buschmann T: **DNABarcodes: an R package for the systematic construction of DNA**
819 **sample tags.** *Bioinformatics* 2017, **33:**920-922.
820 44. Bagnoli JW, Ziegenhain C, Janjic A, Wange LE, Vieth B, Parekh S, Geuder J, Hellmann I, Enard
821 W: **Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq.** *Nature*
822 *Communications* 2018, **9**.
823 45. Dynerman D, Lyden A, Quan J, Caldera S, McGeever A, Dimitrov B, King R, Cirolia G, Tan M,
824 Sit R, et al: **Designing and implementing programmable depletion in sequencing libraries**
825 **with DASHit.** *bioRxiv* 2020**:**2020.2001.2012.891176.
826 46. Smith T, Heger A, Sudbery I: **UMI-tools: modeling sequencing errors in Unique Molecular**
827 **Identifiers to improve quantification accuracy.** *Genome Research* 2017, **27:**491-499.
828 47. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras
829 TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.
830 48. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for**
831 **assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30:**923-930.
832 49. Wood DE, Lu J, Langmead B: **Improved metagenomic analysis with Kraken 2.** *Genome Biol*
833 2019, **20:**1-13.
834 50. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential**
835 **expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
836 51. Yu GC, Wang LG, Han YY, He QY: **clusterProfiler: an R Package for Comparing Biological**
837 **Themes Among Gene Clusters.** *Omics-a Journal of Integrative Biology* 2012, **16:**284-287.
838 52. Bailey TL, Grant CE: **SEA: Simple Enrichment Analysis of motifs.** *bioRxiv*
839 2021**:**2021.2008.2023.457422.
840 53. Zhao WH, Zhang S, Zhu YM, Xi XC, Bao PF, Ma ZY, Kapral TH, Chen SY, Zagrovic B, Yang YCT,
841 Lu ZJ: **POSTAR3: an updated platform for exploring post-transcriptional regulation**
842 **coordinated by RNA-binding proteins.** *Nucleic Acids Research* 2022, **50:**D287-D294.
843 54. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: A sequence logo generator.**
844 *Genome Research* 2004, **14:**1188-1190.
845 55. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M: **pROC: an open-**
846 **source package for R and S plus to analyze and compare ROC curves.** *Bmc Bioinformatics*
847 2011, **12:**1-8.
848

849

850 # Figure Legends

851

852 **Figure 1 | Depletion-assisted multiplexed cell-free total RNA sequencing.**

853 **(A)** Bioanalyzer trace of cfRNA fragment lengths in a human plasma sample. (**B)** The relative
854 proportion of reads for various RNA biotypes detected by total RNA sequencing averaged by
855 three human plasma samples. **(C)** Distribution of reads' insert size for the fragmented rRNAs
856 and mtRNAs, derived from the above sequencing data. **(D)** Distribution of reads' coverage.
857 Blue bars on top represent sgRNA target sites. **(E)** The designed sgRNAs tiling the fragmented
858 rRNA and mtRNA sequences. **(F)** Schematic overview of DETECTOR-seq workflow. First,
859 cfRNAs are reverse transcribed with random primers and TSO. Sample barcodes and UMIs
860 are introduced during this step. Second, after calibrating input amounts, samples are pooled
861 and pre-amplified. Third, cDNAs of rRNAs and mtRNAs are depleted by CRISPR-Cas9.
862 Subsequently, DETECTOR-seq library is further amplified, then sequenced on an Illumina
863 platform. rRNA: ribosomal RNA; mtRNA: mitochondrial RNA; TSO: template switching oligo;
864 UMI: unique molecular identifier.

865

866 **Figure 2 | Efficient and specific depletion of rRNA and mtRNA sequences.**

867 **(A)** The read distributions and **(B)** coverages of untreated and rRNA/mtRNA-depleted
868 DETECTOR-seq libraries. Read coverage was normalized to total mapped reads. Pearson
869 correlation of cfRNA expression levels between **(C)** untreated and rRNA/mtRNA-depleted
870 DETECTOR-seq libraries, and **(D)** DETECTOR-seq versus SMARTer-seq. TPM: transcripts

871 per million mapped reads (rRNA/mtRNA reads were removed).

872

873 **Figure 3 | Comparing DETECTOR-seq with other cfRNA-seq methods.**

874 **(A)** Average percentages of genome-aligned reads mapping to exonic, intronic, and intergenic

875 regions for four different cfRNA-seq methods. **(B)** Average coverage across all mRNAs' 5' and

876 3' exon boundary sites flanking upstream and downstream by 50 bp. **(C)** Average percentages

877 of reads located in the sense and antisense strands of mRNAs' exons, introns, and promoters.

878 **(D)** Average percentages of clean reads (after trimming low-quality and adapter sequences)

879 assigned to different sources. **(E)** Numbers of raw sequencing reads and human genome-

880 aligned reads with a fixed budget of $300 for each method. **(F)** Summary of key techniques

881 used in the four cfRNA-seq approaches. Numbers of used samples: Phospho-seq: 15;

882 SILVER-seq: 128; SMARTer-seq: 373; DETECTOR-seq: 113.

883

884 **Figure 4 | Distinct human and microbial RNA signatures in *Plasma* versus *EV*.**

885 **(A)** Illustration of sequencing *Plasma* cfRNAs and *EV* cfRNAs in paired plasma samples. (B)

886 Plasma extracellular vesicles were characterized by nanoparticle tracking analysis and

887 transmission electron microscopy (scale bar represents 200 nm). **(C)** Distribution of reads

888 mapped to human genome and microbiome in *Plasma* and *EV* cfRNA datasets. Left: RNA

889 spectrum mapping to human genome; Right: relative abundance of reads aligned to different

890 phyla. **(D)** Differential human RNA species between *Plasma* and *EV* cfRNAs. **(E)** Pie charts

891 show the average fractional contributions of various cell types to the *Plasma* and *EV*

892 transcriptomes. Box plots show the diversity of cell type contributions to the *Plasma* and *EV*
893 transcriptomes measured by the ratio of non-blood cells and Simpson's index. **(F)** Boxplots
894 represent the enrichment of cfRNAs derived from transposable elements in *Plasma* cfRNAs
895 compared to *EV* cfRNAs. The different transposable element categories, including short
896 interspersed elements (SINEs), long interspersed elements (LINEs), long interspersed
897 elements with long terminal repeats (LTR), and DNA transposons, are represented. **(G)** The
898 fractions of reads aligned to microbe and virus. *Plasma*: 44 samples; *EV*: 44 samples (all
899 samples paired). ****: *P*-value < 0.0001, **: *P*-value < 0.01, *: *P*-value < 0.05, Wilcoxon rank
900 sum test, two-tailed.
901
902 **Figure 5 | Distinct functional pathways, motifs, and binding proteins of the selective**
903 ***Plasma* and *EV* cfRNAs.**
904 **(A)** Definition of the selective cfRNAs enriched in *Plasma* or *EV*. Cutoff: |Fold-change|>1 and
905 FDR<0.1. **(B)** Top enriched GO pathways of the selective cfRNAs. **(C)** Top enriched motifs
906 and their corresponding RNA binding proteins (RBPs) of the selective cfRNAs. *Plasma*: 44
907 samples; *EV*: 44 samples (all samples paired).
908
909 **Figure 6 | Cancer-relevant cfRNA signatures in *Plasma* and *EV*.**
910 **(A)** Cancer-relevant ones (differentially expressed between cancer patients and normal
911 controls, |$\log_2$fold-change|>1 and FDR<0.05) in the selective and non-selective human cfRNAs.
912 Cancer: colorectal cancer (CRC) and lung cancer (LC); NC: normal control. **(B)** Enriched GO

913 terms related to cancer-relevant human cfRNAs. Performances (average of 20 bootstrap
914 procedures) of cancer-relevant human cfRNAs distinguishing cancer patients from normal
915 controls when excluding **(C)** and including **(D)** non-selective cfRNAs. **(E)** AUROCs of cancer
916 type classification (CRC vs. LC) using human- or microbe-derived reads in *Plasma* and *EV*
917 cfRNAs. **(F)** Numbers of microbial features (genus) with significantly differential abundance
918 ($|\log_2$fold-change$|>1$ and FDR<0.1) between CRC and LC in 20 bootstrap procedures. **(G)**
919 Distinct cancer type-specific microbial features (genus) identified in *Plasma* and *EV* cfRNAs.
920 Heatmaps show z-scores of the abundance levels of these microbial RNA features; bar plots
921 illustrate their average $\log_2$FCs and FDRs between CRC and LC. FC: fold-change; FDR: false
922 discovery rate. ****: *P*-value < 0.0001, ***: *P*-value < 0.001, *: *P*-value < 0.05, Wilcoxon rank
923 sum test, two-tailed. CRC samples: *Plasma* (n=23), *EV* (n=19), 19 of them paired; LC samples:
924 *Plasma* (n=19), *EV* (n=19), 18 of them paired; NC samples: *Plasma* (n=19), *EV* (n=14), 7 of
925 them paired.

Figure 1

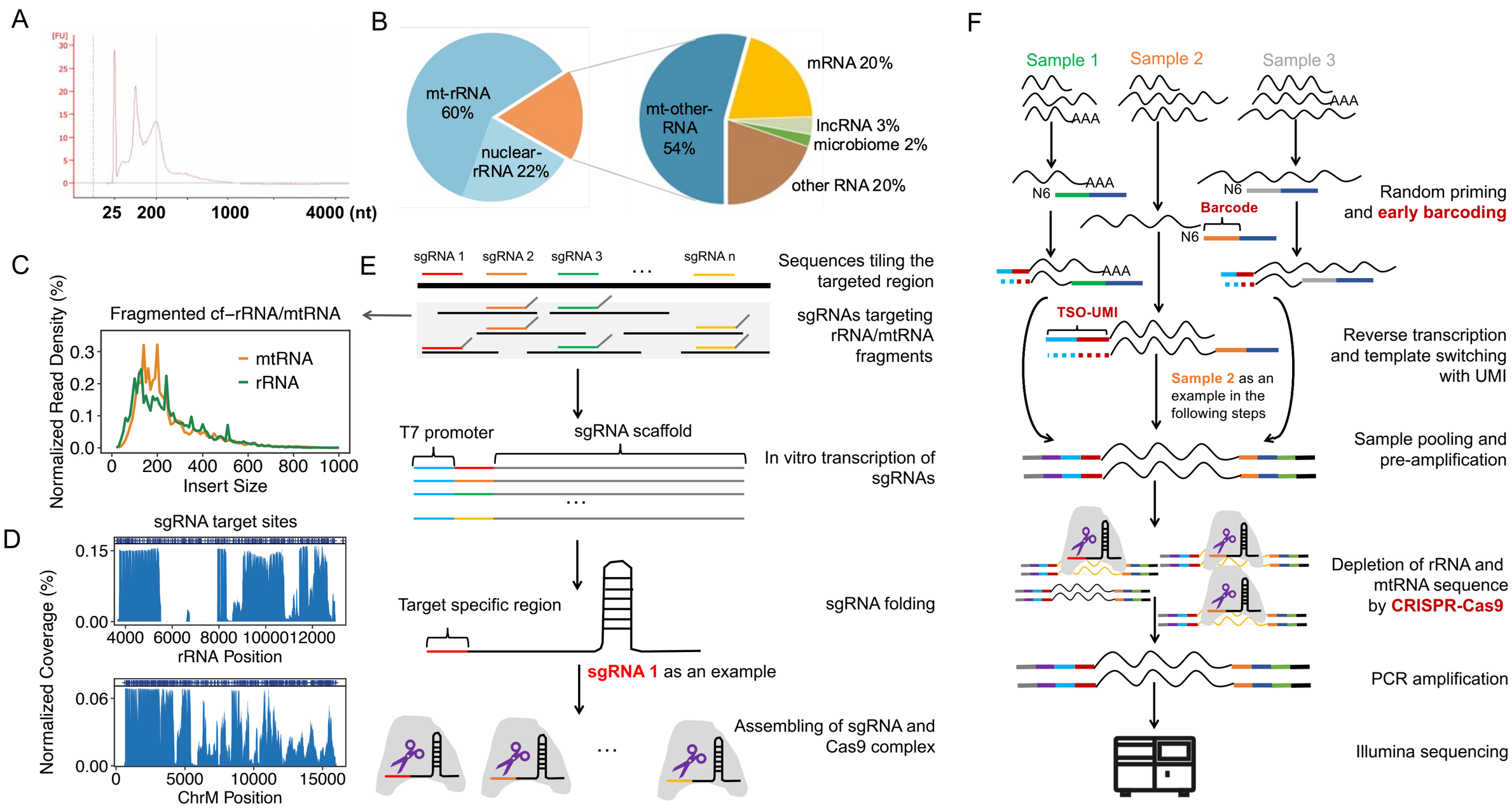

A
[FU]
25 200 1000 4000 (nt)
B
mt-rRNA 60%
nuclear-rRNA 22%
mt-other-RNA 54%
mRNA 20%
lncRNA 3%
microbiome 2%
other RNA 20%
C
Fragmented cf–rRNA/mtRNA
Normalized Read Density (%)
mtRNA
rRNA
Insert Size
D
sgRNA target sites
Normalized Coverage (%)
rRNA Position
ChrM Position
E
sgRNA 1
sgRNA 2
sgRNA 3
sgRNA n
Sequences tiling the targeted region
sgRNAs targeting rRNA/mtRNA fragments
T7 promoter
sgRNA scaffold
In vitro transcription of sgRNAs
Target specific region
sgRNA folding
sgRNA 1 as an example
Assembling of sgRNA and Cas9 complex
F
Sample 1
Sample 2
Sample 3
AAA
N6
Barcode
Random priming and early barcoding
TSO-UMI
Reverse transcription and template switching with UMI
Sample 2 as an example in the following steps
Sample pooling and pre-amplification
Depletion of rRNA and mtRNA sequence by CRISPR-Cas9
PCR amplification
Illumina sequencing

Figure 2

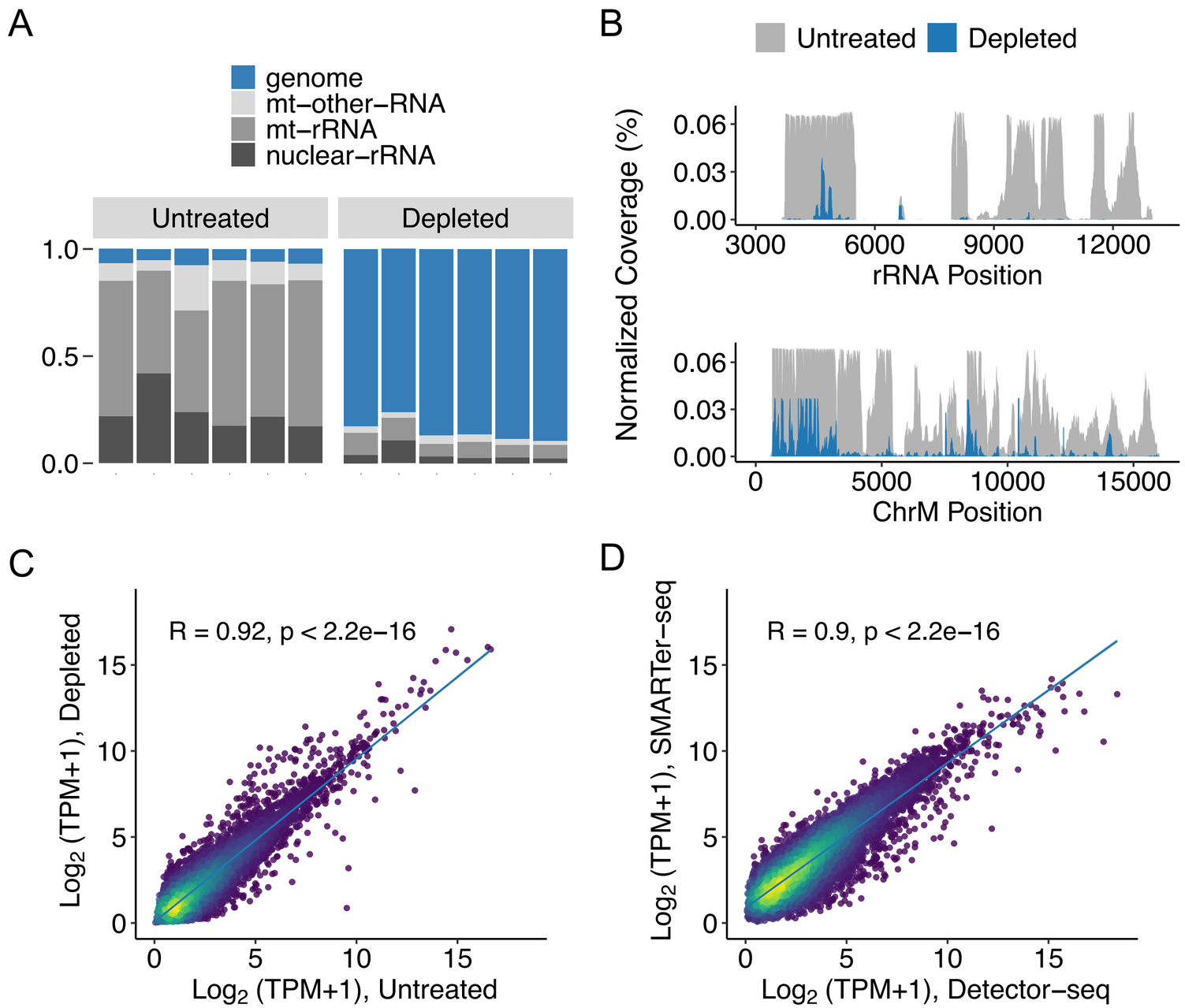

A
genome
mt−other−RNA
mt−rRNA
nuclear−rRNA
Untreated
Depleted
1.0
0.5
0.0
B
Untreated
Depleted
Normalized Coverage (%)
0.06
0.03
0.00
3000
6000
9000
12000
rRNA Position
0.06
0.03
0.00
0
5000
10000
15000
ChrM Position
C
R = 0.92, p < 2.2e−16
Log2 (TPM+1), Depleted
Log2 (TPM+1), Untreated
D
R = 0.9, p < 2.2e−16
Log2 (TPM+1), SMARTer−seq
Log2 (TPM+1), Detector−seq

Figure 3

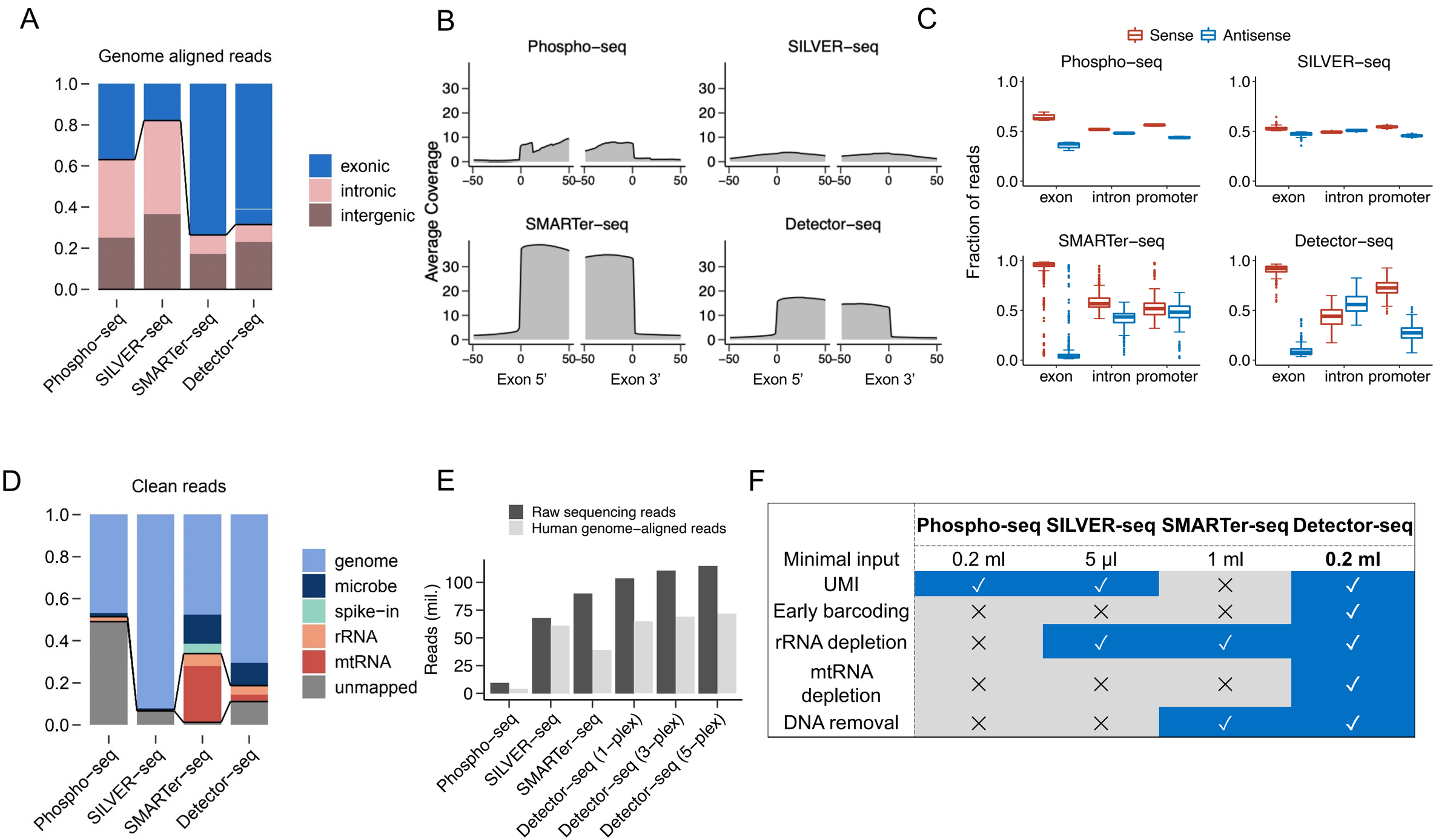

A
Genome aligned reads
1.0
0.8
0.6
0.4
0.2
0.0
exonic
intronic
intergenic
Phospho−seq
SILVER−seq
SMARTer−seq
Detector−seq
B
Phospho−seq
SILVER−seq
SMARTer−seq
Detector−seq
Average Coverage
Exon 5'
Exon 3'
C
Sense
Antisense
Fraction of reads
exon
intron
promoter
D
Clean reads
genome
microbe
spike−in
rRNA
mtRNA
unmapped
E
Raw sequencing reads
Human genome−aligned reads
Reads (mil.)
Detector−seq (1−plex)
Detector−seq (3−plex)
Detector−seq (5−plex)
F
Phospho-seq SILVER-seq SMARTer-seq Detector-seq
Minimal input
0.2 ml
5 µl
1 ml
0.2 ml
UMI
Early barcoding
rRNA depletion
mtRNA depletion
DNA removal

Figure 4

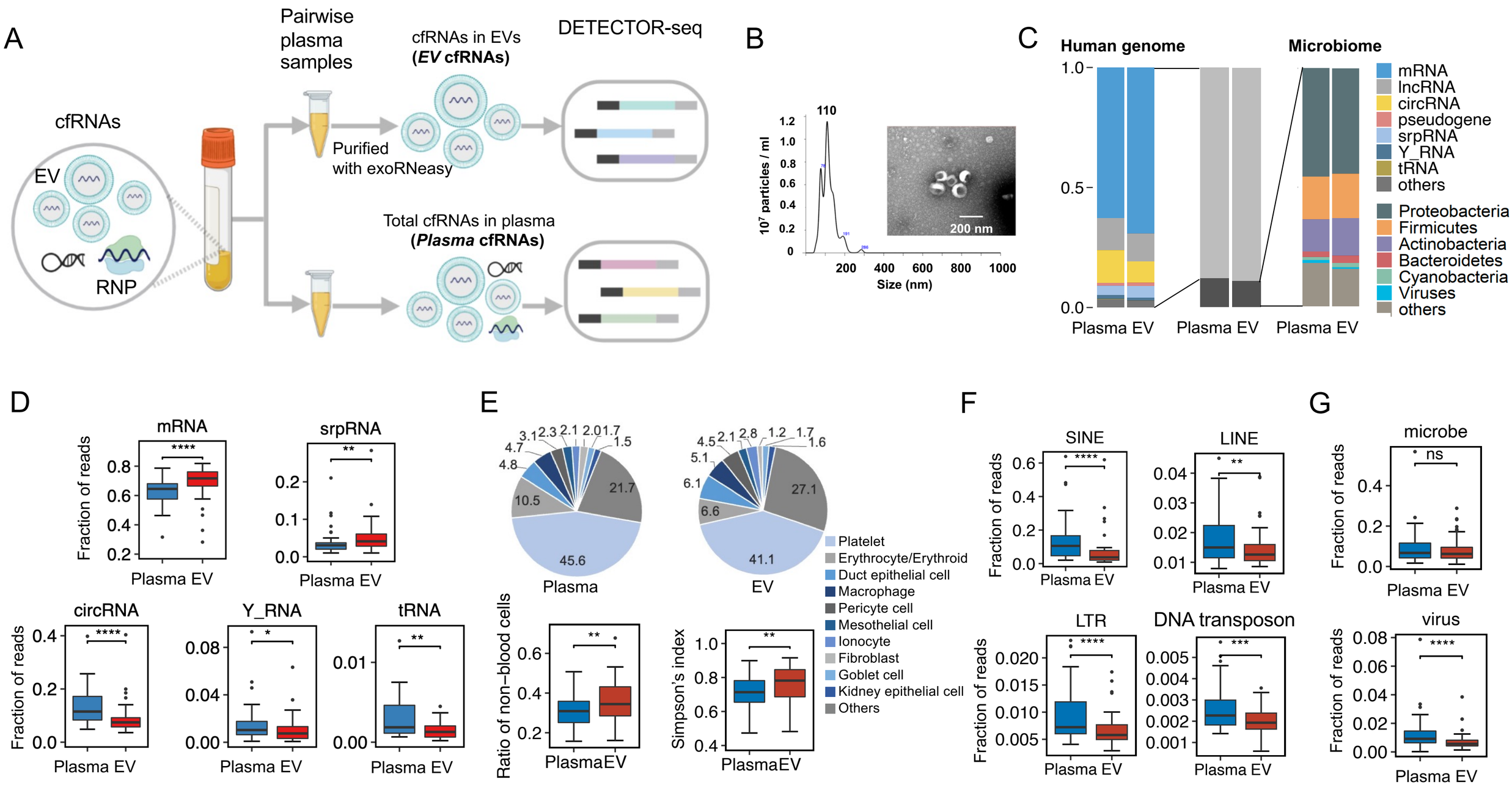
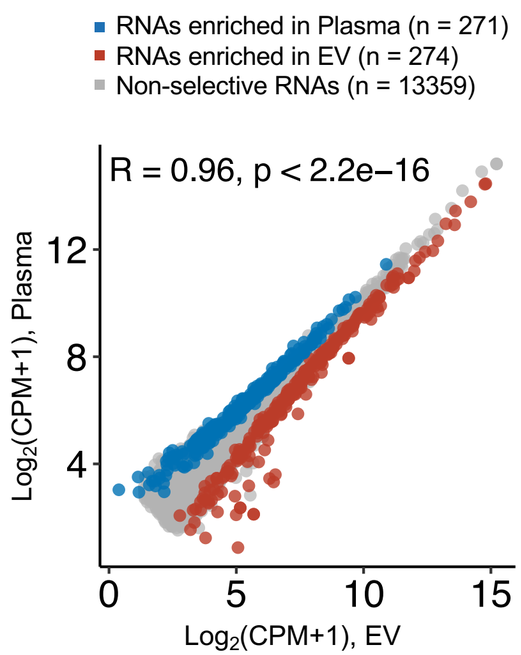
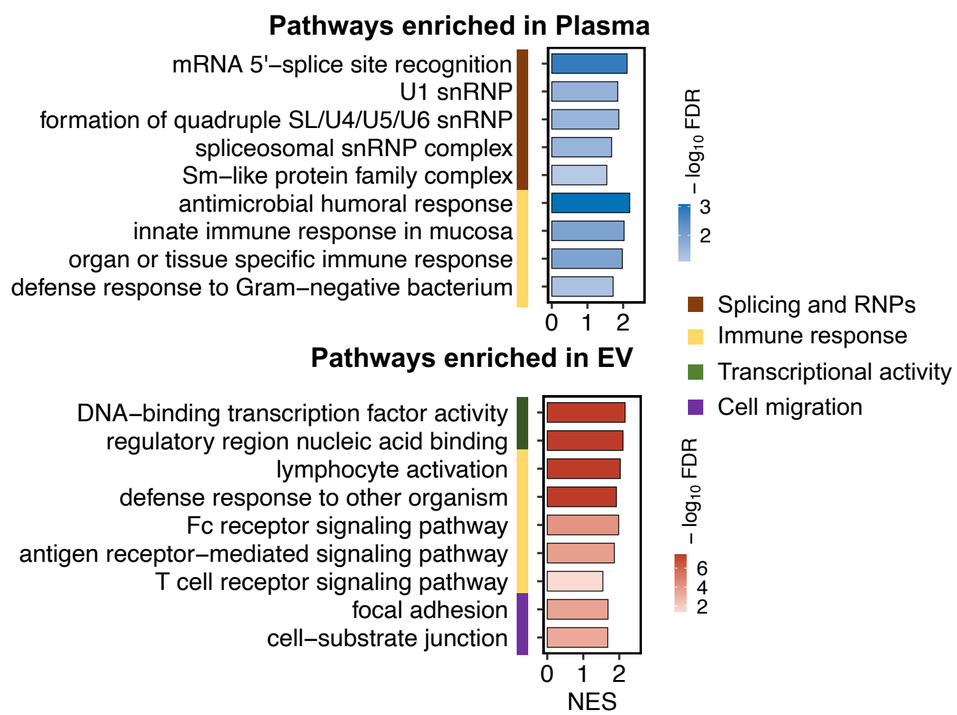

A
cfRNAs
EV
RNP
Pairwise plasma samples
cfRNAs in EVs (EV cfRNAs)
DETECTOR-seq
Purified with exoRNeasy
Total cfRNAs in plasma (Plasma cfRNAs)
B
110
10^7 particles / ml
Size (nm)
200 nm
C
Human genome
Microbiome
Plasma EV
mRNA
lncRNA
circRNA
pseudogene
srpRNA
Y_RNA
tRNA
others
Proteobacteria
Firmicutes
Actinobacteria
Bacteroidetes
Cyanobacteria
Viruses
others
D
mRNA
srpRNA
circRNA
Y_RNA
tRNA
Fraction of reads
Plasma EV
E
Plasma
EV
Platelet
Erythrocyte/Erythroid
Duct epithelial cell
Macrophage
Pericyte cell
Mesothelial cell
Ionocyte
Fibroblast
Goblet cell
Kidney epithelial cell
Others
Ratio of non-blood cells
Simpson's index
PlasmaEV
F
SINE
LINE
LTR
DNA transposon
G
microbe
virus

Figure 5

A



B



C

	Motif	E-value	RBP	Annotation
Enriched in Plasma		1.64E-35	ABCF1	innate immune
		1.46E-29	SFPQ	splicing factor
		2.31E-26	LARP4	La RNP
		4.51E-23	TROVE2	Y RNA binding protein
		1.77E-20	DKC1	snoRNP
Enriched in EV		2.57E-61	PUM1	innate immune
		4.7E-57	BCLAF1	transcription repressor
		2.4E-55	HNRNPU	transcription regulator
		9.47E-55	PCBP1	immune checkpoint
		1.12E-51	APOBEC3C	RNA editing

Figure 6

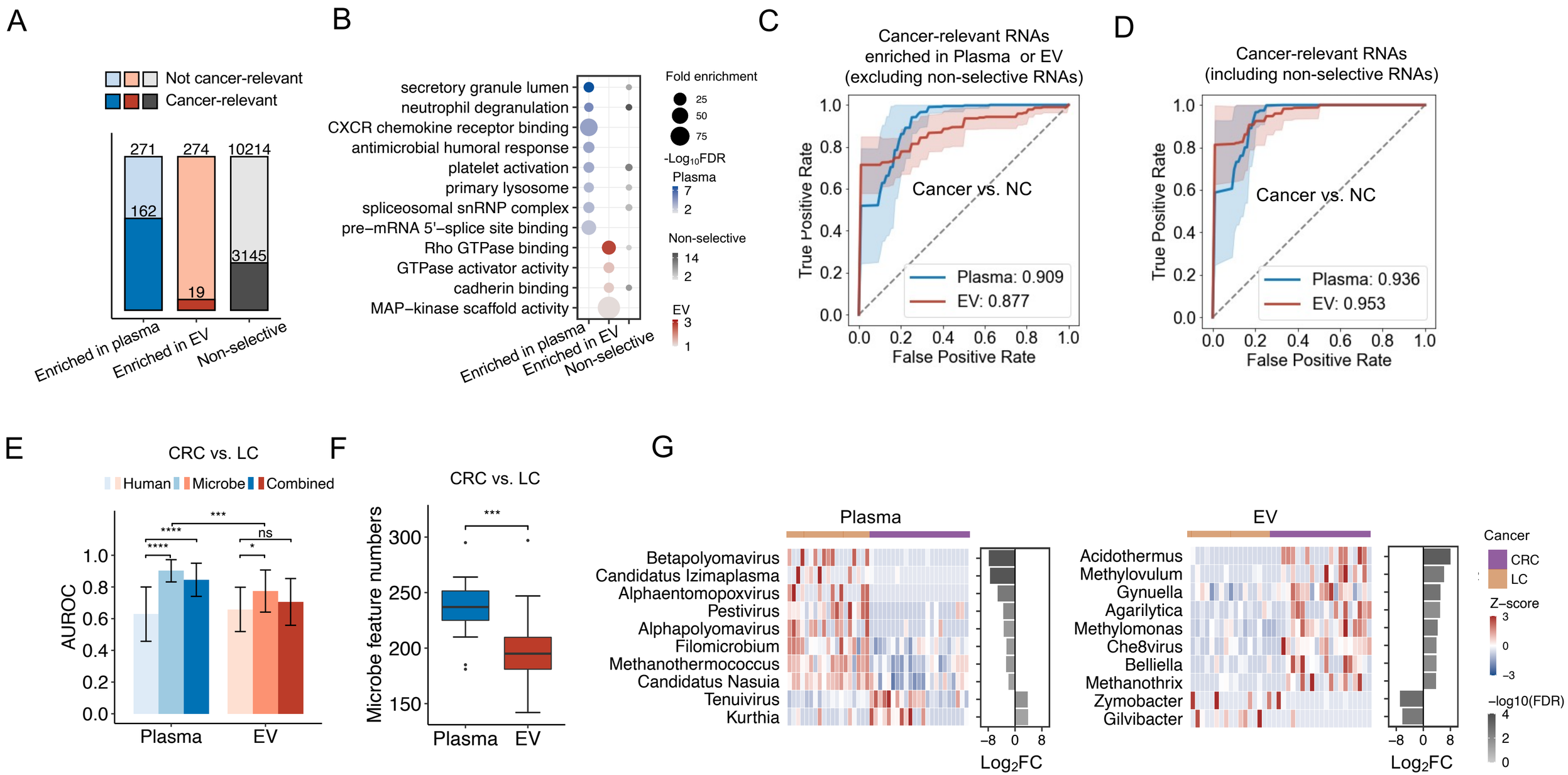

A
Not cancer-relevant
Cancer-relevant
271
274
10214
162
19
3145
Enriched in plasma
Enriched in EV
Non-selective
B
secretory granule lumen
neutrophil degranulation
CXCR chemokine receptor binding
antimicrobial humoral response
platelet activation
primary lysosome
spliceosomal snRNP complex
pre−mRNA 5'−splice site binding
Rho GTPase binding
GTPase activator activity
cadherin binding
MAP−kinase scaffold activity
Enriched in plasma
Enriched in EV
Non-selective
Fold enrichment
25
50
75
-Log10FDR
Plasma
7
2
Non-selective
14
2
EV
3
1
C
Cancer-relevant RNAs
enriched in Plasma or EV
(excluding non-selective RNAs)
Cancer vs. NC
Plasma: 0.909
EV: 0.877
True Positive Rate
False Positive Rate
D
Cancer-relevant RNAs
(including non-selective RNAs)
Cancer vs. NC
Plasma: 0.936
EV: 0.953
True Positive Rate
False Positive Rate
E
CRC vs. LC
Human
Microbe
Combined
AUROC
Plasma
EV
F
CRC vs. LC
Microbe feature numbers
Plasma
EV
G
Plasma
Betapolyomavirus
Candidatus Izimaplasma
Alphaentomopoxvirus
Pestivirus
Alphapolyomavirus
Filomicrobium
Methanothermococcus
Candidatus Nasuia
Tenuivirus
Kurthia
Log2FC
EV
Acidothermus
Methylovulum
Gynuella
Agarilytica
Methylomonas
Che8virus
Belliella
Methanothrix
Zymobacter
Gilvibacter
Log2FC
Cancer
CRC
LC
Z−score
−log10(FDR)

Table 1. Practical reference for cfRNA-seq in human plasma

	***Plasma* cfRNA**	***EV* cfRNA**
EV Purification	**No**	**Yes**
Cost of plasma volume, experimental time and reagents[1]	relatively less	relatively more
Enriched RNA species	circRNA, tRNA, Y RNA	mRNA, srpRNA
Enriched microbes	viruses	intestinimonas, etc.
Diversity of cell-types-of-origin	relatively low	relatively high
TE RNAs[2]	relatively more	relatively less
Cancer detection	good (AUC: 0.94)	good (AUC: 0.95)
Cancer type-specific microbes	relatively more	relatively less
Cancer type classification	relatively good (AUC: 0.90)	relatively poor (AUC: 0.77)

[1] Different cost is due to the EV purification step.
[2] Cell-free RNAs derived from transposable elements.