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

Hongke Wang[1,2,†], Qing Zhan[1,2,†], Hongjie Guo[3,†], Jiuliang Zhao[4,5], Shaozhen Xing[1,2], Shanwen Chen[3], Shuai Zuo[3], Mengtao Li[4,5,*], Pengyuan Wang[3,*], 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] Translational Cancer Research Center, Division of General Surgery, Peking University First Hospital, Beijing 100034, China

[4] Department of Rheumatology and Clinical Immunology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China

[5] National Clinical Research Center for Dermatologic and Immunologic Diseases (Ministry of Science & Technology), MOE Key Laboratory of Rheumatology and Clinical Immunology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Beijing 100730, 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

In biofluid, long RNAs are more informative than microRNAs in terms of gene number and variation type. Therefore, cell-free long RNAs have shown promising potential as biomarkers in liquid biopsy, while they are mostly fragmented. In order to investigate these fragmented cell-free RNAs (cfRNAs), we developed a cost-effective cfRNA sequencing method, DETECTOR-seq (depletion-assisted multiplexing cell-free total RNA sequencing). It utilized a set of customized guide RNAs to remove large amounts of unwanted RNAs (i.e., fragmented ribosomal and mitochondrial RNAs) in human plasma. Early barcoding was also incorporated to save cost and plasma volume. After demonstrating its superior performance to other methods, we used DETECTOR-seq to investigate cell-free transcriptomes in whole human plasma and extracellular vesicles (EVs) it contains. We first observed different type distributions: structured circular RNAs, tRNAs, Y RNAs, and virus RNAs were enriched in plasma, while mRNAs and srpRNAs were enriched in EVs. We also uncovered distinct functional pathways: RNA splicing-related ribonucleoproteins (RNPs) and antimicrobial humoral response genes were enriched in plasma, while transcriptional activity, cell migration, and antigen receptor-mediated immune signals were enriched in EVs. Subsequently, we compared the performances of these distinct cfRNAs in whole plasma versus EVs on classifying cancer patients. The accuracies were comparable when discriminating cancer patients from healthy donors (AUCs: 0.936 versus 0.953). Meanwhile, cancer types (i.e., colorectal versus lung cancer) were better classified with microbial cfRNAs in plasma

than in EV (AUCs: 0.898 versus 0.772). Overall, by investigating total and EV cfRNAs in the pairwise plasma samples, our work provides practical guidance for the proper decision of EV purification when launching a cfRNA-based study. Furthermore, as a cost-effective method, DETECTOR-seq would facilitate transcriptome-wide studies in the fields of extracellular RNA biology and clinical liquid biopsy.

Key Points

1. *DETECTOR-seq enables efficient and specific depletion of sequences derived from fragmented ribosomal and mitochondrial RNAs in plasma.*
2. *Distinct cfRNA signatures in whole plasma versus EVs were revealed.*
3. *Both Plasma and EV cfRNAs were capable of distinguishing cancer patients from normal individuals.*
4. *Microbial RNAs in Plasma cfRNAs enabled better classification of cancer types than EV cfRNAs.*

Key Words

Cell-free RNA; Extracellular vesicle; Exosome; Cancer classification; Liquid biopsy

Introduction

In recent years, liquid biopsy has emerged as a non-invasive approach for assessing circulating biomarkers in various body fluids to monitor physiologic and disease states [1]. Cell-free RNAs (cfRNAs), given their virtue of being highly dynamic, hold great potential to reflect the pathophysiological processes, thus offering unique opportunities for disease monitoring. Previous reports have suggested that cfRNAs are packaged into various extracellular complexes, such as extracellular vesicles (EVs, including micro-vesicles and exosomes) and non-vesicular ribonucleoproteins (RNPs) [2]. Due to the protection of EV, RNA binding proteins, and/or their self-structures, cfRNAs are capable of being stably present in human bloodstream [3]. Most cfRNA studies investigated either total [4-6] or EV [7-9] RNAs in plasma, while gain and loss of EV purification in liquid biopsy are still under debate. For instance, it is not clear whether the EV purification step is needed in cfRNA-based cancer screening tests.

Efforts in characterizing cfRNAs were initially focused on small RNAs like microRNAs (miRNAs) because of the nature of RNA degradation and fragmentation in biofluids. However, miRNAs represent only a small proportion of the human transcriptome [10]. Therefore, a broader space of cfRNAs, such as messenger RNAs (mRNAs), long non-coding RNAs (lncRNAs), and circular RNAs (circRNAs) have started to be investigated later [4-7, 11]. Due to RNases in human blood, these cell-free long RNA species (>50 nt) have relatively low concentrations. They are usually fragmented (~50–200 nucleotides), lacking intact RNA ends [12]. The conventional small RNA-seq

83 approach, which ligates sequencing adapters based on RNA ends of 5' phosphate (5' P)
84 and 3' hydroxyl (3' OH), does not work well for these fragmented cfRNAs [13].

85 Recently, several sequencing approaches have been developed to profile cell-free
86 long RNA fragments. Phospho-RNA-seq incorporates T4 polynucleotide kinase into
87 ligation-based TruSeq small RNA-seq [12]. Thus, it can recover mRNA and lncRNA
88 fragments lacking 5' P and/or 3' OH ends. However, as the authors mentioned, the
89 libraries of Phospho-RNA-seq contained high fractions of ribosomal RNAs (rRNAs) and
90 Y RNAs, reducing the capacity to detect other informative RNA species [12]. Another
91 method, SILVER-seq, captures both small and long cfRNAs from extremely low-input
92 serum samples [14]. However, substantial DNA contamination seemed to be an issue of
93 SILVER-seq [15]. Recently, SMARTer stranded total RNA-seq (hereafter called
94 SMARTer-seq) was used in several cfRNA studies [4-7], where unwanted ribosomal
95 sequences were depleted using a proprietary R-probe-based system called ZapR [16,
96 17]. However, as a commercial kit, SMARTer-seq was not specifically optimized for
97 cfRNA library in plasma. It is not cost-efficient either. Overall, the current cfRNA
98 sequencing approaches were limited by unwanted RNAs, DNA contamination, and high
99 cost.

100 In this study, we present an optimized cfRNA sequencing method, DETECTOR-seq
101 (depletion-assisted multiplexing cell-free total RNA sequencing), which utilizes early
102 barcoding and CRISPR-Cas9 to reduce the cost and high-abundant, fragmented rRNAs
103 and mitochondrial RNAs (mtRNAs) in human plasma. Then, we used DETECTOR-seq to

104 investigate 113 plasma cfRNA samples (including 61 plasma total RNA and 52 EV RNA
105 libraries) derived from healthy donors, lung and colorectal cancer patients. To the best of
106 our knowledge, this study is the first to compare pairwise total and EV-selected
107 transcriptomes in the same plasma samples, suggesting their distinct signatures and
108 different utilities in the liquid biopsy of cancer.
109
110

111 # Methods

112 **Issues of sequencing cell-free RNAs**

113 The sequencing of cfRNAs in plasma and other biofluids usually meets the following
114 obstacles. First, consistent with previous reports [10], we observed that plasma cfRNAs
115 were degraded with a fragment length of <200 nucleotides **(Figure 1A)**. These
116 fragmented cfRNAs are hard to be detected by many RNA-seq protocols based on
117 ligation techniques requiring intact RNA ends. Second, ribosomal RNAs (rRNAs) and
118 mitochondrial RNAs (mtRNAs) accounted for ~92% of all clean reads (reads after
119 removing adapters and filtering low-quality reads), while messenger RNAs (mRNAs) and
120 long non-coding RNAs (lncRNAs) collectively made up only a small fraction (~4%) of
121 cell-free transcriptome **(Figure 1B)**. It is worth noting that microbe-derived RNAs can
122 also be detected in human plasma with a relatively small fraction (~0.4%) **(Figure 1B)**.
123 The high fractions of rRNAs and mtRNAs hamper the detection of other informative RNA
124 species. And they are fragmented into pieces in plasma, making them hard to be

125 removed **(Figures 1C, D)**. Third, cfRNAs are usually in the range of hundred picograms
126 to several nanograms per ml of human plasma [14], which can be easily lost and
127 contaminated during purification and amplification. For instance, low cfRNA input usually
128 requires 20-24 PCR amplification cycles for library preparation, which produces a high
129 duplication ratio of raw reads. Meanwhile, DNA contamination ignorable in conventional
130 RNA-seq is often over-amplified, causing a big issue in cfRNA-seq [15].
131
132 **DETECTOR-seq**
133 To improve the efficiency and reliability of cfRNA detection, we developed DETECTOR-
134 seq (depletion-assisted multiplexing cell-free total RNA sequencing) to profile cell-free
135 transcriptome in human plasma **(Figures 1E, F)**. DETECTOR-seq captures fragmented
136 cfRNAs with unbiased random priming and template-switching. Then, it uses CRISPR-
137 Cas9 to remove the abundant sequences derived from ribosomal and mitochondrial
138 RNAs in the complementary DNA (cDNA) library. In this step, guide RNAs (sgRNAs) in
139 the CRISPR-Cas9 are specifically optimized for human plasma cfRNAs **(Supplementary**
140 **Figures 1,2)**, covering the fragmented rRNA and mtRNA sequences **(Figures 1D, E)**.
141 The sgRNAs are in vitro transcribed using T7 RNA polymerase, then bind with Cas9
142 nuclease to form ribonucleoprotein (RNP) complex and induce site-specific cleavage
143 with the endonuclease activity of Cas9 **(Figure 1E)**, thus preventing further amplification
144 of cDNAs derived from rRNAs and mtRNAs in the final sequencing library. Meanwhile,
145 DETECTOR-seq utilizes early barcoding during reverse transcription. The multiplexed

library will cope with low content of plasma cfRNAs, and reduce experimental time and cost as well. It is also worth mentioning that unique molecular identifiers (UMIs) are added to every sequence in the reverse transcription step, hence DETECTOR-seq is capable of removing PCR duplicates to avoid RNA quantification bias **(Figure 1F)**. In addition, we also optimized cfRNA extraction **(Supplementary Figure 3)** and residual DNA digestion **(Supplementary Figure 4)** protocols.

Depletion of rRNA and mtRNA sequences in human plasma

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

well correlated (Pearson correlation, R=0.90, *P*-value < $2.2\times^{-16}$; **Figure 2D**). In summary, the above results demonstrate the efficient and specific depletion of unwanted sequences in DETECTOR-seq.

Results

Analytical validation analysis demonstrating high-quality reads of DETECTOR-seq

To evaluate the performance of DETECTOR-seq, we prepared cfRNA libraries in a 3-plex, 4-plex, or 5-plex manner determined by RNA concentrations. The total read numbers of different barcoded samples in one multiplexing pool were relatively uniform, varying less than 1.5-fold in the 3- and 4-plex samples and less than 3-fold in the 5-plex samples **(Figure 3A)**. In addition, the UMI strategy in DETECTOR-seq retained significantly more reads than the non-UMI approach after duplicated reads were removed **(Figure 3B)**. And a sharp edge of reads' distribution across exon-intron splice junctions suggested that the majority of DNA contamination was effectively removed **(Figure 3C)**. To evaluate the impact of plasma input volume on the number of detected genes, we sequenced cfRNAs with 200, 400, 600, 800, and 1000 μL of plasma aliquots from the same individual with five biological replicates. Around 4000 genes were detected with the minimum (i.e., 200 μL) volume. The detected gene number linearly increased until a plateau between 800 and 1000 μL, suggesting the detected genes would be saturated after 1 mL of plasma **(Figure 3D)**. While highly correlated cfRNA

188 expression levels were observed within technical triplicates (R1-R3), the correlations
189 were slightly decreased between biological triplicates (N1-N3) **(Figure 3E)**. Furthermore,
190 based on ERCC RNA Spike-In Mix, we found a high correlation between expected and
191 observed levels of transcript abundance (Pearson correlation, R=0.91, *P*-value < $2.2\times^{-16}$;
192 **Figure 2F**). These results not only demonstrate DETECTOR-seq's high quality and
193 reproducibility but also suggest its capability of capturing subtle differences in cfRNA
194 profiles between different individuals.

195 Then, we randomly subsampled a dataset (n=24) of DETECTOR-seq for saturation
196 analyses of detected UMIs (transcripts) and genes. Although the detected UMIs kept
197 increasing when more reads in 1ml plasma were sequenced **(Figure 3G)**, the detected
198 gene numbers were quickly saturated at approximately 5 million genome-aligned reads
199 **(Figure 3H)**. These results indicate that DETECTOR-seq achieves saturation of cfRNA
200 detection at a low sequencing depth.

201

202 **Better contamination control and cost-effectiveness of DETECTOR-seq than other**
203 **cfRNA-seq methods**

204 We benchmarked the performance of DETECTOR-seq compared to three other cfRNA-
205 seq methods, including Phospho-RNA-seq [12], SILVER-seq [14], and SMARTer-seq
206 [18]. Within the total genome-aligned reads, DETECTOR-seq and SMARTer-seq had
207 comparable ratios of exonic reads (~70%), while those of SILVER-seq and Phospho-
208 RNA-seq were under 40% **(Figure 4A)**. The lower ratio of exonic reads for SILVER-seq

was presumably due to severe DNA contamination according to a previous report [15]. We also visualized the read coverage across exon boundary sites flanked upstream and downstream by 50 bp, where DETECTOR-seq and SMARTer-seq showed more evident decreases of read coverage from exon to intron/intergenic region than SILVER-seq and Phospho-RNA-seq **(Figure 4B)**. As far as we know, all of the four cell-free RNA-seq methods should preserve the strand specificity of RNAs. Thus, the enrichment of exons' sense over antisense reads of DETECTOR-seq and SMARTer-seq further confirmed their reads' quality **(Figure 4C)**. The above results demonstrate that DETECTOR-seq and SMARTer-seq have better DNA contamination control than SILVER-seq. It was worth noting that Phospho-RNA-seq was developed from a small RNA-seq method, and the read coverage across exon boundary sites and the enrichment of exons' sense over antisense reads may be affected by the read distribution of small RNAs.

In addition, we showed that DETECTOR-seq displayed a higher ratio of reads mapped to human genome (~71%) than those of SMARTer-seq (~48%) because DETECTOR-seq removed mitochondrial RNAs more efficiently than SMARTer-seq **(Figure 4D)**. Furthermore, because of its early barcoding and multiplexing strategy, DETECTOR-seq can produce more raw reads and genome-aligned reads than the other cfRNA-seq approaches **(Figure 4E, Supplementary Figure 5)**. Overall, by summarizing and comparing key characteristics of these approaches **(Figure 4F)**, we collectively demonstrate that DETECTOR-seq has better contamination control and more efficient cost than the other cfRNA-seq methods.

Distinct human and microbial RNA signatures in plasma versus extracellular vesicle

Subsequently, we utilized DETECTOR-seq to comprehensively investigate total cfRNAs and EV cfRNAs pairwise in human plasma **(Figure 5A)**. A proportion of cfRNAs are enclosed inside EVs such as MVs and exosomes [19]. Meanwhile, it is also reported that a significant proportion of cfRNAs are not within EVs but associated with proteins to form non-vesicular RNPs [20]. Although both total cfRNAs [4-6] and EV cfRNAs [7, 9] have been used in the liquid biopsy studies, their distinct signals and utilities have not been pairwise compared yet.

In total, we sequenced 139 plasma cfRNA samples derived from healthy donors, lung cancer and colorectal cancer patients (**Supplementary Figure 6**). Then, 113 datasets passed quality control (QC) procedures of RNA samples and sequencing data (**Supplementary Figures 6-8**). Among them, 61 were total cfRNA-seq and 52 were EV cfRNA-seq, where 44 were paired from the same plasma samples. In the following description, total cfRNA-seq of plasma and EV cfRNA-seq of plasma will be abbreviated to *Plasma* cfRNA and *EV* cfRNA, respectively.

From a general view, there was a high degree of similarity between *Plasma* and *EV* cfRNAs, with ~90% of aligned reads mapping to human genome and ~10% mapping to microbe genomes **(Figure 5B)**. For human cfRNAs, mRNA, lncRNA, and circRNA were the major RNA types. For microbial cfRNAs, the most abundant phylum was

Proteobacteria, followed by *Firmicutes* and *Actinobacteria*. The human and microbial RNA compositions resembled previous reports [18, 21].

In addition, distinctive signatures were revealed for the first time by our pairwise comparison between *Plasma* and *EV* cfRNAs. We first observed that *Plasma* cfRNAs had more short fragments (50~100 nt), while *EV* cfRNAs had more long fragments (>100 nt) **(Supplementary Figure 12)**. We also observed that structured tRNAs, Y RNAs, and circRNAs were significantly enriched in *Plasma* cfRNAs, while mRNAs and signal recognition particle RNAs (srpRNAs) were significantly enriched in *EV* cfRNAs **(Figure 5C)**. This is consistent with a previous study reporting that tRNA and Y RNA fragments were significantly enriched in extracellular RNPs [2]. Moreover, we also found that the relative abundance of circRNAs was significantly higher in *Plasma* cfRNAs than *EV* cfRNAs **(***P*-value < 0.0001, Wilcoxon rank sum test; **Figure 5C**, **Supplementary Figure 9)**, perhaps due to its circle-like structure resisting degradation outside of EVs. We totally identified 13 circRNAs differentially enriched in *Plasma* versus *EV* cfRNAs. Only one of them, hsa_circ_0048555, was enriched in EVs **(Supplementary Figure 10)**. Reads mapped to the back-spliced junction were used to calculate the enrichment.

A recent study provided a framework to infer cell types of origin of the cell-free transcriptome [22]. We utilized this method and found a high similarity of the cell types of origin between *Plasma* and *EV* transcriptomes **(Figure 5D)**. Platelets and erythrocytes were inferred as the major origins for both *Plasma* and *EV* cfRNAs, which was in agreement with the previous study [22]. Intriguingly, we found non-blood cells

contributed more to *EV* cfRNAs than to *Plasma* cfRNAs (*P*-value < 0.01, Wilcoxon rank sum test; **Figure 5D**). Therefore, the diversities of cell types of origin (measured by Simpson's index) of *EV* cfRNAs were significantly higher than those of *Plasma* cfRNAs (*P*-value < 0.01, Wilcoxon rank sum test; **Figure 5D**).

We also identified distinct microbe genera in *Plasma* and *EV* cfRNAs **(Supplementary Figure 11)**. Although there was no significant difference between the ratio of microbe reads in *Plasma* and *EV* cfRNAs, we found cfRNAs mapped to virus genomes were significantly elevated in *Plasma* cfRNAs **(Figure 5E)**. Meanwhile, viruses such as *Senecavirus*, *Cheravirus*, *Orthopoxvirus*, *Tenuivirus*, and *Rhadinovirus* were significantly enriched in *Plasma* cfRNAs, while *Intestinimonas*, *Mordavella*, and *Jonquetella* were significantly enriched in *EV* cfRNAs **(Figure 5F)**. In summary, the above comparison results have revealed distinct molecular characteristics between *Plasma* and *EV* cfRNAs in terms of fragment size, RNA species, cell types of origin, and microbe genera.

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

To find selective functions and motifs of cfRNAs in EVs, we identified 545 selectively distributed RNAs showing significantly differential abundance between *Plasma* and *EV* transcriptomes **(**|Fold-change| >1 and FDR < 0.1; **Figure 6A, Supplementary Figure 13)**. Among them, 271 cfRNAs were enriched in *Plasma*, while 274 cfRNAs were enriched in *EVs*. We investigated the functional roles and biological pathways of these

293 selective cfRNAs **(Figure 6B, Supplementary Figure 14)**. Based on KEGG pathway
294 enrichment analysis, we found that the selective RNAs elevated in *Plasma* were
295 significantly enriched in terms associated with RNA splicing, RNP (e.g., mRNA 5' splice
296 site recognition, U1 snRNP, spliceosomal snRNP complex and Sm-like protein family
297 complex), antimicrobial and innate immune responses. Meanwhile, the selective RNAs
298 that were enriched in *EVs* were primarily associated with DNA binding transcription
299 factor activity, focal adhesion, cell-substrate junction, and T cell receptor signaling
300 immune pathway. Notably, we found different immune pathways enriched in the selective
301 cfRNAs of *Plasma* versus *EVs* **(Figure 6B, Supplementary Figure 14)**.

302 We further investigated sequence motifs and their associated RNA binding proteins
303 (RBPs) for the selective cfRNAs **(Figure 6C, Supplementary Figure 15)**. And we found
304 that the selective cfRNAs enriched in *Plasma* contained binding motifs/sites for ABCF1,
305 a protein that plays a role in innate immune response [23]; SFPQ, a splicing factor;
306 LARP4, a La RNP; TROVE2, a Y RNA binding protein; and DKC1, a snoRNP. Meanwhile,
307 the selective cfRNAs enriched in *EVs* contained binding motifs/sites for PUM1, a protein
308 that participates in human innate immune response [24]; BCLAF1, a transcription factor;
309 HNRNPU, a transcription suppressor; PCBP1, a previously reported immune checkpoint
310 [25]; APOBEC3C, an RNA editing enzyme. These enriched motifs and their associated
311 RBPs were consistent with the biological functions of the selective cfRNAs revealed
312 above.

313

Specific cancer signals revealed in *Plasma* and *EV* cfRNAs

In order to demonstrate whether the EV purification step is needed for a cfRNA-based cancer test, we compared the cancer classification performance of *Plasma* cfRNAs and *EV* cfRNAs. We sequenced cfRNAs in the plasma samples of lung cancer (LC) and colorectal cancer (CRC) patients, as they are two major cancer types **(Supplementary Figure 6)**. Based on differential expression analysis between cancer and normal control (NC) ($|\log_2$fold-change$|>1$ and FDR<0.05), we defined a set of cancer-relevant cfRNAs **(Supplementary Figure 16)**. Interestingly, we found that they were remarkably enriched in *Plasma* compared to *EVs* **(Figure 7A).** We also found that enriched functions of these cancer-relevant *Plasma* cfRNAs were termed as RNA splicing, snRNP signals, etc **(Figure 7B)**. This is consistent with the enriched pathways of *Plasma* cfRNAs revealed in **Figure 6B**.

Based on these selectively distributed cancer-relevant cfRNAs, we used a random forest classifier to discriminate cancer patients from NCs. Although the selective cfRNAs in *Plasma* performed slightly better than those in *EVs* (average AUROC: 0.909 versus 0.877, **Figure 7C, Supplementary Figure 17)**, comparable performances were observed between *Plasma* and *EV* cfRNAs when a large number of non-selective cfRNAs **(Figures 6A, 7A)** were included as well (average AUROC: 0.936 versus 0.953, **Figure 7D, Supplementary Figure 18)**. Collectively, these results imply that the EV purification step can reveal distinct cancer signals, but it has a very subtle effect on the accuracy of detection of cancer patients from healthy controls.

335

336 **Microbial cfRNAs in *Plasma* classify cancer types better than *EV* cfRNAs**

337 Cancer screening test not only requires detecting cancer patients from normal controls
338 but also needs to determine primary tumor locations. Therefore, we further compared the
339 performances of human cfRNAs in *Plasma* and *EV* for classifying CRC from LC.

340 First, we found none of them did a good job (average AUROC: 0.628 versus 0.659,
341 **Figure 7E, Supplementary Figure 19)**. Fortunately, a recent study of our group
342 revealed that microbe-derived cfRNAs in human plasma reflect cancer-type-specific
343 information [18]. Based on the RNA abundance levels of the contamination-filtered
344 microbe genera, we found the microbial cfRNAs significantly improved the classification
345 for both *Plasma* and *EV* cfRNA data (average AUC: 0.898 versus 0.772, **Figure 7E,**
346 **Supplementary Figure 19)**.

347 Notably, the microbial reads in *Plasma* cfRNAs performed better than those in *EV*
348 cfRNAs. Consistently, we also found more cancer-type-specific features in *Plasma*
349 cfRNAs than in *EV* cfRNAs **(Figure 7F)**. We identified the microbial features recurrently
350 showing differential abundance between CRC and LC in all of the 20 bootstrap
351 samplings. The abundance of top recurrent microbe genera, along with fold-change and
352 false discovery rates were illustrated **(Figure 7G)**. For instance, more *Methanothrix*
353 reads were found in CRC than in LC using *EV* cfRNA-seq data. This is consistent with a
354 previous study reporting that *Methanothrix soehngenii* was enriched in gut microbiome of
355 CRC patients [26]. Meanwhile, many cancer-relevant virus RNAs in *Plasma* cfRNAs

356 classified cancer types, consistent with the observation of more virus RNAs detected in
357 *Plasma* than in *EVs* **(Figure 5E)**. For instance, more reads of *Alpha-polyomavirus* and
358 *Beta-polyomavirus* were found in LC than in CRC using *Plasma* cfRNA-seq data.
359 Supportively, some polyomaviruses were also reported to be detectable in
360 gastrointestinal tract and respiratory aspirates [27]. These studies and results suggest
361 that microbe-derived cfRNAs in *Plasma* and *EV* present promising but yet poorly
362 investigated signatures for specific cancer types.
363
364

365 Conclusion and Discussion

366 **Conclusions**. In summary, this study introduced a depletion-assisted cost-effective
367 cfRNA profiling approach, termed DETECTOR-seq, which utilized multiple technologies
368 such as early barcoding, template-switching, UMI, and sgRNA/CRISPR-Cas9. Using
369 DETECTOR-seq, we recapitulated molecular characteristics of *Plasma* and *EV* cfRNAs
370 and identified their distinct human and microbial signatures, thus illustrating the gain and
371 loss of certain cfRNA signals due to EV purification. Our work provides a practical guide
372 for cfRNA-based liquid biopsy **(Table 1)**. Moreover, we envision that DETECTOR-seq
373 would be a useful tool to facilitate further studies in extracellular RNA biology.
374

375 **Technologies utilized and optimized in DETECTOR-seq**. Plasma cell-free
376 transcriptome remains challenging to study owing to the low quantity and quality of

fragmented RNAs [11]. Over-represented rRNA and mtRNA species [12], DNA contamination [15], and high cost are still the major issues of cfRNA sequencing. Multiple technologies were included in DETECTOR-seq to address these issues **(Figure 4F)**. First, DETECTOR-seq captures fragmented cfRNAs with random priming and template-switching strategies, which have been proven to be highly efficient in single-cell RNA-seq [28]. Second, the early barcoding protocol of DETECTOR-seq enables us to prepare cfRNA libraries in a multiplexing manner, thus reducing the volume of required plasma and experimental costs. In fact, DETECTOR-seq is capable of detecting cfRNAs with a low input volume of 0.2 to 1 mL plasma with a 2- to 6-fold cost saving compared to existing approaches. Third, with UMIs tagging to cDNAs of RNA fragments, DETECTOR-seq can accurately quantify the low-quantity cfRNAs. Fourth, by optimizing the procedures of RNA extraction and residual DNA digestion (**Supplementary Figures 3-4**), DETECTOR-seq avoids the potential contamination of genomic DNAs. Fifth, DETECTOR-seq uses CRISPR-Cas9 technology to deplete rRNA and mtRNA sequences. A CRISPR-based depletion strategy, DASH (Depletion of Abundant Sequences by Hybridization) [29] has been utilized in other fields, such as ATAC-seq [30], small RNA-seq [31], bacterial RNA-seq [32] and single-cell total RNA-seq [33]. Here, we applied this CRISPR-based method to cfRNA sequencing and designed a specific set of sgRNAs for human plasma **(Supplementary Figures 1,2)**.

***Plasma* vs. *EV* in cancer detection and cancer type classification**. Researchers have used both *Plasma* cfRNA-seq [4-6] and *EV* cfRNA-seq [7, 19, 34-36] to identify

disease biomarkers. But which one is better is still not clear. By pairwise comparison between *Plasma* and *EV* cfRNA-seq, we found that both of them can distinguish cancer patients from controls with comparable performance. However, cancer types can be better classified with microbe-derived features in *Plasma* cfRNAs than those in *EV* cfRNAs.

Distinct signatures in *Plasma* vs. *EV* cfRNAs. Moreover, this study has brought new insights into distinct cfRNA signatures in *Plasma* versus *EVs*. *Plasma* contains miscellaneous cfRNAs released from alive or apoptotic cells, while RNAs in *EV* cargos are considered to be secreted actively by cells for functional roles in intercellular communications [37]. This study revealed distinct biological pathways, enriched motifs, and RBP-binding sites in *Plasma* vs. *EV* cfRNAs. We also found that short RNA fragments (50 to 100 nt) associated with RNPs were enriched in *Plasma* cfRNAs, indicating higher degradation extent of non-vesicular RNAs than those of EV RNAs.

Limitations of this study. Though DETECTOR-seq provides several advantages when compared with other approaches, it needs to be further improved. For example, the efficiency of random priming of DETECTOR-seq is proportional to the fragment length of RNAs, which will bias the library. Meanwhile, DETECTOR-seq includes several purification steps to remove by-products, such as empty library constructs, adapter dimers, and superfluous primers. Because these purification procedures retain longer products, RNA fragments shorter than 50 nucleotides are largely discarded along with by-products. Thus, DETECTOR-seq can be modified based on other strategies like

poly(A) tailing to obtain a complete spectrum of cfRNAs, including both small and long fragments [38, 39].

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.

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).

Competing interests

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

Funding

This work is supported by Tsinghua University Spring Breeze Fund (2021Z99CFY022), National Natural Science Foundation of China (81972798, 32170671, 81902384), National Key Research and Development Plan of China (2019YFC1315700), National Science and Technology Major Project of China (2018ZX10723204, 2018ZX10302205),

Tsinghua University Guoqiang Institute Grant (2021GQG1020), Tsinghua University Initiative Scientific Research Program of Precision Medicine (2022ZLA003), Bioinformatics Platform of National Center for Protein Sciences (Beijing) (2021-NCPSB-005). This study is also supported by Beijing Advanced Innovation Center for Structural Biology, Bio-Computing Platform of Tsinghua University Branch of China National Center for Protein Sciences, Interdisciplinary Clinical Research Project of Peking University First Hospital, and the Capital Health Research and Development of Special, Open Research Fund Program of Beijing National Research Center for Information Science and Technology. *Funding for open access charge: Tsinghua University Spring Breeze Fund [2021Z99CFY022].*

References

1. Heitzer E, Haque IS, Roberts CES, Speicher MR: **Current and future perspectives of liquid biopsies in genomics-driven oncology**. *Nature Reviews Genetics* 2019, **20:**71-88.

2. Wei ZY, Batagov AO, Schinelli S, Wang JT, Wang Y, El Fatimy R, Rabinovsky R, Balaj L, Chen CC, Hochberg F, et al: **Coding and noncoding landscape of extracellular RNA released by human glioma stem cells.** *Nature Communications* 2017, **8:**1-15.

3. Gruner HN, McManus MT: **Examining the evidence for extracellular RNA function in mammals.** *Nature Reviews Genetics* 2021, **22:**448-458.

4. Moufarrej MN, Vorperian SK, Wong RJ, Campos AA, Quaintance CC, Sit RV, Tan M, Detweiler AM, Mekonen H, Neff NF, et al: **Early prediction of preeclampsia in pregnancy with cell-free RNA.** *Nature* 2022, **602:**689-694.

5. Rasmussen M, Reddy M, Nolan R, Camunas-Soler J, Khodursky A, Scheller NM, Cantonwine DE, Engelbrechtsen L, Mi JD, Dutta A, et al: **RNA profiles reveal signatures of future health and disease in pregnancy.** *Nature* 2022, **601:**422-427.

6. Ngo TTM, Moufarrej MN, Rasmussen MLH, Camunas-Soler J, Pan WY, Okamoto J, Neff NF, Liu KL, Wong RJ, Downes K, et al: **Noninvasive blood tests for fetal development predict gestational age and preterm delivery.** *Science* 2018, **360:**1133-1136.

7. Yu S, Li Y, Liao Z, Wang Z, Wang Z, Li Y, Qian L, Zhao J, Zong H, Kang B, et al: **Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma.** *Gut* 2020, **69:**540-550.

8. Li YC, Zhao JL, Yu SL, Wang Z, He XG, Su YH, Guo TAN, Sheng HY, Chen J, Zheng QP, et al: **Extracellular Vesicles Long RNA Sequencing Reveals Abundant mRNA, circRNA, and lncRNA in Human Blood as Potential Biomarkers for Cancer Diagnosis.** *Clinical Chemistry* 2019, **65**:798-808.

9. Ji J, Chen R, Zhao L, Xu YL, Cao Z, Xu H, Chen X, Shi XL, Zhu YS, Lyu J, et al: **Circulating exosomal mRNA profiling identifies novel signatures for the detection of prostate cancer.** *Molecular Cancer* 2021, **20**:58.

10. Larson MH, Pan WY, Kim HJ, Mauntz RE, Stuart SM, Pimentel M, Zhou YQ, Knudsgaard P, Demas V, Aravanis AM, Jamshidi A: **A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection.** *Nature Communications* 2021, **12**:1-11.

11. Cabus L, Lagarde J, Curado J, Lizano E, Perez-Boza J: **Current challenges and best practices for cell-free long RNA biomarker discovery.** *Biomarker Research* 2022, **10**:62.

12. Giraldez MD, Spengler RM, Etheridge A, Goicochea AJ, Tuck M, Choi SW, Galas DJ, Tewari M: **Phospho-RNA-seq: a modified small RNA-seq method that reveals circulating mRNA and lncRNA fragments as potential biomarkers in human plasma.** *EMBO J* 2019, **38**:e101695.

13. Akat KM, Lee YA, Hurley A, Morozov P, Max KE, Brown M, Bogardus K, Sopeyin A, Hildner K, Diacovo TG, et al: **Detection of circulating extracellular mRNAs by modified small-RNA-sequencing analysis.** *JCI Insight* 2019, **5**:e127317.

14. Zhou Z, Wu Q, Yan Z, Zheng H, Chen CJ, Liu Y, Qi Z, Calandrelli R, Chen Z, Chien S, et al: **Extracellular RNA in a single droplet of human serum reflects physiologic and disease states.** *Proc Natl Acad Sci U S A* 2019, **116:**19200-19208.

15. Verwilt J, Trypsteen W, Van Paemel R, De Preter K, Giraldez MD, Mestdagh P, Vandesompele J: **When DNA gets in the way: A cautionary note for DNA contamination in extracellular RNA-seq studies.** *Proc Natl Acad Sci U S A* 2020, **117:**18934-18936.

16. Stark R, Grzelak M, Hadfield J: **RNA sequencing: the teenage years.** *Nature Reviews Genetics* 2019, **20:**631-656.

17. Farmer AA, Betts C, Bolduc N: **Methods of depleting a target molecule from an initial collection of nucleic acids, and compositions and kits for practicing the same.** 2018.

18. Chen S, Jin Y, Wang S, Xing S, Wu Y, Tao Y, Ma Y, Zuo S, Liu X, Hu Y, et al: **Cancer type classification using plasma cell-free RNAs derived from human and microbes.** *Elife* 2022, **11:**e75181.

19. Moller A, Lobb RJ: **The evolving translational potential of small extracellular vesicles in cancer.** *Nat Rev Cancer* 2020, **20:**697-709.

20. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, Mitchell PS, Bennett CF, Pogosova-Agadjanyan EL, Stirewalt DL, et al: **Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma.** *Proc Natl Acad Sci U S A* 2011, **108:**5003-5008.

21. Pan WY, Ngo TTM, Camunas-Soler J, Song CX, Kowarsky M, Blumenfeld YJ, Wong RJ,

Shaw GM, Stevenson DK, Quake SR: **Simultaneously Monitoring Immune Response and Microbial Infections during Pregnancy through Plasma cfRNA Sequencing.** *Clinical Chemistry* 2017, **63**:1695-1704.

22. Vorperian SK, Moufarrej MN, Tabula Sapiens C, Quake SR: **Cell types of origin of the cell-free transcriptome.** *Nat Biotechnol* 2022, **40**:855-861.

23. Lee MN, Roy M, Ong SE, Mertins P, Villani AC, Li WB, Dotiwala F, Sen J, Doench JG, Orzalli MH, et al: **Identification of regulators of the innate immune response to cytosolic DNA and retroviral infection by an integrative approach.** *Nature Immunology* 2013, **14**:179-185.

24. Liu YH, Qu LL, Liu YY, Roizman B, Zhou GG: **PUM1 is a biphasic negative regulator of innate immunity genes by suppressing LGP2.** *Proceedings of the National Academy of Sciences of the United States of America* 2017, **114**:6902-6911.

25. Ansa-Addo EA, Huang HC, Riesenberg B, Iamaswat S, Borucki D, Nelson MH, Nam JH, Chung 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.

26. 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.

27. 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.

28. 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.

29. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL: **Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications.** *Genome Biol* 2016, **17:**1-13.

30. 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.

31. 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.

32. 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.

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

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. Isakova A, Neff N, Quake SR: **Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states.** *Proceedings of the National Academy of Sciences of the United States of America* 2021, **118**.

Figure Legends

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

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

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

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

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

Figure 3 | Analytical validation analysis of DETECTOR-seq's performance.

(A) Number of sequenced reads of each barcoded sample in each multiplexing library. The dashed line represents an expected number. **(B)** The number of collapsed reads with PCR duplicates removed by UMI or non-UMI methods. ****: *P*-value<0.0001, Wilcoxon rank sum test, two-tailed. **(C)** Average coverage across all the 5' and 3' exon boundary sites flanking upstream and downstream by 50 bp. **(D)** The number of detected genes in DETECTOR-seq libraries (n=5) with different input volumes of plasma. **(E)** Pearson correlation matrix of plasma samples from biological triplicates (N1–N3) and technical triplicates (R1–R3). **(F)** Pearson correlation between spike-in molecules and their reads sequenced by DETECTOR-seq for ERCC spike-in controls. **(G)** Numbers of detected UMIs and **(H)** detected genes (defined by three different minimum counts) at various subsampled genome-aligned read depths. The error bar represents the standard deviation of multiple samples (n=24). M: million.

Figure 4 | Comparing DETECTOR-seq with other cfRNA-seq methods.

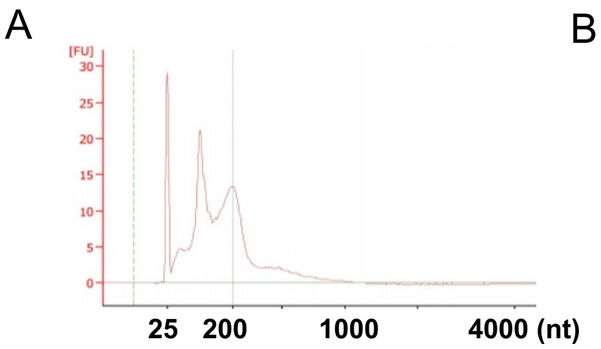
(A) Average percentages of genome-aligned reads mapping to exonic, intronic, and intergenic regions for four different cfRNA-seq methods. **(B)** Average coverage across all mRNAs' 5' and 3' exon boundary sites flanking upstream and downstream by 50 bp. **(C)** Average percentages of reads located in the sense and antisense strands of mRNAs' exons,

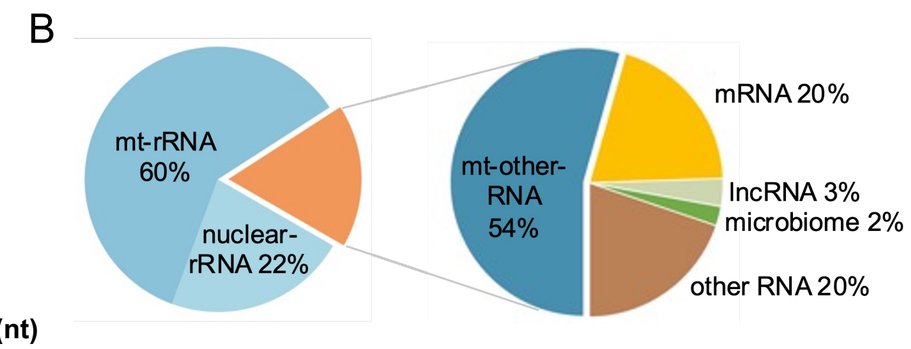
615 introns, and promoters. **(D)** Average percentages of clean reads (after trimming low-quality
616 and adapter sequences) assigned to different sources. **(E)** Numbers of raw sequencing
617 reads and human genome-aligned reads with a fixed budget of $300 for each method. **(F)**
618 Summary of key techniques used in the four cfRNA-seq approaches. Numbers of used
619 samples: Phospho-seq: 15; SILVER-seq: 128; SMARTer-seq: 373; DETECTOR-seq: 113.
620
621 **Figure 5 | Distinct human and microbial RNA signatures in *Plasma* versus *EV*.**
622 **(A)** Illustration of sequencing *Plasma* cfRNAs and *EV* cfRNAs in pairwise plasma samples.
623 **(B)** Distribution of reads mapped to human genome and microbiome in *Plasma* and *EV*
624 cfRNA datasets. Left: RNA spectrum mapping to human genome; Right: relative abundance
625 of reads aligned to different phyla. **(C)** Differential human RNA species between *Plasma* and
626 *EV* cfRNAs. **(D)** Pie charts show the average fractional contributions of various cell types to
627 the *Plasma* and *EV* transcriptomes. Box plots show the diversity of cell type contributions to
628 the *Plasma* and *EV* transcriptomes measured by the ratio of non-blood cells and Simpson's
629 index. **(E)** The fractions of reads aligned to microbe and virus. **(F)** Differential microbe genera
630 between *Plasma* and *EV* cfRNAs. *Plasma*: 44 samples; *EV*: 44 samples (all samples paired).
631 ****: *P*-value < 0.0001, **: *P*-value < 0.01, *: *P*-value < 0.05, Wilcoxon rank sum test, two-
632 tailed.
633
634 **Figure 6 | Distinct functional pathways, motifs, and binding proteins of the selective**
635 ***Plasma* and *EV* cfRNAs.**

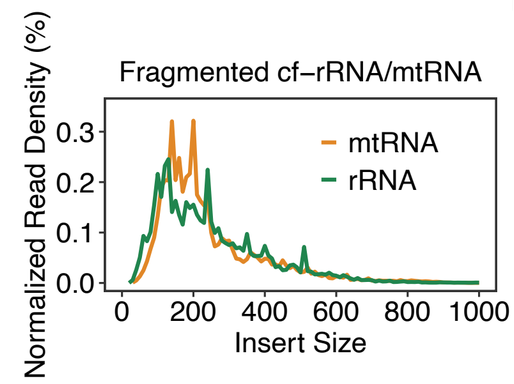
636 **(A)** Definition of the selective cfRNAs enriched in *Plasma* or *EV*. Cutoff: |Fold-change|>1 and
637 FDR<0.1. **(B)** Top enriched KEGG pathways of the selective cfRNAs. **(C)** Top enriched
638 motifs and their corresponding RNA binding proteins (RBPs) of the selective cfRNAs.
639 *Plasma*: 44 samples; *EV*: 44 samples (all samples paired).
640
641 **Figure 7 | Cancer classification using *Plasma* cfRNAs and *EV* cfRNAs.**
642 **(A)** Cancer-relevant ones (differentially expressed between cancer patients and normal
643 controls, $|\log_2$fold-change$|>1$ and FDR<0.05) in the selective and non-selective human
644 cfRNAs. Cancer: colorectal cancer (CRC) and lung cancer (LC); NC: normal control. **(B)**
645 Enriched GO terms related to cancer-relevant human cfRNAs. Performances (average of 20
646 bootstrap procedures) of cancer-relevant human cfRNAs distinguishing cancer patients from
647 normal controls when excluding **(C)** and including **(D)** non-selective cfRNAs. **(E)** AUROCs of
648 cancer type classification (CRC vs. LC) using human- or microbe-derived reads in *Plasma*
649 and *EV* cfRNAs. **(F)** Numbers of microbial features (genus) with significantly differential
650 abundance ($|\log_2$fold-change$|>1$ and FDR<0.1) between CRC and LC in 20 bootstrap
651 procedures. **(G)** Distinct cancer type-specific microbial features (genus) identified in *Plasma*
652 and *EV* cfRNAs. Heatmaps show z-scores of the abundance levels of these microbial RNA
653 features; bar plots illustrate their average $\log_2$FCs and FDRs between CRC and LC. FC: fold-
654 change; FDR: false discovery rate. ****: *P*-value < 0.0001, ***: *P*-value < 0.001, *: *P*-value <
655 0.05, Wilcoxon rank sum test, two-tailed. CRC samples: *Plasma* (n=23), *EV* (n=19), 19 of
656 them paired; LC samples: *Plasma* (n=19), *EV* (n=19), 18 of them paired; NC samples:

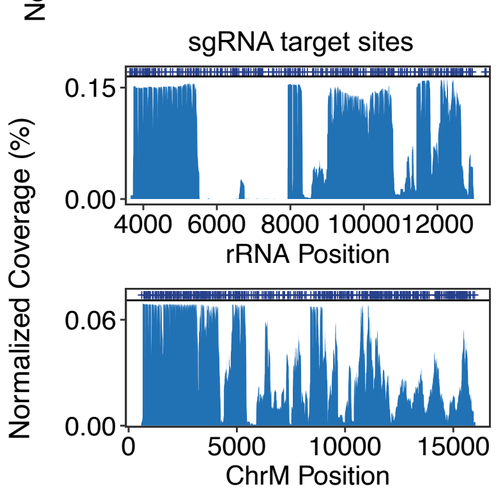
657 *Plasma* (n=19), *EV* (n=14), 7 of them paired.

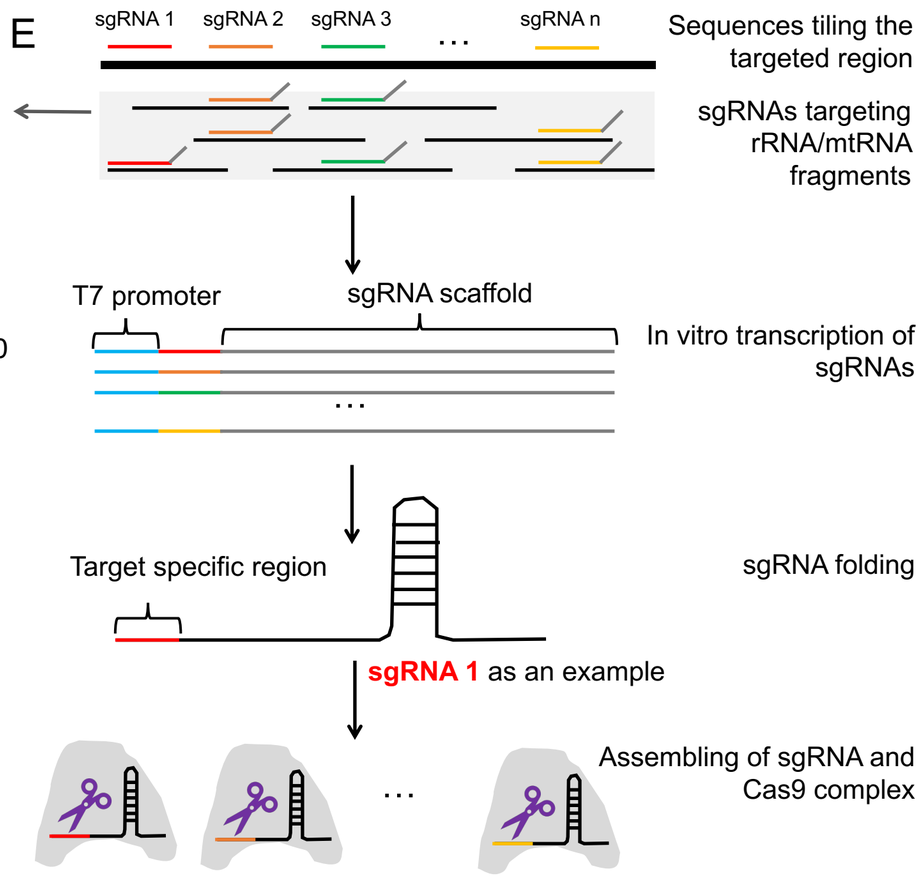
Figure 1


A
[FU]
30
25
20
15
10
5
0
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 (%)
0.3
0.2
0.1
0.0
mtRNA
rRNA
0
200
400
600
800
1000
Insert Size


D
sgRNA target sites
Normalized Coverage (%)
0.15
0.00
4000
6000
8000
10000
12000
rRNA Position
0.06
0.00
0
5000
10000
15000
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

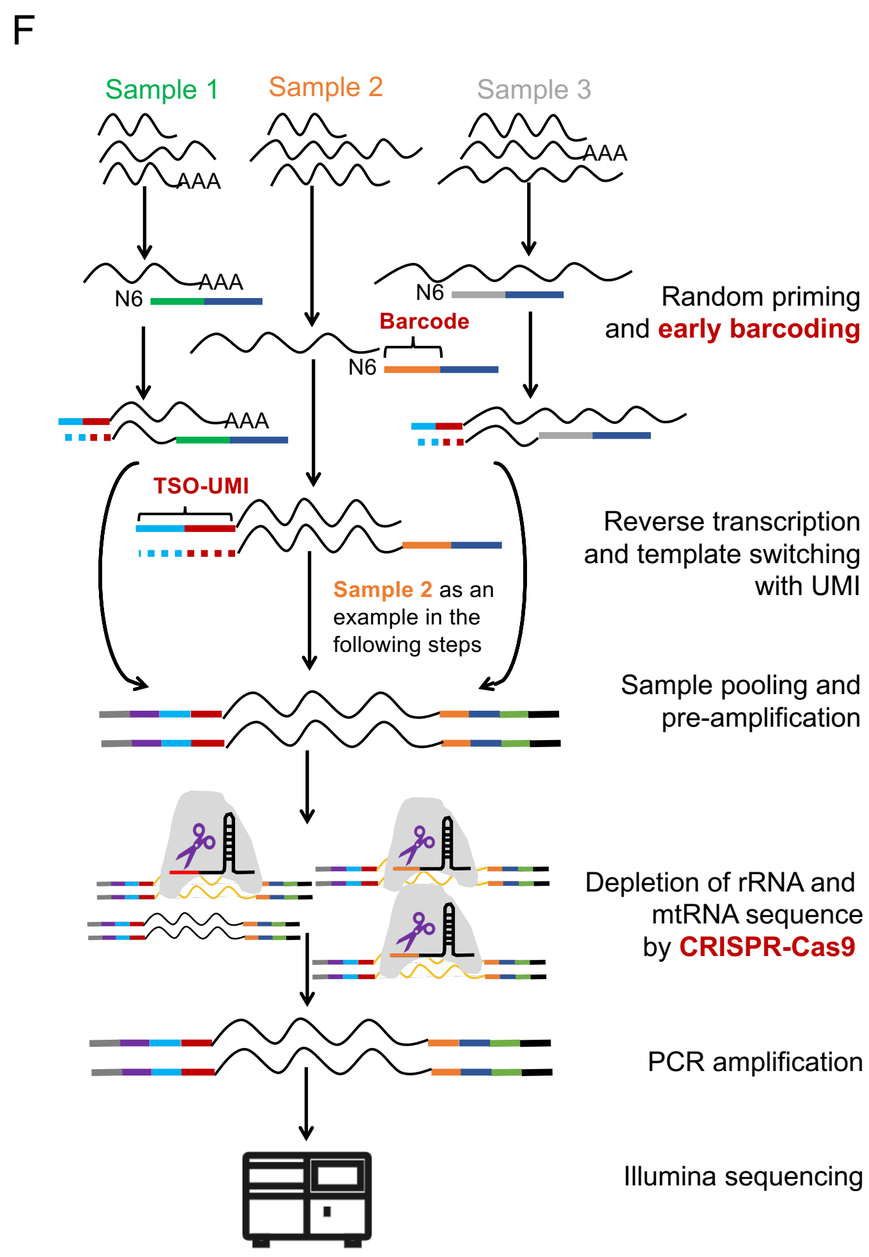

F
Sample 1
Sample 2
Sample 3
AAA
N6
Barcode
Random priming and early barcoding
TSO-UMI
Sample 2 as an example in the following steps
Reverse transcription and template switching with UMI
Sample pooling and pre-amplification
Depletion of rRNA and mtRNA sequence by CRISPR-Cas9
PCR amplification
Illumina sequencing

Figure 2

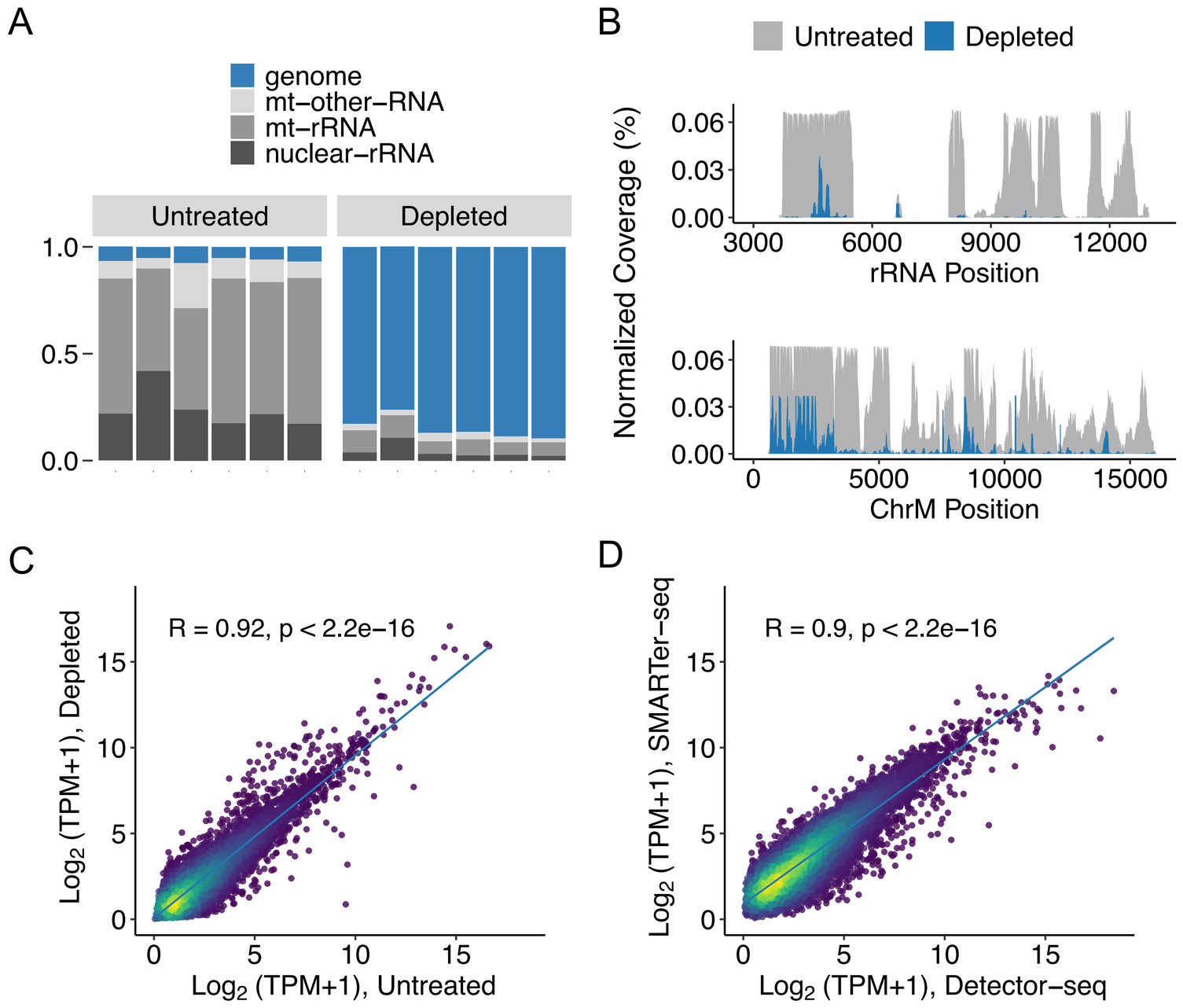
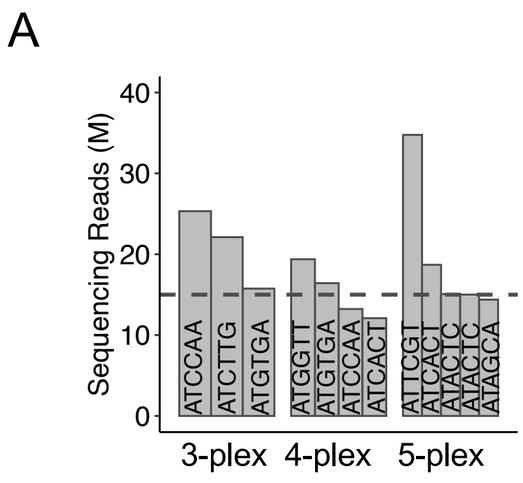
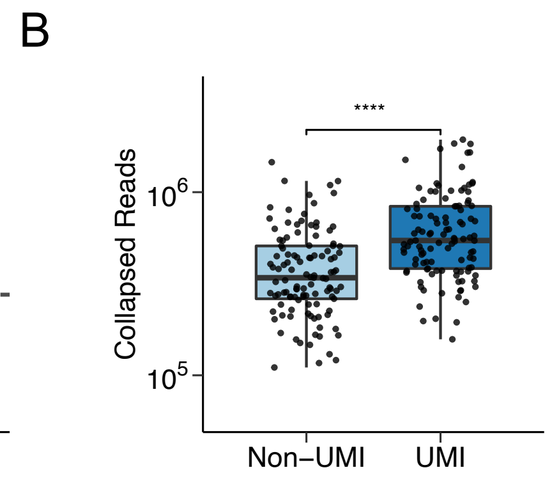
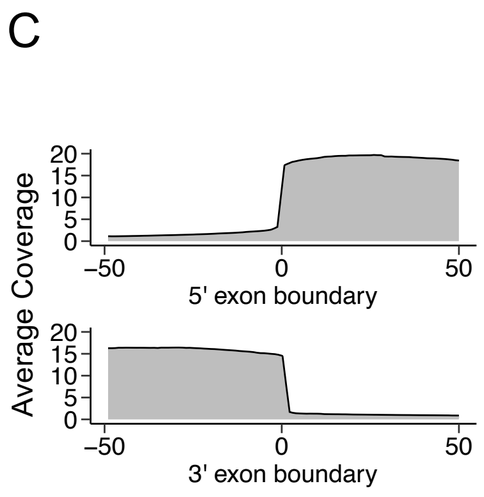
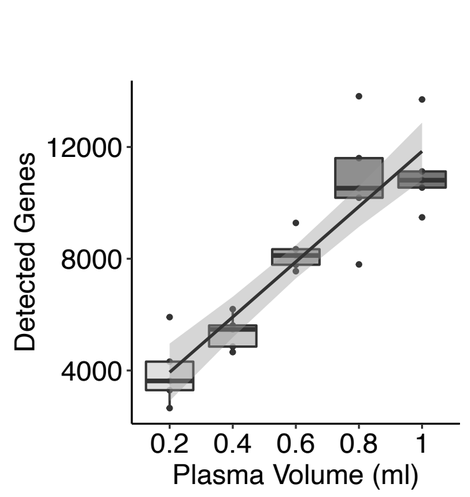

A
genome
mt–other–RNA
mt–rRNA
nuclear–rRNA
Untreated
Depleted
1.0
0.5
0.0
B
Untreated
Depleted
Normalized Coverage (%)
0.06
0.03
0.00
3000
6000
9000
12000
rRNA Position
0.06
0.03
0.00
0
5000
10000
15000
ChrM Position
C
R = 0.92, p < 2.2e–16
Log2 (TPM+1), Depleted
Log2 (TPM+1), Untreated
D
R = 0.9, p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

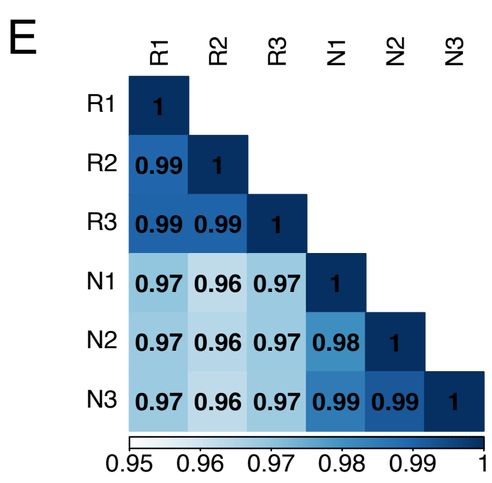
Figure 3

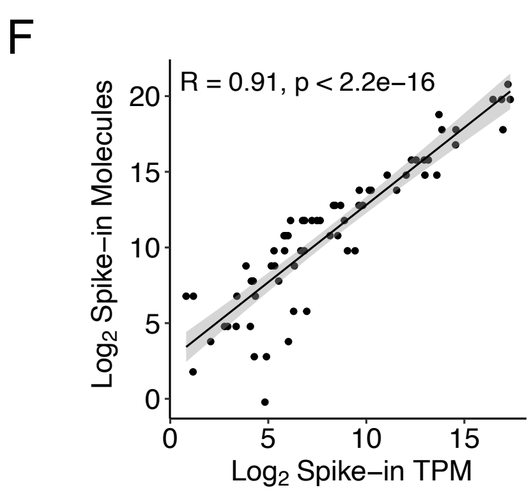

A
Sequencing Reads (M)
40
30
20
10
0
ATCCAA
ATCTTG
ATGTGA
ATGGTT
ATGTGA
ATCCAA
ATCACT
ATTCGT
ATCACT
ATACTC
ATACTC
ATAGCA
3-plex
4-plex
5-plex

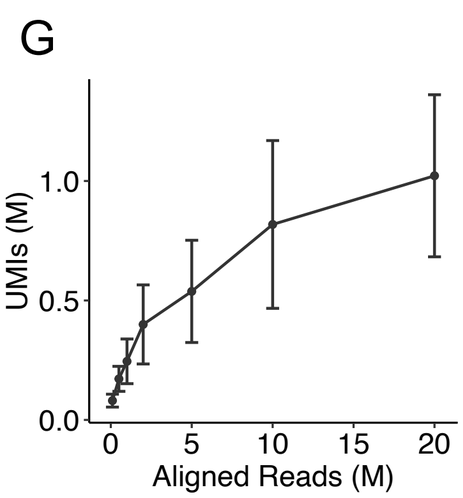

B
Collapsed Reads
10^6
10^5
****
Non−UMI
UMI


C
Average Coverage
20
15
10
5
0
−50
0
50
5' exon boundary
3' exon boundary


D
Detected Genes
12000
8000
4000
0.2
0.4
0.6
0.8
1
Plasma Volume (ml)


E
R1
R2
R3
N1
N2
N3
1
0.99
0.97
0.96
0.98
0.95


F
R = 0.91, p < 2.2e−16
Log_2 Spike−in Molecules
Log_2 Spike−in TPM


G
UMIs (M)
1.0
0.5
0.0
Aligned Reads (M)

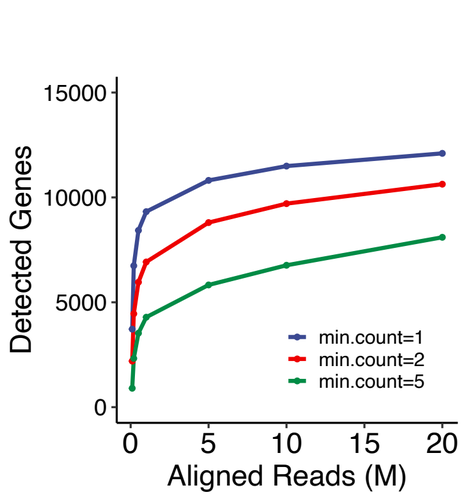
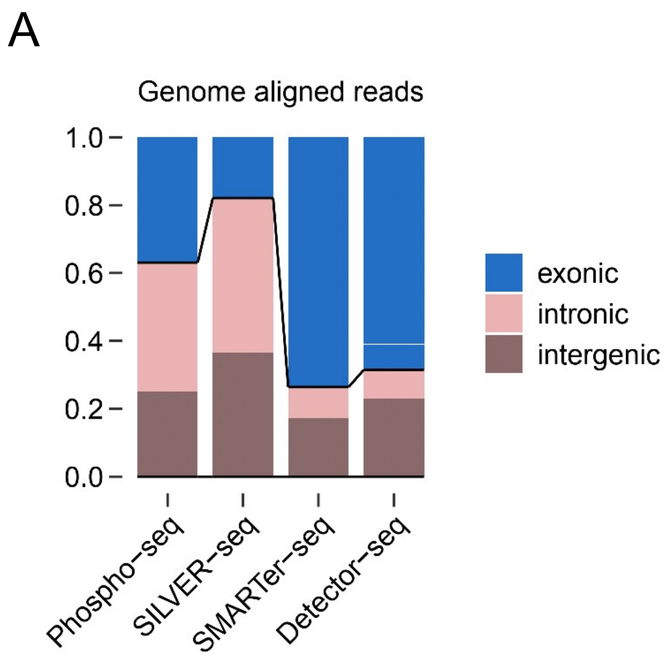
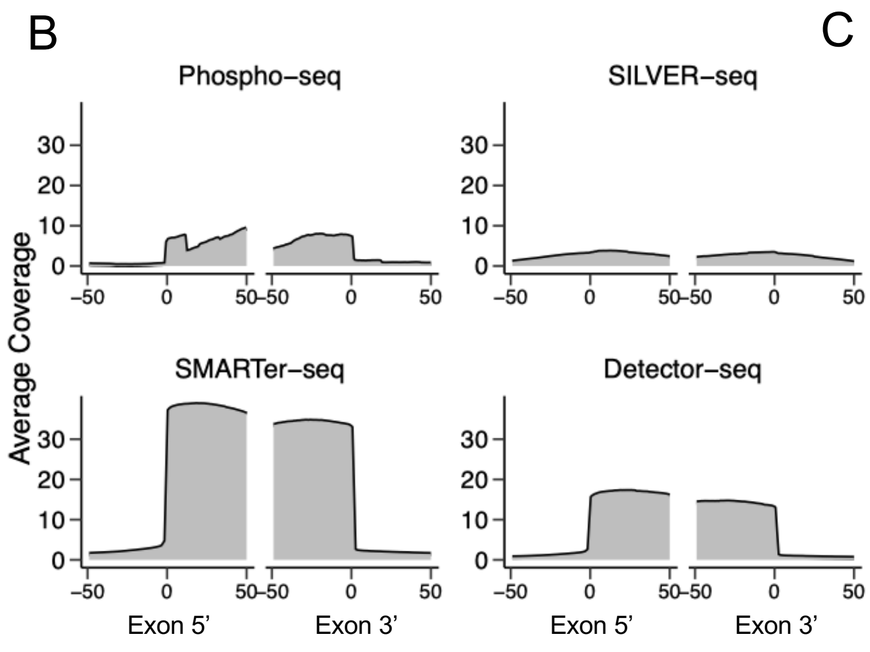
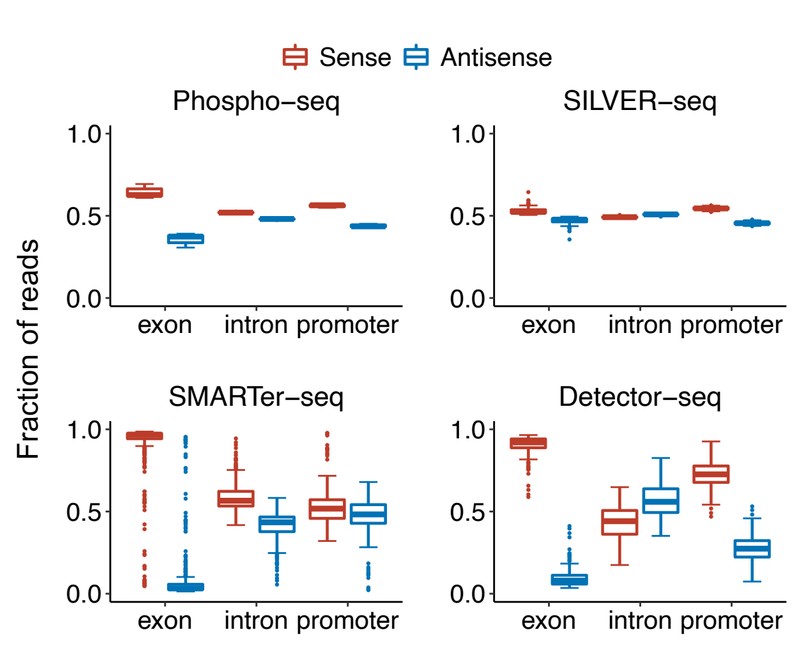
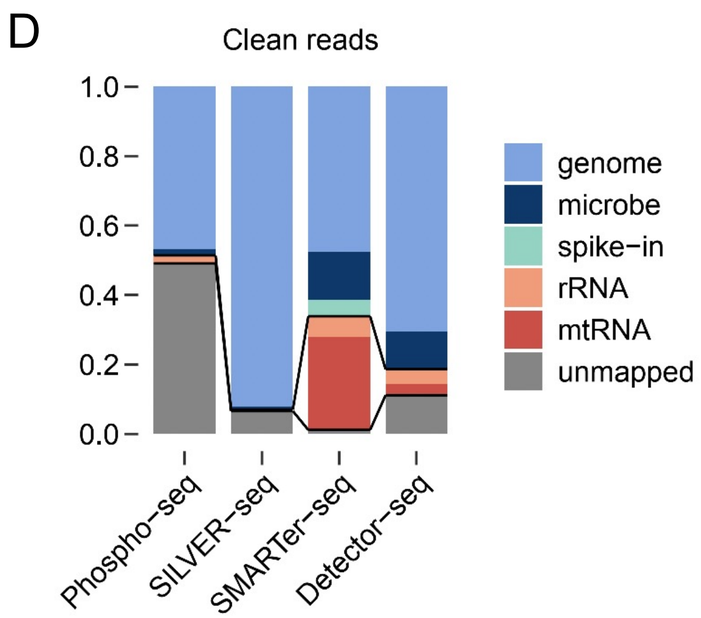

H
Detected Genes
15000
10000
5000
0
min.count=1
min.count=2
min.count=5
Aligned Reads (M)

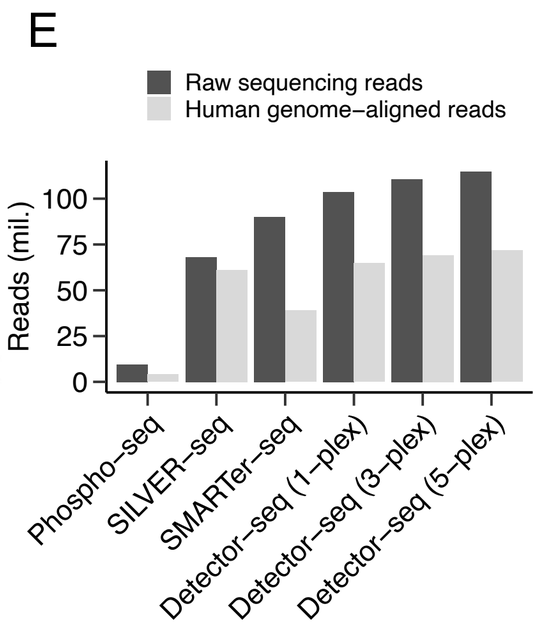
Figure 4







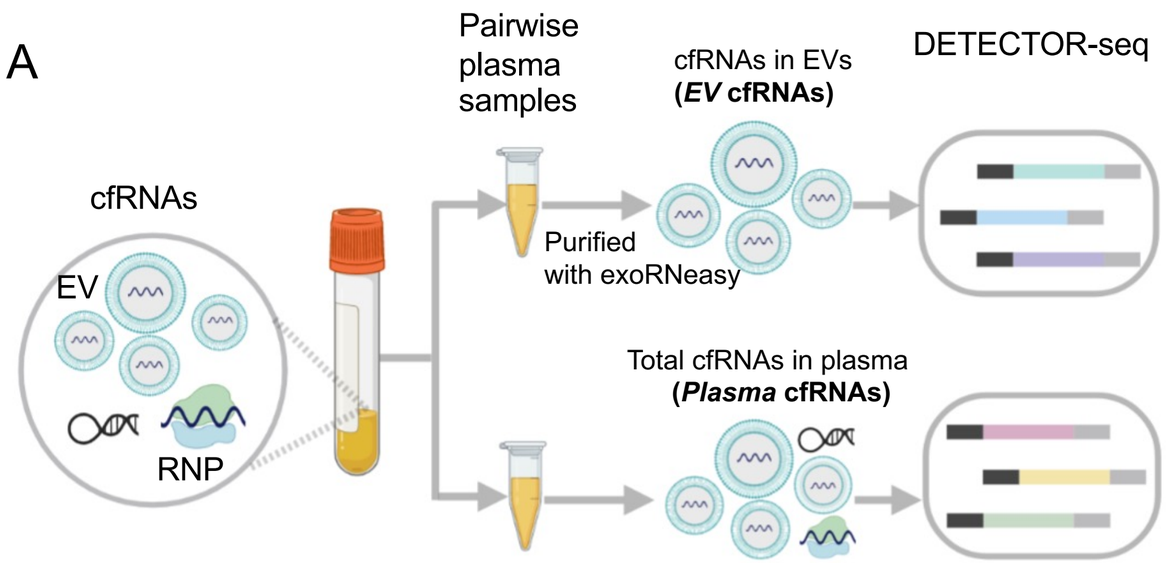


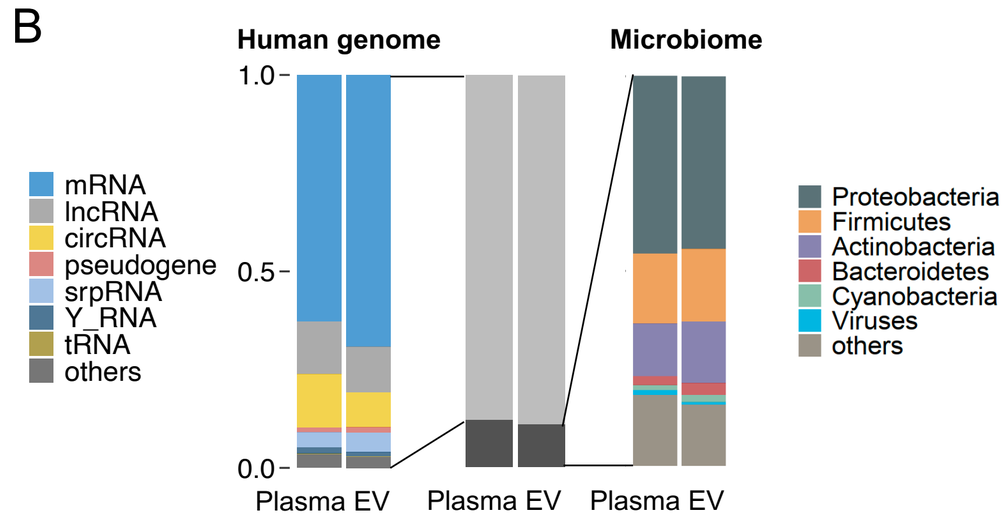


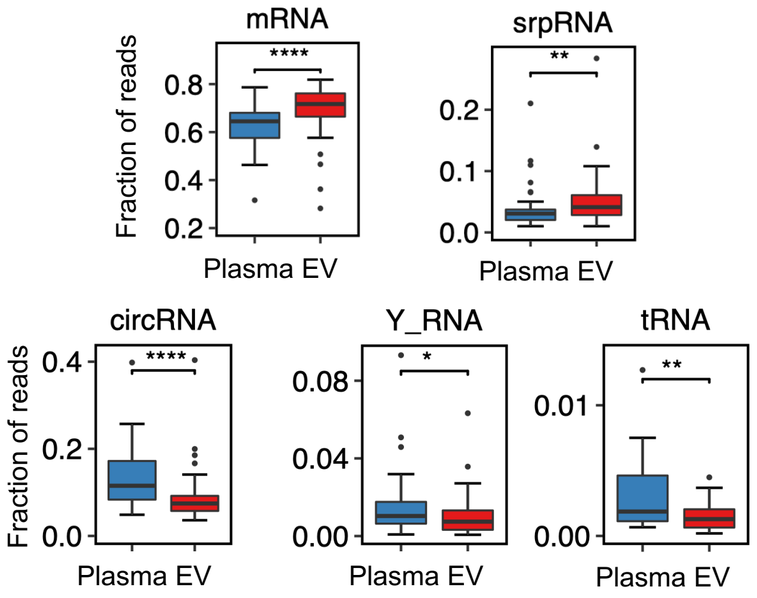
F

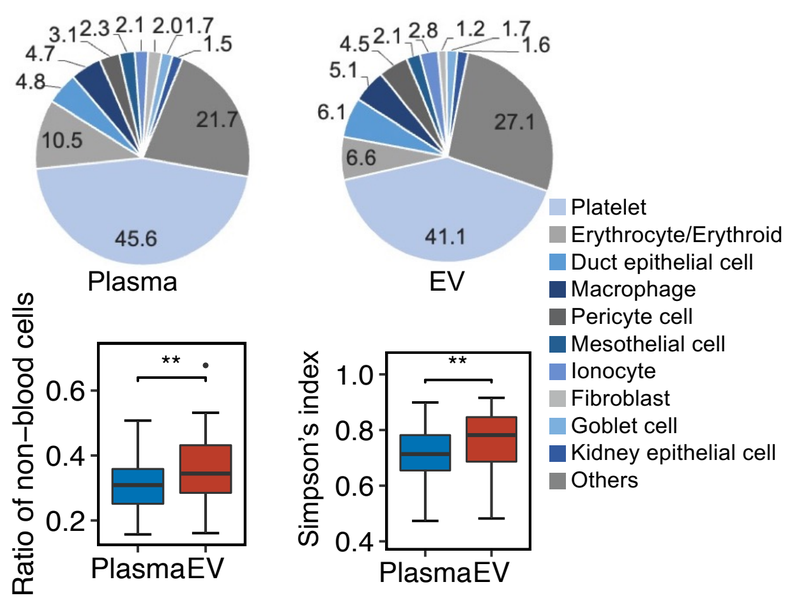
	Phospho-seq	**SILVER-seq**	**SMARTer-seq**	**Detector-seq**
Minimal input	0.2 ml	5 µl	1 ml	**0.2 ml**
UMI	✓	✓	×	✓
Early barcoding	×	×	×	✓
rRNA depletion	×	✓	✓	✓
mtRNA depletion	×	×	×	✓
DNA removal	×	×	✓	✓

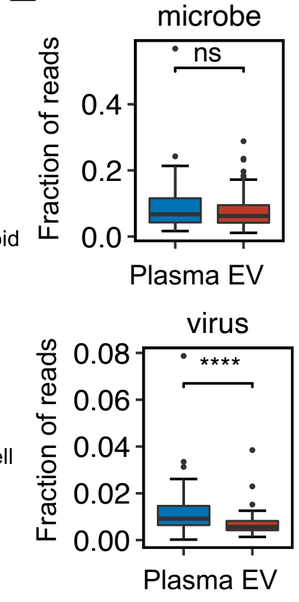
Figure 5


A
Pairwise plasma samples
cfRNAs in EVs (EV cfRNAs)
DETECTOR-seq
cfRNAs
EV
RNP
Purified with exoRNeasy
Total cfRNAs in plasma (Plasma cfRNAs)


B
Human genome
Microbiome
mRNA
lncRNA
circRNA
pseudogene
srpRNA
Y_RNA
tRNA
others
Proteobacteria
Firmicutes
Actinobacteria
Bacteroidetes
Cyanobacteria
Viruses
others
Plasma EV
Plasma EV
Plasma EV


C
mRNA
srpRNA
circRNA
Y_RNA
tRNA
Fraction of reads
Plasma EV


D
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


E
microbe
virus
Fraction of reads
Plasma EV

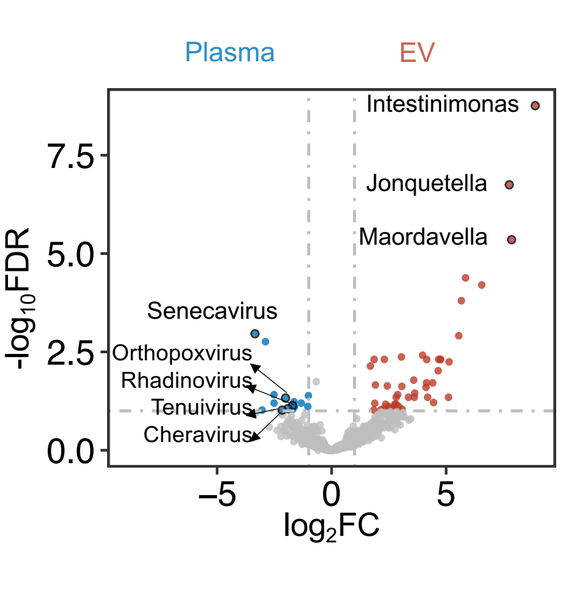
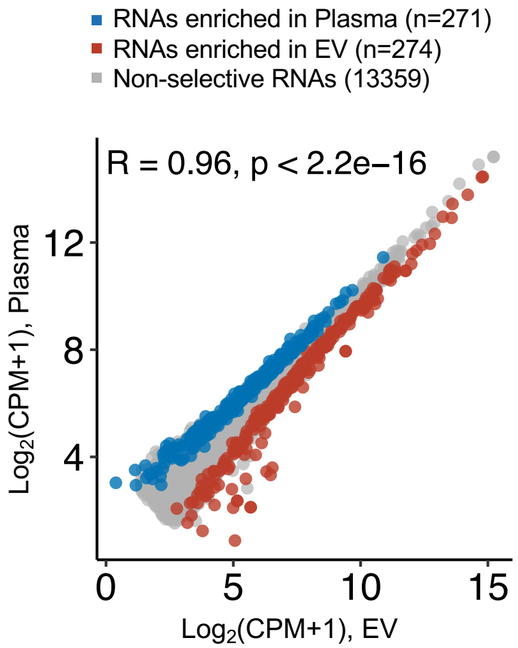
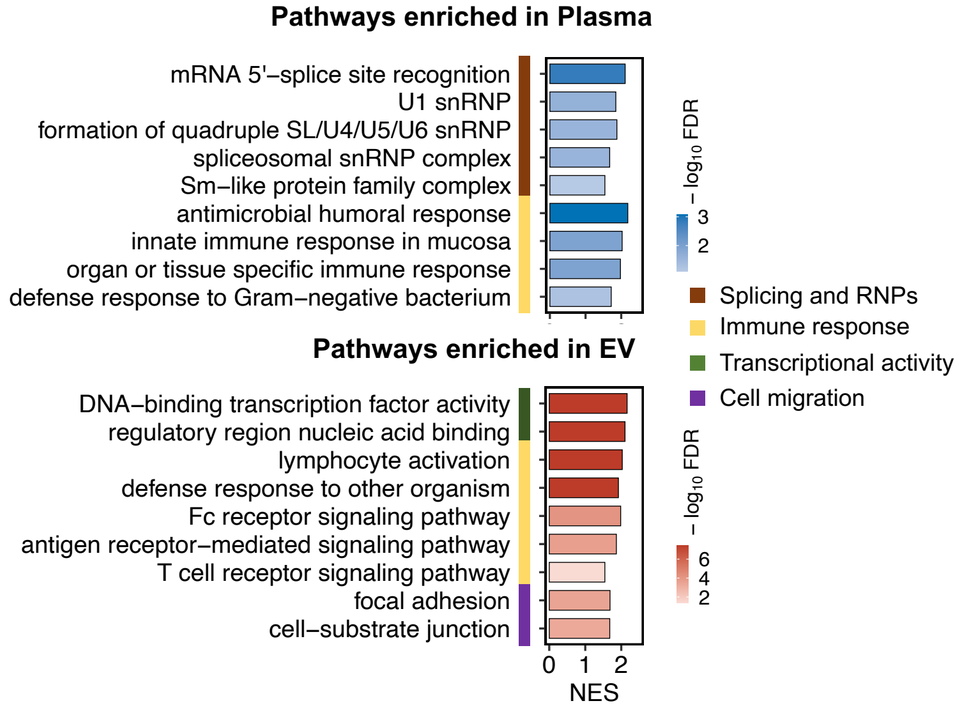

F
Plasma
EV
Intestinimonas
Jonquetella
Maordavella
Senecavirus
Orthopoxvirus
Rhadinovirus
Tenuivirus
Cheravirus
-log10FDR
log2FC

Figure 6

A



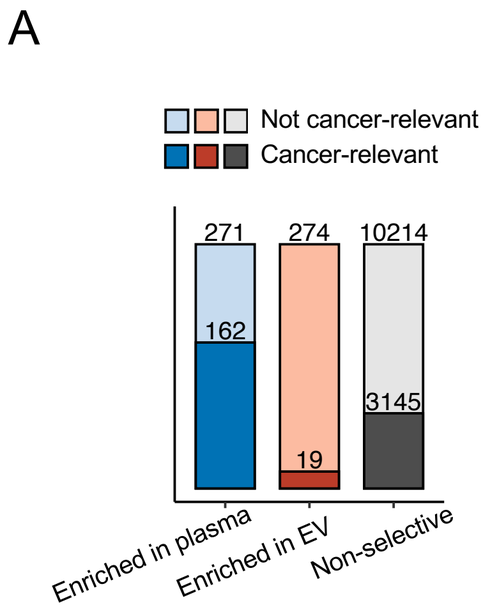
B

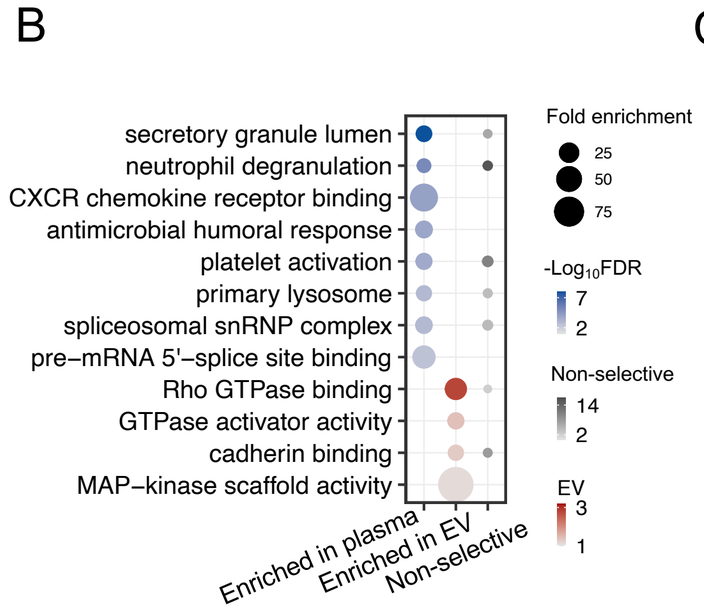


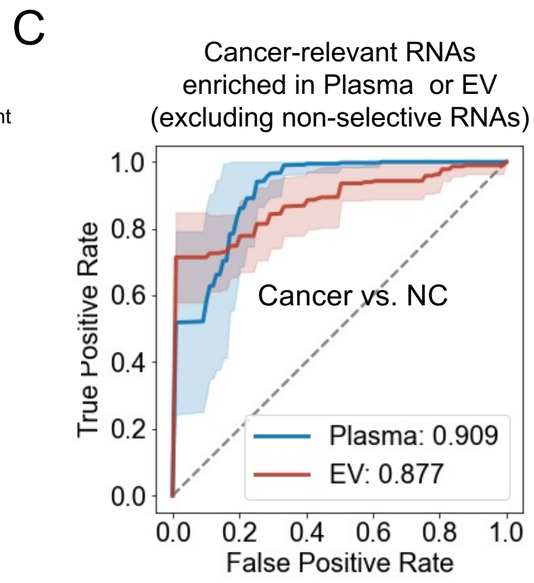
C

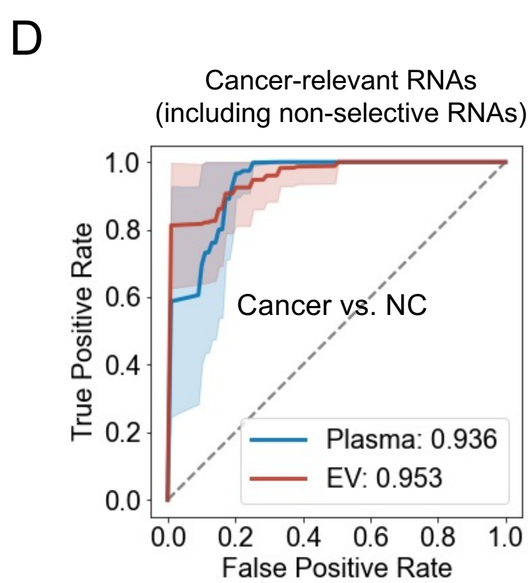
	Motif	E-value	RBP	Annotation
Enriched in Plasma		1.64E-35	ABCF1	innate immune
		1.46E-29	SFPQ	splicing factor
		2.31E-26	LARP4	La RNP
		4.51E-23	TROVE2	Y RNA binding protein
		1.77E-20	DKC1	snoRNP
Enriched in EV		2.57E-61	PUM1	innate immune
		4.7E-57	BCLAF1	transcription repressor
		2.4E-55	HNRNPU	transcription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

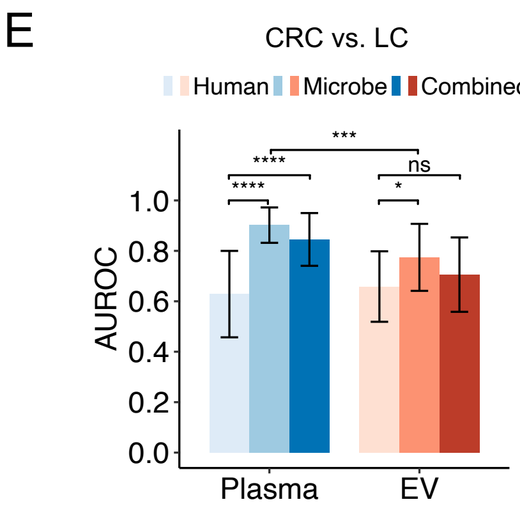
Figure 7

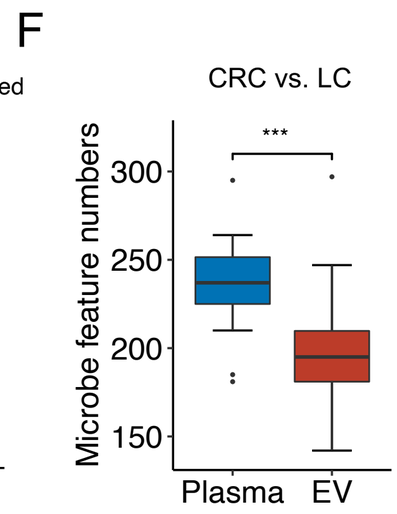

A
Not cancer-relevant
Cancer-relevant
271
274
10214
162
19
3145
Enriched in plasma
Enriched in EV
Non-selective


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
7
2
Non-selective
14
2
EV
3
1


C
Cancer-relevant RNAs
enriched in Plasma or EV
(excluding non-selective RNAs)
True Positive Rate
False Positive Rate
Cancer vs. NC
Plasma: 0.909
EV: 0.877


D
Cancer-relevant RNAs
(including non-selective RNAs)
True Positive Rate
False Positive Rate
Cancer vs. NC
Plasma: 0.936
EV: 0.953


E
CRC vs. LC
Human
Microbe
Combined
AUROC
Plasma
EV


F
CRC vs. LC
Microbe feature numbers
Plasma
EV

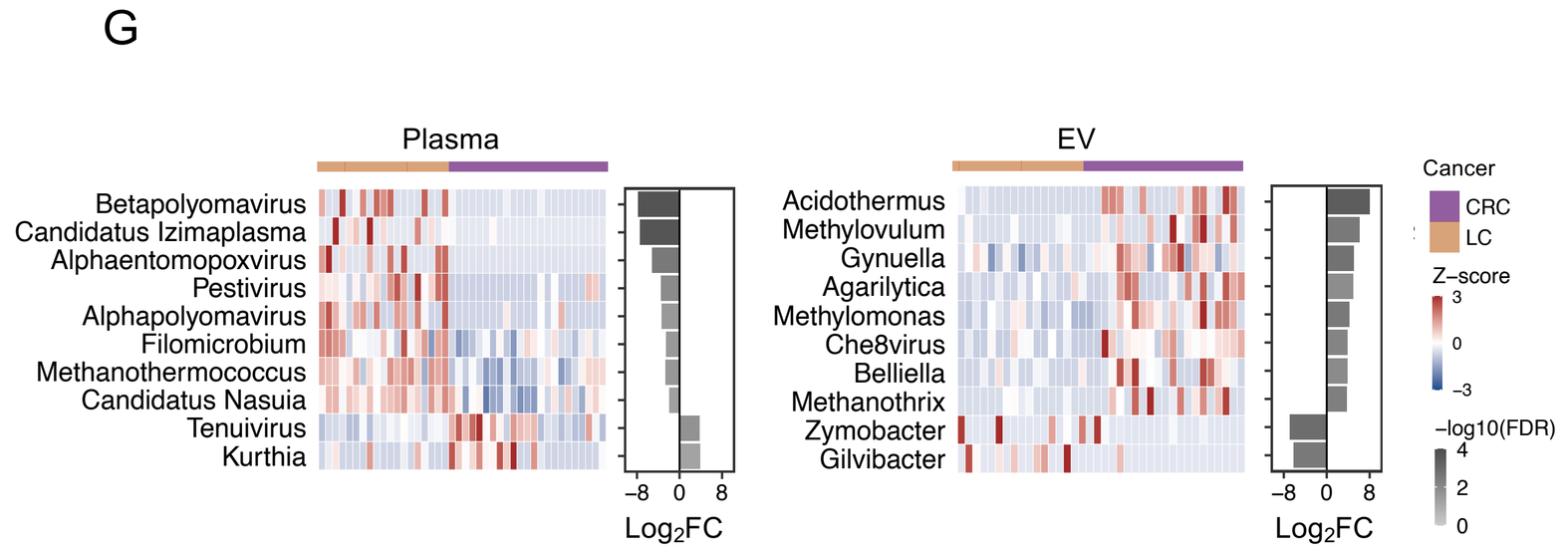

G
Plasma
Betapolyomavirus
Candidatus Izimaplasma
Alphaentomopoxvirus
Pestivirus
Alphapolyomavirus
Filomicrobium
Methanothermococcus
Candidatus Nasuia
Tenuivirus
Kurthia
Log2FC
EV
Acidothermus
Methylovulum
Gynuella
Agarilytica
Methylomonas
Che8virus
Belliella
Methanothrix
Zymobacter
Gilvibacter
Log2FC
Cancer
CRC
LC
Z–score
–log10(FDR)

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

	Total cfRNA-seq (*Plasma* cfRNA-seq)	***EV* cfRNA-seq**
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
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.