Genome Biology

RESEARCH

Open Access

Peak analysis of cell-free RNA finds recurrently protected narrow regions with clinical potential

Pengfei Bao[1,5,6†], Taiwei Wang[1,2,3,7†], Xiaofan Liu[1,5†], Shaozhen Xing[1,5], Hanjin Ruan[4], Hongli Ma[1], Yuhuan Tao[1,5], Qing Zhan[1,5], Efres Belmonte-Reche[8,9,10], Lizheng Qin[4], Zhengxue Han[4], Minghui Mao[4*], Mengtao Li[2,3*] and Zhi John Lu[1,5,6,7,11*]

†Pengfei Bao, Taiwei Wang and Xiaofan Liu contributed equally to this work.

*Correspondence: mmh_hover@163.com; mengtao.li@cstar.org.cn; zhilu@tsinghua.edu.cn

[1] MOE Key Laboratory of Bioinformatics, State Key Lab of Green Biomanufacturing, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China
[2] Department of Rheumatology and Clinical Immunology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China
[4] Department of Oral and Maxillofacial & Head and Neck Oncology, Beijing Stomatological Hospital, Capital Medical University, Beijing, China
Full list of author information is available at the end of the article

Abstract

Background: Cell-free RNAs (cfRNAs) can be detected in biofluids and have emerged as valuable disease biomarkers. Accurate identification of the fragmented cfRNA signals, especially those originating from pathological cells, is crucial for understanding their biological functions and clinical value. However, many challenges still need to be addressed for their application, including developing specific analysis methods and translating cfRNA fragments with biological support into clinical applications.

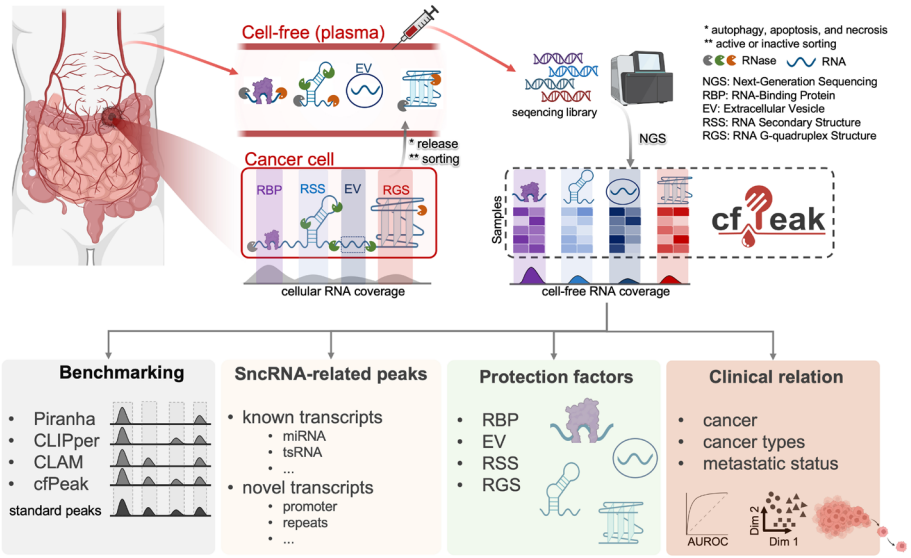
Results: We present cfPeak, a novel method combining statistics and machine learning models to detect the fragmented cfRNA signals effectively. When test in real and artificial cfRNA sequencing (cfRNA-seq) data, cfPeak shows an improved performance compared with other applicable methods. We reveal that narrow cfRNA peaks preferentially overlap with protein binding sites, vesicle-sorting sites, structural sites, and novel small non-coding RNAs (sncRNAs). When applied in clinical cohorts, cfPeak identified cfRNA peaks in patients' plasma that enable cancer detection and are informative of cancer types and metastasis.

Conclusions: Our study fills the gap in the current small cfRNA-seq analysis at fragment-scale and builds a bridge to the scientific discovery in cfRNA fragmentomics. We demonstrate the significance of finding low abundant tissue-derived signals in small cfRNA and prove the feasibility for application in liquid biopsy.

Keywords: Cell-free RNA, CfRNA fragment, Peak calling, SncRNA, Liquid biopsy biomarker, Tissue-of-origin, Noninvasive cancer detection

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Graphical Abstract



Background

Cell-free RNAs (cfRNAs), also known as extracellular RNAs (exRNAs), refer to RNA molecules that are released or sorted from various cell types due to natural or pathological events like active EV secretion [1], autophagy, apoptosis, and necrosis [2]; they mixed together and can circulate in biofluids. CfRNA molecules are usually degraded into fragments due to the digestion of various types of RNase [3, 4]. However, some short or even long fragments can survive being fully degraded due to different protection factors like protein-binding, formation of local RNA secondary structure (RSS), and extracellular vesicles (EV) [3, 5].

Recently, noncanonical small RNA species have been found in the cellular environment, including tRNA-derived-small fragments (tsRNAs, tRFs) [6–9], ribosomal-derived-small RNA fragments (rsRNAs) [6, 8, 10], mRNA-derived-small fragments (msRNA) [8, 11, 12], lncRNA-derived-small fragments (lncsRNA) [8, 12], and other unannotated small RNA species [5, 13], many of which have been shown to play important roles in cellular regulation and disease pathogenesis and have specific biogenesis pathways that closely related to RNase degradation [8, 9, 11]. Some tools and databases have been developed to study these classes of sncRNAs in (small) RNA-seq data [14, 15]. As most cfRNA fragments are derived from multiple tissue-of-origin (TOO), far more than well-known cell-free miRNA [16–18], diverse sncRNA species are also present in cell-free biofluids and can be detected in cfRNA-seq data [11–13], and some have been proposed as promising biomarkers for a wide range of diseases, including cancer, cardiovascular, and neurological disorders [17, 19–23], which was also stated in our previous work [24]. Two available cfRNA/exRNA sequencing data analysis pipelines utilized a similar strategy; they map clean reads to transcriptome and calculate read counts for each annotated transcript [25, 26]. However, this cannot fully explore the locally

fragmented profile of cfRNA; read counts from full-length transcripts (tx) could not always represent the cfRNA's abundance and biological significance at the local fragment-scale. On the other hand, many pathological cell-derived cfRNA fragments of low abundance or in unannotated regions may be discarded or overlooked, some of which have been found with diagnostic potential [13, 27, 28]. In summary, the lack of proper tools to identify and quantify fragmented cfRNA signals limits scientific discovery and clinical application.

In addition to the traditional analysis methods, several cfRNA studies tried to find local read clusters (peaks) in long- or short-range using peak calling methods, which are usually implemented in cross-linking immunoprecipitation sequencing (CLIP-seq) or RNA immunoprecipitation sequencing (RIP-seq) data analysis. Yao et al. [5] explored total cfRNA profile in healthy donors' plasma by MACS2 using their well-developed library workflow and found a series of interesting results, including the report of many protein-binding and structured cfRNA peaks in plasma. However, the study was not extended to broad validation and clinical application. Another drawback is that MACS2 was mainly designed for genomic peaks like chromatin immunoprecipitation sequencing (ChIP-seq) peaks, which might be unsuitable for cfRNA. Felden et al. [13] identified cfRNA/exRNA peaks in unannotated regions from patients' plasma, validated their presence in EV, and tested their diagnostic performance in large cancer cohorts, which was helpful for small cfRNA's clinical applications. This work has two limitations. First, it only focused on peaks in unannotated regions, and the vast number of peaks in annotated regions was not considered. Second, the peak identification procedures lacked background correction and filtering steps, which might lead to redundant and noisy peaks of low confidence. As far as we know, no systematic peak analysis method evaluated for cfRNA considering their specific properties has been reported, and the cfRNA peaks in biofluids for clinical applications need further biological explanation and in-depth evaluation.

Here we present a fragment-scale peak calling method, cfPeak, which considers cfRNA's characteristics. When applied in real and artificial cfRNA datasets, cfPeak was more sensitive than other methods in the identification of true peaks as well as applicable to multiple data types. When applied to clinical cohorts, cfPeak identified narrow peaks that could detect cancer, identify cancer types, and predict metastatic status. Furthermore, we demonstrated the importance of rescuing low abundant cfRNA peaks with potential cancer tissue-origin in liquid biopsy.

Results

Narrow peaks can be recurrently detected in small cfRNA-seq data

We compared the read coverage profiles of cellular CLIP-seq (CL-CLIP-seq), cellular small RNA-seq (CL-smRNA-seq), and cell-free small RNA-seq (CF-smRNA-seq). Clean reads were mapped to known transcripts and then called peaks traditionally using CLIPper [29]. Four example regions in transcripts (peak precursors) of known cfRNA species were shown. Generally, for non-captured or non-immunoprecipitated sequencing data like CL- and CF-small RNA-seq, peaks also exist in different transcript species across multiple library strategies, which is consistent with previous studies [10, 30]. Moreover, we observed some notable differences. Peaks in CL- and

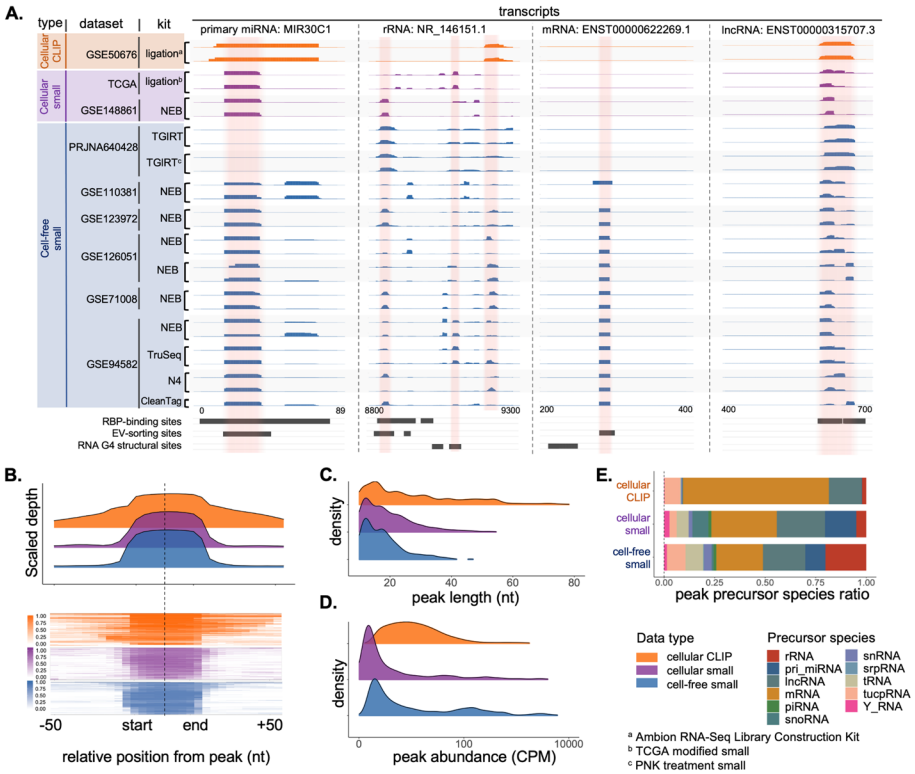


Fig. 1 Narrow peaks can be recurrently detected in small cfRNA-seq data. **A** Four example regions with recurrent peaks. X axis and Y axis of each region stand for transcriptomic coordinate relative to full-length transcript and min–max scaled read depth. Two random samples/tracks were plotted for each dataset. **B**–**E** Comparison of representative datasets (GSE50676, GSE148861, and GSE71008) from three different data types. **B** Ridge plot showed read clusters' distribution around peak regions. X axis stands for extended 50 nt from scaled peak regions. Top: Y axis means scaled read depth. Bottom: Y axis of heatmap means regions ranked by mean of scaled depth each data type, color intensity means scaled count signal. **C** Ridge plot showed peak length distribution. **D** Ridge plot showed peak abundance distribution. **E** Stacked bar plot showed ratio of different peak precursor (transcript) species. NEB, NEBnext small RNA-seq Kit; TGIRT, thermostable group II intron reverse transcriptases

CF-smRNA-seq have relatively sharper peak ranges and more diverse peak locations than CL-CLIP-seq. Some peaks in CF-smRNA-seq were recurrently detected and overlapped with annotated regions, which implied potential biological significance. For example, peaks in the rRNA example overlapped with RBP-binding sites (RBPBS), EV-sorting sites, and RNA G-quadruplexes structural (RGS) sites. We also identified a unique peak across CF-smRNA-seq datasets in mRNA precursor that overlapped with EV-sorting sites, which could be explained by the cell-free-specific EV-sorting mechanism (Fig. 1A).

To further evaluate the difference among the three data types statistically, we focused on one representative dataset for each data type with a similar library strategy. The meta coverage plot in the extended peak regions showed CL-CLIP-seq libraries have a relatively broader reads clustering pattern. Peak length distribution demonstrated a slightly smaller peak size in CF-smRNA-seq. We also found that CF-smRNA-seq has more diverse peak precursor species (Fig. 1B–E). In conclusion, recurrent peaks with potential biological significance can be detected across different

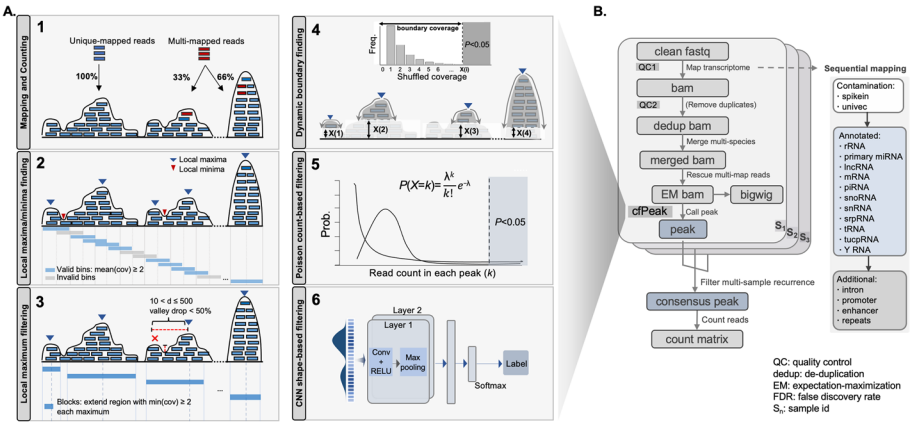


Fig. 2 Overview of cfPeak and analysis pipeline for fragmented cfRNA. **A** Peak calling modules of cfPeak. First, multi-mapped reads rescued by EM are extracted and converted to 1D vector of depth for each transcript. Second, full-length transcript is divided into half-overlapped bins and valid bins are merged; positions of local maxima and minima are located for each merged bin. Third, redundant maxima are removed iteratively and block region is located for each reduced maximum. Fourth, peak boundaries are located by shuffling observed reads to background region for each local maximum. Fifth, Poisson test to remove insignificant peaks supported by fewer reads. Sixth, optional CNN-assisted filtering based on peak shape in a fixed local window. **B** Whole pipeline of peak analysis for fragmented cfRNA

CF-smRNA-seq samples and datasets, many of which tend to be in narrow sizes and low abundance while with more positional diversities.

Overview of cfPeak and analysis pipeline for fragmented cfRNA

Considering the differences we observed, one may face the risk of improper modeling or inadequate parameters when exploring small cfRNA-seq with traditional peak callers developed for CLIP- and RIP-seq. Thus, we introduce a peak calling method, cfPeak, to fill this gap. CfPeak is organized into six steps with multiple advantages over other methods (Fig. 2A, Additional file 1: Fig. S1B–F) and is wrapped into a computational pipeline (Fig. 2B). In short, clean reads are sequentially mapped to contamination sequences, known annotated RNA transcripts, and optionally other regions that have been shown to host potential novel transcripts [31–34]. Multi-mapped reads have been reported to occupy a considerable part of available reads; thus, EM-based reads' reassignment is deployed before cfPeak since it can significantly improve the identification of peaks that are located in repetitive regions [35] (Additional file 1: Fig. S1A). After peak calling of each sample by cfPeak, recurrent peaks among samples (consensus peaks) can be obtained, and a count matrix is reported by counting consensus peaks for each sample and compiling all samples (Additional file 1: Fig. S2). More details were described in Additional file 1: "Supplementary Methods" section.

CfPeak sensitively finds peaks in the public and artificial datasets

We next systematically compared the performance of cfPeak with other methods in identifying cfRNA peaks in known transcripts. An in silico pooled cell-free sample from healthy donors in the reference plasma small RNA-seq dataset (GSE71008) was generated to avoid heterogeneity. Basic statistics of peaks in this sample were summarized, including peak number, peak abundance, and peak length (Fig. 3A, Additional

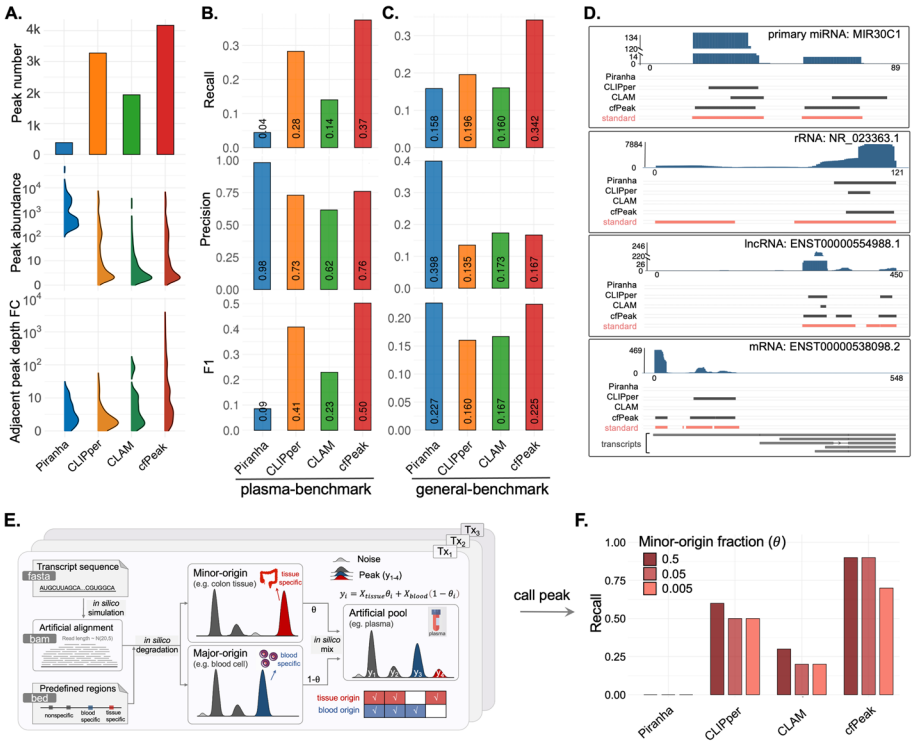


Fig. 3 CfPeak sensitively finds peaks in the public and artificial datasets. **A** Peak number, abundance (CPM), and fold change of adjacent peak depth of four methods (consensus peak). **B** Recall, precision, and F1 score of four methods using plasma-benchmark. **C** Recall, precision, and F1 score of four methods using general-benchmark. **D** Example regions with peaks were variably detected by four methods. Y axis: min–max scaled depth. X axis: transcriptomic coordinates in full-length transcript. Alternative transcripts were plotted below if exist. **E** Illustration of artificial data generation. **F** Recall of four methods in artificial plasma samples with different minor-origin fraction. Tx, transcripts; FC, fold change

file 1: Fig. S3A). CfPeak identified more peaks than other methods, and all their abundance was at a similar scale, except Piranha which was higher. For adjacent peaks with varying differences in peak depth or abundance, the smaller ones that tended to be discarded by other methods were rescued by cfPeak (Fig. 3A). We compared the overlaps of peak regions with each other; cfPeak identified more unique peaks ($n = 1027$) than CLIPper ($n = 301$), CLAM ($n = 124$), and Piranha ($n = 0$). CfPeak, CLIPper, and CLAM seemed to have the highest number of overlapped peaks, which implied the consensus results of different methods (Additional file 1: Fig. S3C). Moreover, peaks from AGO CLIP-seq typically host miRNA seed motifs [34, 35], and therefore we hypothesized a similar enrichment might exist since recent studies have revealed the presence of AGO2 protein in cell-free plasma [36, 37]. As expected, we obtained a similar pattern in AGO2-binding-sites-overlapped peaks, and they enriched more seed motifs from top abundant miRNAs in plasma (Additional file 1: Fig. S3D).

To better compare the performance of different methods, we defined a gold-standard true peak set (noisy truth) by filtering peaks above the predefined recurrence of samples and datasets using available public CF-smRNA-seq data. This strategy is referred to as plasma-benchmark and is applicable to plasma small RNA-seq datasets. We also developed another complementary general-benchmark based on the theory

and rationale of unique-mapped reads tend to map to multiple false sites after being shortened [35, 38], which is applicable to all datasets (Additional file 1: Fig. S4). In such a way, we found cfPeak recalled more true peaks (Fig. 3B, C), some of which were low abundant and less detectable for other methods (Fig. 3D). We also evaluated the precision, F1 score, and false positive rate (FPR) of the four methods. CfPeak presented the higher F1 score in plasma-benchmark despite rescuing many low abundant peaks (Fig. 3B), and also showed higher FPR and highest recall in general-benchmark (Additional file 1: Fig. S5B).

To better evaluate the performance and limit-of-detection of different methods in an ideal situation, we simplified the multi-origin model of plasma small cfRNA as consisting of one major-origin contribution (e.g., blood-derived) and one minor-origin contribution (e.g., colon-derived), and artificially generated such a dataset (Fig. 3E). Details were described in Additional file 1: "Supplementary Methods" section. We further tested cfPeak in this artificial dataset and observed similar results; minor-origin-derived low abundant peaks in RNA admixture could be detected sensitively by cfPeak (90%, 90%, and 70% recall at 50%, 5%, and 0.5% minor-origin fraction, respectively), many of which were discarded by traditional methods (Fig. 3F).

Various library strategies have been used in generating CF-smRNA-seq datasets; thus, we sought to test the performance of cfPeak beyond the reference dataset. CfPeak acquired similarly competent performance in different datasets using two benchmark metrics, implying its generality and robustness (Additional file 1: Fig. S5A,B). Also, we found that cfPeak was also potentially applicable to CLIP-seq dataset; additional peaks of interest were identified from a previous study [39] (Additional file 1: Fig. S5C). Together, these observations suggested cfPeak is a promising and generalized tool for peak analysis that highlights fragmented and low abundant peak regions with heterogeneous cellular origins for multiple bulk sequencing data types in broad biofluids.

CfPeak is capable of identifying cfRNA peaks in known and novel transcripts

We next investigated the extensive profile of cfRNA peaks in known transcripts using cfPeak. First, focused on mature miRNA (peak) with known boundaries, we found the proportion of boundaries that were precisely determined by cfPeak (GSE71008 average: 49.9%) was higher than other methods (Additional file 1: Fig. S3B). While most CF-smRNA-seq studies focused on canonical small RNA species like miRNA, many reads that cannot be mapped to miRNA transcripts (GSE71008 pool: 74.2%) were discarded. Therefore, we next tested whether peaks in previously known cell-free sncRNA species other than miRNA can be appropriately located using cfPeak. Previous studies have reported tsRNA as a prominent sncRNA species in cell-free environments [30, 40, 41]. We thus checked tRNA-mapped reads and tRNA-hosted peaks in the reference dataset. Grouping reads by tRNA amino acid classes, we found lengths of most reads spanned between 15 and 40 nt, which were slightly longer than the average lengths of all available reads. We also summarized read numbers by known tsRNA classes [42] and found tRNA transcripts tended to have different preferences in fragmentation position. The read number was also distributed in a biased way; Glu and Gly tRNA occupied the majority (GSE71008 pool: 80.9%) of tRNA-mapped reads (Gly 5p-tR-halves: 36.1%,

Glu 5p-tR-halves: 28.1%, Glu misc-tRFs: 7.0%), which has been reported to form stable homodimers or heterodimers in cell-free environments that are resistant to RNase [40] (Additional file 1: Fig. S6B). We also found that the density distribution pattern of peaks was similar to that of reads (Additional file 1: Fig. S6C). Collectively, these implied cfPeak could accurately locate small cfRNA peaks in known transcripts.

Current human transcript reference is still expanding, and many novel transcripts may exist and could be detectable in sequencing data [5, 43], and reads unmappable (GSE71008 pool: 17.2%) to known transcripts might still be mappable to the human genome. To test whether cfPeak could find some cfRNA peaks that are located in those unannotated or novel transcripts, we further mapped left reads in the reference dataset to introns, promoters, enhancers, and repeats; these regions have been reported to host unannotated transcripts or newly identified species (e.g., intron-derived miRNA, intron-derived kink-turn RNA, promoter-derived transcription initiation RNA, enhancer RNA, repeats-derived small RNA) [31–34, 44] (Fig. 4A). We surprisingly found many peaks overlapped with novel sncRNA candidates, including miRNA(-like), snoRNA(-like), tRNA(-like), and ktRNA(-like) transcripts, most of which can bind RBPs or form locally stable structures (Fig. 4B–C, Additional file 1: Fig. S7B–F). The number of novel sncRNA candidates discovered in peak-related regions (± 100 nt from boundaries) is significantly higher than that in randomly shuffled background regions (Additional file 1: Fig. S7A). We also noticed peak regions with potential stem-loop or G-quadruplex (G4) structure, and some even with evolutionary conservation, suggesting potential unknown transcripts (Additional file 1: Fig. S7F). Interestingly, some predicted sncRNA candidates have already been curated in the latest transcript annotation (GENCODE v43 and UCSC RefSeq track) (Fig. 4C, Additional file 1: Fig. S7 F). Their existence in cell-free environments was also reported but less studied [13, 43]. Some notes need to be mentioned when interpreting the results above. First, some peak regions seemed shorter than the predicted novel sncRNA, probably due to the degradation or inefficient capture of full-length sncRNA in miRNA-orientated small RNA libraries, consistent with a previous study [13]. Second, we defined those as novel (sncRNA) transcripts for simplicity, but they are not always produced by ab initio transcription from the genome since one major biogenesis pathway of sncRNA (candidate) is degradation from longer transcript precursors [8]. All these observations showed cfPeak's ability to locate small cfRNA peaks in novel transcripts and highlighted its potential to assist in identifying novel sncRNA.

CfRNA peaks are recurrently protected by proteins, vesicles, and local structures

CfDNA has been reported under the protection of protein-binding and local structures [45–51], while recurrently protected cfRNA fragments are seldom evaluated systematically. Previous efforts have been made to individually explain these recurrently protected cfRNA fragments by the protection of RBP-binding, EV encapsulation, and RNA secondary structure [5, 40]. We hypothesized that noncanonical secondary structures (e.g., RGS) could be another potential protection factor. Inspired by cfDNA studies [45, 52], we defined the tWPS (transcriptomic window protection score) to describe the protection level in transcriptomic regions (Fig. 5A) (details in Additional file 1: "Supplementary Methods" section). We first focused on peaks in known transcripts and found similar patterns in the curve of tWPS near the center of cfRNA peaks detected by cfPeak in the

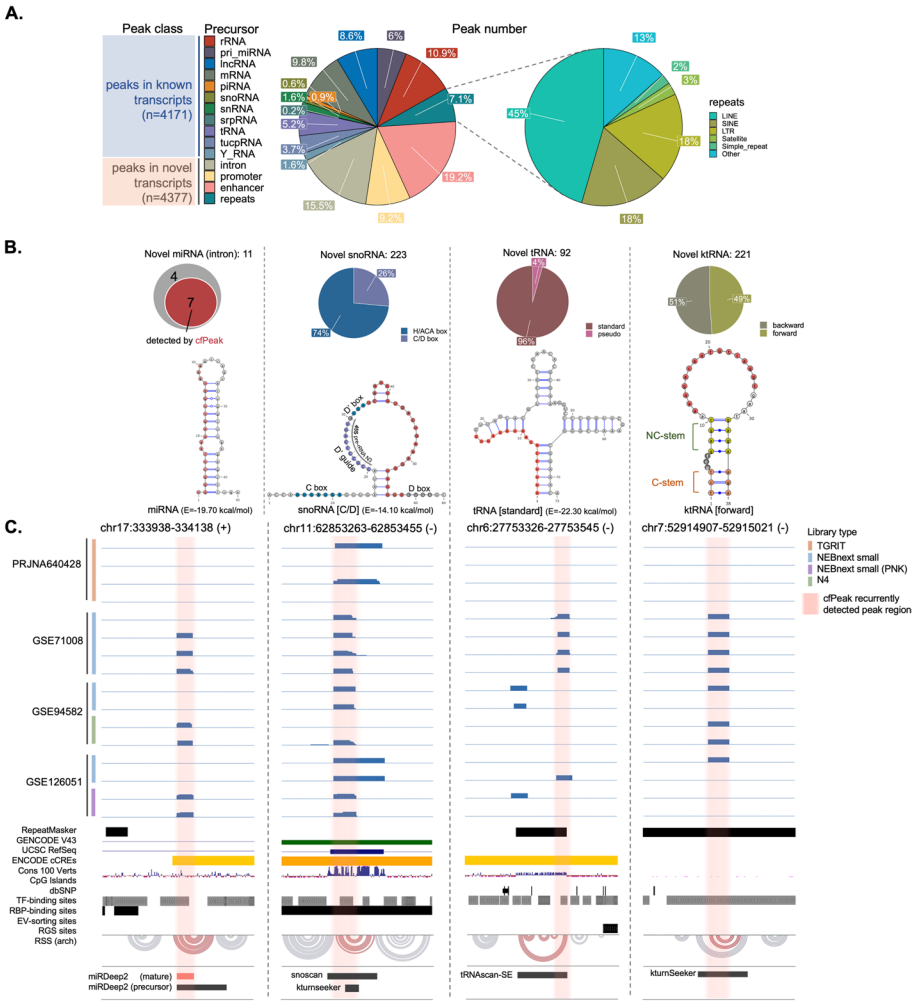


Fig. 4 CfPeak is capable of identifying cfRNA peaks in novel transcripts. **A** Peak number distribution across known transcripts from 11 RNA species and novel transcripts from four regions. Peaks in repeats regions were shown in detail. **B** First column: Venn plot showed the number of intron-derived novel miRNA candidates newly identified by miRDeep2 and detected by cfPeak. Stem-loop RNA structure plot showed an example of novel miRNA candidate. Second column: pie plot showed the number of novel snoRNA candidates of two types identified by snoscan and snoGPS. Stem-loop RNA structure plot showed an example of novel C/D box snoRNA candidate. Third column: pie plot showed the number of novel tRNA candidates of two types identified by tRNAscan-SE. Stem-loop RNA structure plot showed an example of standard tRNA candidate. Fourth column: pie plot showed number of novel ktRNA candidates of two types identified by kturnSeeker. Stem-loop RNA structure plot showed an example of forward ktRNA candidate. Parts of peak regions detected in cfRNA sample (GSE71008 pool) were filled with red color in stem-loop RNA structure plot. **C** Coverage of extended regions in different cfRNA-seq datasets and several UCSC tracks near four candidates mentioned above. Y axis: min–max scaled depth. X axis: hg38 genomic coordinate. TGIRT, thermostable group II intron reverse transcriptases sequencing library; N4, optimized small RNA library (four degenerate nucleotides at the ligation ends); RGS, RNA G4 structural sites; RSS, RNA secondary structure

reference plasma small RNA-seq dataset (GSE71008), implying these narrow peaks were also protected from degradation (Fig. 5B). To confirm our hypothesis, we summarized peaks that overlapped with annotated potential binding sites of protection factors and observed significantly more peak regions that overlapped with RBP-binding, EV-sorting, RGS, and RSS sites compared to size-matched background regions (fold change with 95% CI. RBP: 1.73 [1.68, 1.78]; EV: 72.13 [65.39, 78.87]; RGS: 1.17 [1.09, 1.25]; RSS-MFE: 1.25

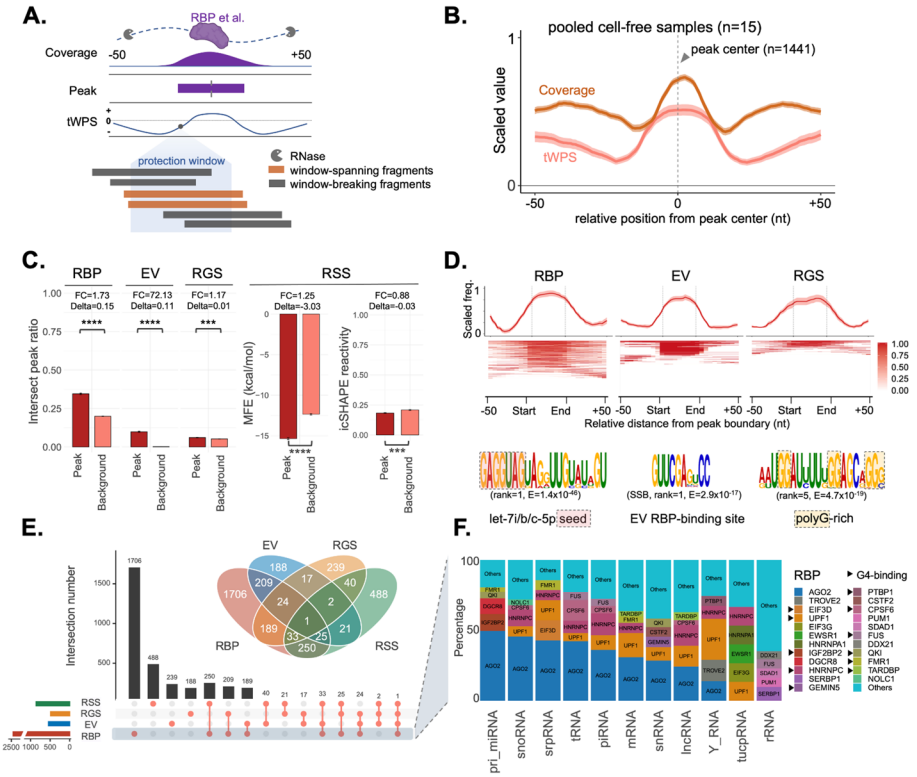


Fig. 5 CfRNA peaks are recurrently protected by proteins, vesicles, and local structures. Recurrently protected regions in the reference dataset (GSE71008). **A** Illustration of (RBP) protected peak region in cell-free environments. **B** Scaled meta plot of reads coverage and tWPS around peak center. Dash line means peak center position. Ribbon area means 95% CI calculated from multiple regions. **C** Left: bar plot showed ratio of RBP-, EV-, and RGS-intersected peak number to total in peak and size-matched background regions. Right: bar plot showed MFE and mean mRNA icSHPAE reactivity value in the extended peak and size-matched background regions. Error bar means SEM calculated from multiple samples. **D** Scaled frequency meta plot and heatmap of annotated RBP-binding sites, EV-sorting sites, and RGS sites occurrence in the extended peak regions. Ribbon area means 95% CI calculated from multiple samples. Regions in heatmap were ranked by mean of scaled occurrence. Representative motif patterns enriched from AGO2-binding-sites-overlapped, EV-sorting, and RGS peak regions were shown below. **E** Upset plot showed overlapped contribution of four protection factors in all annotation-overlapped consensus peaks. Venn plot was shown in the top right. **F** Stacked bar plot showed the percentage of top 5 RBPs (ranked by overlapped peak number) by transcript species in all RBP-binding-sites-overlapped consensus peaks. RGS related RBPs (G4RBP) were highlighted in black triangle. **P* value < 0.05, ***P* value < 0.01, ****P* value < 0.001, *****P* value < 0.0001, Wilcoxon rank sum test, one-tailed (peak has higher ratio and lower MFE or reactivity than background). MFE, minimum free energy; RGS, RNA G4 structure; RSS, RNA secondary structure; tWPS, transcriptomic window protection score. FC: fold change (peak/background). Delta: difference (peak-background)

[1.22, 1.27]; RSS-reactivity: 0.88 [0.84, 0.92]) (Fig. 5C). The meta coverage plot and heatmap of these four annotations' occurrences in the extended peak regions showed that peak regions overlapped with more annotation records than the adjacent flank regions. Further motif analysis in peak regions found that peaks overlapped with AGO2-binding sites hosted seed motifs from let-7*−5p (one of the top abundant miRNA families in plasma). Peaks overlapped with EV-sorting sites significantly enriched binding motifs of EV-abundant RBPs like recently reported SSB protein [53]. Peaks overlapped with RGS

sites enriched repetitive G-rich motifs (Fig. 5D). Notably, peaks in novel transcripts also significantly overlapped with these annotation regions (Additional file 1: Fig. S8B).

As the protection contribution of these factors may overlap, one peak region may be explained by more than one protection factor. To find their overlapping involvement, we defined an overlapping score, and the peak was considered protected by one factor if the overlapping score was higher than the given threshold. The protection factor with the highest overlapping score was assigned to the peak if multiple overlapping records exist (details in " Methods" section). In such a way, we found most of the cfRNA peaks (53.7%) identified by cfPeak were defined as protected by one of the four factors mentioned above (Fig. 5E). In around 38.1% of peaks that were protected by RBPs ($n = 2437$), we found AGO2 was frequently reported as one of the top-ranked RBPs that protected most of the cfRNA peaks, especially primary miRNA. This could be explained by abundant AGO2 protein in cell-free environments reported as small cfRNA carrier (miRNA-mediated silencing complex) [36, 37, 53]. Also, we found many RBPs were annotated to bind RGS, which implied the combined involvement of multiple protection factors [54, 55] (Fig. 5F). Similar results could also be observed from other peak callers (Additional file 1: Fig. S8C–E) and plasma cfRNA-seq datasets. Collectively, these results implied that a considerable fraction of cfRNA fragments are protected by proteins, vesicles, and local structures.

CfRNA peaks in plasma are informative of cancer and cancer types

Recently, well-established cfDNA's metrics that are based on the protection of protein-binding and local structures have shed light on the rising of cfDNA fragmentomics for clinical utilities [48–50], but the clinical potential of recurrently protected cfRNAs is rarely evaluated broadly. CfPeak can sensitively detect cfRNA peaks derived from the minor-origin tissue, and this advantage can be exploited in liquid biopsy where the low fraction of pathological tissue-derived cfDNA or cfRNA usually hinders the noninvasive detection of pathological onset or progression [27, 56]. Thus, we further tested cfPeak's application potential in serving as a diagnostic approach using a public colorectal cancer (CRC) plasma small RNA-seq cohort (PRJNA540919) that includes CRC and normal control (NC) individuals. We first used the logistic regression (LR) models to classify plasma samples between CRC and NC utilizing all peaks in known and novel transcripts. We found models using all peaks in known transcripts had a better classification performance (AUROC = 0.996) than those using miRNA peaks only (AUROC = 0.967), and peaks in novel transcripts also had a high diagnostic value (AUROC = 0.998) (Fig. 6A). A detailed AUROC bar plot by transcript species showed peaks from previously discarded species like mRNA also had a high AUROC, implying the necessity of inclusion of other transcript species in small cfRNA studies of liquid biopsy (Additional file 1: Fig. S9A). We next explored the relationship between cfRNA peaks' abundance in plasma and classification importance in the LR models, and noticed a substantial fraction of low abundant peaks with high importance, while those highly expressed in CRC tissue relative to primary blood cell showed significantly higher classification importance, which implied cancer tissue-derived peaks contributed to a better classification (Fig. 6B).

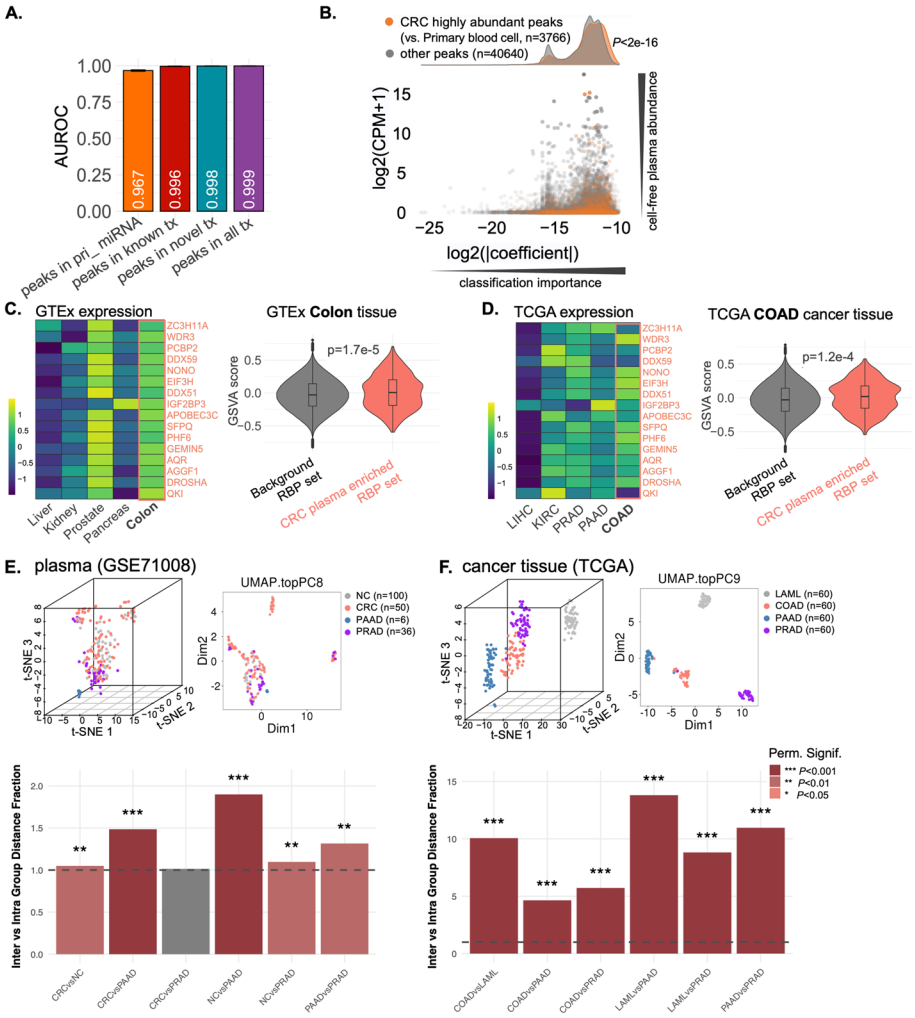


Fig. 6 CfRNA peaks in plasma are informative of cancer and cancer types. **A** AUROC bar plot by different peak matrices in the CRC plasma small RNA-seq cohort (PRJNA540919). **B** Scatter plot showed the relation between peak abundance and classification importance in the CRC plasma cohort (PRJNA540919). X axis means log-transformed absolute coefficient in the LR models (higher value means more importance). Y axis means log-transformed CPM value. CRC highly expressed (relative to primary blood cell) peaks were highlighted in orange. **C** Left: expression profiles of CRC plasma-enriched RBP set across several human tissue types from GTEx. Right: expression profiles (GSVA score) of CRC plasma-enriched RBP set were overexpressed compared with those of randomly shuffled background RBP set across GTEx colon tissue data ($n = 779$). **D** Left: expression profiles of CRC plasma-enriched RBP set across several human cancer types from TCGA. Right: expression profiles (GSVA score) of CRC plasma-enriched RBP set were overexpressed compared with those of randomly shuffled background RBP set across TCGA COAD cancer tissue data ($n = 469$). CRC plasma-enriched RBP set: 16 RBPs enriched from high abundant peaks (top 500 *P* value ranked) in the CRC relative to NC plasma (PRJNA540919). Background RBP set: randomly shuffled RBP set with the same number ($n = 16$, permutation = 2000). Wilcoxon's rank sum test, two-sided. The centers of the boxplots are median values; the bounds of the boxes are 25% and 75% quantiles. The minima are 25% quantile − 1.5*interquartile range (IQR) and the maxima are 75% quantile + 1.5*IQR. **E** 3D t-SNE plot showed the top 500 *P* value ranked cfRNA peaks identified by cfPeak in plasma are informative of cancer types in the three-cancer plasma small RNA-seq discovery cohort (GSE71008). Significance of group distance fraction below was calculated by permutation, same as right. **F** 3D t-SNE plot showed the top 500 *P* value ranked cfRNA peaks from plasma are informative of cancer types in the three-cancer tissue small RNA-seq validation cohort (TCGA)

To investigate whether cfRNA peaks from cancer tissue alone could be used for robust cancer detection, we defined the subset of peaks with differentially higher abundance and frequency in specific tissue type relative to primary blood cell as tissue dominantly contributed peaks (TDCPs) (details in " Methods" section). A total of 1745 CRC TDCP candidates were compiled from the CRC tissue small RNA-seq discovery cohort (TCGA) (Additional file 1: Fig. S10A–D). Based on the assumption that plasma samples in CRC patients constitute more colon- or CRC-derived cfRNA fragments than those in NC, we proposed a TDCP candidates-based peak-index without training for noninvasive CRC detection and observed the CRC peak-index kept high performance (AUROC = 0.99) in separating CRC from NC samples in the CRC plasma small RNA-seq validation cohort (PRJNA540919) (Additional file 1: Fig. S10E).

We further checked the presence of specific tissue-derived cfRNA peaks in 1745 identified CRC TDCP candidates from TCGA cancer tissue samples, and 109 peaks were reidentified in plasma as CRC TDCPs with potential colon-origin under the same filtering metric of tissue (Additional file 1: Fig. S11A). We found many of these tissue-derived detectable peaks in plasma were low abundant, consistent with previous observations in long RNA transcripts [27]. These TDCPs showed differential coverage with sample heterogeneity between groups in plasma as expected, among which one known mature miRNA (miR-6803-5p, ENST00000615997.1) peak overlapped with annotated protection factors and was reported as CRC biomarker in plasma nanovesicles [57] (Additional file 1: Fig. S11B). These results demonstrated that cfRNA peaks in plasma identified by cfPeak are informative of cancer, and tissue-derived small fragments alone can serve for robust detection in clinical cohort.

CfDNA studies have shown the transcription factor-binding status in cancer tissue could be revealed by plasma cfDNA profiling [45, 46, 51, 58]. As cfRNA fragments with different origins can be protected by multiple protection factors, some of which are involved in cellular post-transcriptional regulation, we next investigated whether some cfRNA peaks can potentially inform the biological status of the original tissue like RBP-binding footprint. We focused on peaks that overlapped with annotated RBP-binding sites in known transcripts. RBP-binding motif enrichment analysis from peaks with differentially higher abundance in CRC plasma found 16 RBPs that could bind CRC patients' cfRNA, many of which also have been reported to associate with CRC development or progression (e.g., WDR5, PCBP2, NONO, IGF2BP3, GEMIN5, AGGF1). Gene set variation analysis showed relatively higher expression of enriched RBPs than size-matched permutated RBPs in GTEx colon tissue samples, and a similar result was observed in TCGA COAD cancer tissue samples for these enriched RBPs (Fig. 6C–D). These implied cfRNA peaks in plasma could potentially be utilized to infer the RBP-binding footprint in the original tissue.

Expanding catalogs of tissue-specific sncRNAs have been reported [4, 59], but a study using cell-free sncRNA for the tissue-of-origin investigation has not been conducted yet. As parts of cfRNA peaks were tissue-derived (Additional file 1: Fig. S11), and some of them were under the protection of RBP in the original tissue, which could be an important pathway in the cellular biogenesis of sncRNA [11], we thus wondered whether these cfRNA peaks could discriminate different cancer tissue types (Additional file 1: Fig. S12A). Using a three-cancer plasma small RNA-seq

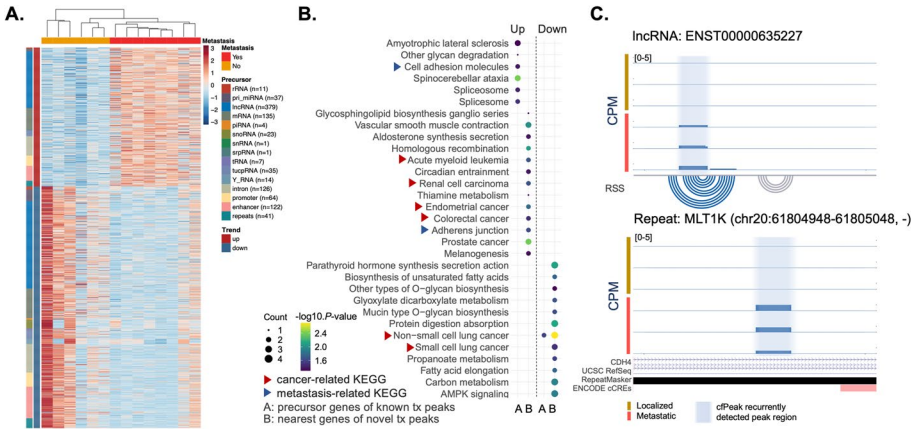


Fig. 7 CfRNA peaks in plasma are informative of cancer types and metastatic status. Identification of informative cfRNA peaks in the metastatic OSCC plasma small RNA-seq cohort (in-house). **A** Heatmap showed the relative abundance (logCPM) of differentially abundant (top 1000 *P* value ranked) peaks in OSCC plasma samples. **B** Dot plot showed enriched KEGG pathways of differentially abundant (top 1000 *P* value ranked) peaks that grouped by the trend of abundance in metastatic relative to localized OSCC plasma. **C** Differential normalized coverage (CPM) between groups in two extended example peak regions

cohort (GSE71008), we next tested the multi-classification potential of cfRNA peaks in plasma. Their abundance levels in each group were compared with those in the rest groups and peaks with differential abundance were selected as group-specific candidates (Additional file 1: Fig. S12B). As expected, t-SNE results of the abundance matrix of candidate peaks showed that these cfRNA peaks in plasma could separate different cancer types in the three-cancer plasma cohort. When validated in TCGA cancer tissue small RNA-seq data, we surprisingly found that the same candidate peaks were also informative of these cancer types (Fig. 6E–F). This highlighted the potential of cfRNA peaks in plasma identified by cfPeak for cancer-type discrimination.

CfRNA peaks in plasma are informative of cancer metastatic status

We next explored whether cancer status in a closer view could be obtained from cfRNA peaks. Oral cancer is characterized by poor prognosis and low survival rate despite sophisticated surgical and radiotherapeutic modalities. Lymphatic metastasis of oral cancer is a complex process involving multiple post-transcriptional biological processes. Thus, noninvasive cancer metastasis detection in plasma could be challenging and beneficial. We recruited a metastatic cohort of oral squamous cell carcinoma (OSCC), a common oral cancer, and profiled small cfRNA in plasma samples. We found that cfRNA peaks with differential abundance in plasma could clearly distinguish patients who had localized OSCC from those who had metastatic OSCC (Fig. 7A). Overrepresentation analysis showed many cancer- and metastasis-related KEGG pathways could be enriched in precursor or nearest genes of upregulated (higher abundance in metastasis group) peaks, especially for those in novel transcripts (Fig. 7B). We observed one of the differentially abundant peaks within lncRNA LINC01108 (ENST00000635227) had differential coverage among samples from two groups, and the peak's locations seemed to overlap with local RSS site, implying RSS

protection might contribute to the peak formation. Also, another peak in the minus strand of repeats was close to CDH4 gene, which has been reported to inhibit ferroptosis in OSCC cells [60], and ferroptosis has been known to relate to cancer metastasis [61, 62] (Fig. 7C). Collectively, these results implied the potential of cfRNA peaks identified by cfPeak for noninvasive detection of the metastatic status of cancer.

Discussion

Peak calling is an essential step in the upstream preprocessing of transcriptomic sequencing data, which aims to identify enriched read clusters against background noise [63]. There are many statistics-based peak callers, such as Piranha [64], CLIPper [29], and CLAM [35], which were developed for the discovery of significant signals in post-transcriptional regulations like RBP-binding and RNA modification events. Piranha internally contains two modes; the original ZTNB (zero-truncated negative binomial) mode determines a fixed depth as the threshold. The other covariates-adjusted ZTNB mode allows correction of different transcripts' abundance at bin scale; both modes generally sacrifice sensitivity for accuracy, leading to much fewer peaks than other methods, as previously stated [65]. CLIPper and CLAM indirectly infer different statistical backgrounds in regions of fixed length near peaks or at full-length transcript-scale to handle background noise (Additional file 1: Table S1). All these traditional RNA-seq peak callers tend to underestimate short fragment size-matched local regions. Meanwhile, with the biological applications of machine learning methods based on deep neural networks, genomic peak callers employing a CNN model were developed to predict high-quality peaks in a supervised manner, like CNN-peaks [66], LanceOtron [67], and DEOCSU [68]. These models rely on peaks labeled by human researchers and learn the latent patterns from visual inspections. We thus expect that newly developed transcriptomic peak caller assisted by machine learning models could also learn transcriptomic-specific peak patterns.

Some essential differences exist among cellular CLIP-seq, cellular small RNA-seq, and cell-free small RNA-seq. First, most small cfRNA-seq datasets were generated using low-input cfRNA samples fragmented by cell-free RNase with no size-matched control and no additional cross-linking or immunoprecipitation step. Second, unlike CLIP-seq that mainly captures RBP-binding sites, small cfRNA-seq captures fragments protected from more diverse factors, additionally including EV-sorting and local structure, but the effect size is relatively small (Additional file 1: Fig. S13). These differences point to the critical characteristics among data types that require additional consideration during peak analysis.

In addition to the risk of improper modeling or inadequate parameters [5, 11], another significant limitation of applying traditional methods to small cfRNA-seq is that most of them are insensitive to the discovery of low-abundance signal sites. However, this overlooked situation is highlighted in cfRNA studies; many cfRNA molecules are fragmented before secreting from cell and during circulating, and adjacent peaks in small cfRNA-seq may have different cellular origins with varying sorting mechanisms and regulatory functions, such as sncRNA fragments that recently reported [5, 69]. Low abundant peaks near high abundant peaks tend to be masked by traditional methods, but they may originate from tissues that contribute

a small fraction to the cfRNA admixture, which happen to be valuable in liquid biopsy since they may carry the information of the tissue of interest and should not be missed out. Furthermore, their abundance in traditional sequencing data types does not always represent their actual biological importance; sometimes the underestimation was caused by insufficient capturing or amplification bias during library preparation [12, 69, 70]. In summary, the fragmented and heterogeneous origin characteristics of cfRNA pose unique challenges for identifying informative cfRNA peaks (Additional file 1: Fig. S14).

Our peak analysis in small cfRNA-seq also has some limitations. First, cfPeak maps cfRNA reads to the transcriptome sequentially before the genome; this strategy is more sensitive to known transcripts and more suitable if some specific transcript species are of interest. Second, we only included three representative methods in the evaluation; many other peak callers can be added and compared if compatible with transcriptome-mapped reads, like CLIPick [65], CTK [71], and PIPE-CLIP [63]. Third, the transcriptome-level EM in our pipeline is adapted from a genomic version [35], and we assumed that the predefined mapping order could be seen as a preference for regions in EM reassignment. Fourth, we mainly focused on short, narrowly protected regions in small RNA-seq (or miRNA-seq), and cfPeak is also applicable to total or long RNA-seq (Figs. 1A, 4C). Fifth, we only considered four protection factors of cfRNA in the analysis for simplicity. For EV-sorting candidate peaks, we also carried out independent validation in plasma to show the rationale and feasibility of our annotation procedures (Additional file 1: Fig. S15A–C). Other annotation types like RBP-binding sites and G4 structural sites were downloaded from published database, and they might also need additional experimental validation if available. In fact, some other RNA carrier types or protection factors exist; further detailed annotation integrating more datasets may explain these cfRNA fragments more precisely and clearly, and the protection contribution might also change accordingly [36, 37, 53, 72]. Also, some longer fragments have been proved to exist in cell-free environments under different mechanisms, like full-length structured intron and pre-tRNA [5, 73]. Meanwhile, inter-molecular interactions were not considered, though heterodimer of tRNA halves has been reported to exist in cell-free environments [40]. In addition to canonical RSS, other advanced structures like i-motif, kink-turn, and inter-molecular G4 may also exist in cfRNA [44]. Sixth, further validation using standard assay is needed. For example, the RBP-binding sites we used were extracted from our previous database work [74], and the RBP-cfRNA binding status in vivo needs confirmation in human biofluids [53]. Notably, RGS may not be stable in EV-free plasma, but theoretically, it can exist in cell and EV with suitable ion concentration. The actual origin of cfRNA in clinical samples also needs further confirmation, like the detection of human cfRNA in plasma samples from the patient-derived xenograft (PDX) mouse model. The performance cfPeak of clinical application might also be affected by multiple conditions like cohort size, and still need further validation in larger cohort.

Conclusions

CfPeak is a new computational method designed for peak analysis in small cfRNA-seq and is potentially applicable to broad data types. It shows an improved ability to identify recurrently protected cfRNA peaks and pave the way to cfRNA fragmentomics. We highlight the significance of rescuing low abundant peaks that potentially derived from tissue in clinical cohorts, and demonstrate the clinical potential of these cfRNA narrow peaks in liquid biopsy.

Methods

Cohort design, sample collection, and processing

The individuals in the metastatic oral squamous cell carcinoma (OSCC) cohort were recruited from Beijing Stomatological Hospital (CMU, Beijing). Informed consent was obtained for all patients prior to the enrolment of this study. We included six localized OSCC cancer patients and eight metastatic OSCC cancer patients in total; the mean age is 59 (male: 58, female: 60). The lymph node metastatic status of OSCC was annotated if TNM stage was proven N + histopathologically (Additional file 1: Table S2). The study was approved by the Ethics Committee of Beijing Stomatological Hospital (CMU, Beijing) (CMU-IRB-KJ-PJ-2022–15) and complied with the Declaration of Helsinki.

Sample collection and processing workflow were performed as previously reported [75]. In short, peripheral whole blood samples were collected from individuals before therapy or surgery using EDTA-coated vacutainer tubes. Next, plasma was separated within 2 h after collection. All plasma samples were aliquoted and stored at − 80 °C. Then, cfRNAs were extracted using miRNeasy serum/plasma miRNA isolation kit (QIAGEN, Shanghai, China), and DNA contamination was removed by Recombinant DNase I (RNase-free) (TAKARA, Beijing, China). Small libraries were prepared according to the manual of QIAseq miRNA library kit (QIAGEN, Shanghai, China). Followed by library quantification with Qubit dsDNA HS Kit. Library fragment size and quality were checked with Agilent 2100 Bioanalyzer. Libraries were sequenced on Illumina HiSeq X-ten with PE150.

For in-house validation of EV-sorting candidate peaks, six healthy donors were recruited in Peking Union Medical College Hospital (PUMCH, Beijing). Informed consent was obtained for all healthy donors prior to the enrolment of this study. The study was approved by the Ethics Committee (JS-3386D) and complied with the Declaration of Helsinki. Total cfRNA-seq preparation in healthy donor's plasma, peripheral whole blood samples were collected from individuals before therapy or surgery using EDTA-coated vacutainer tubes. Plasma was separated within 2 h after collection by centrifuge at 1900 g for 10 min at 4 °C. All plasma samples were aliquoted and stored at − 80 °C. EV-depleted plasma cfRNAs were extracted using QIAzol Lysis Reagent (QIAGEN, Beijing, China), and DNA contamination was removed by Recombinant DNase I (RNase-free) (TAKARA, Beijing, China). The total cfRNA libraries were prepared using our in-house protocol that improved from our previously published method, DETECTOR-seq [76]. Library quantification was performed by Qubit dsDNA HS Kit. Library fragment size and quality were checked using Agilent 2100 Bioanalyzer. Libraries were sequenced on DNBSEQ-T7 (MGI Tech.) with PE150.

Peak calling in cfRNA peak analysis pipeline

We implement four peak caller options in the peak analysis pipeline: Piranha, CLIPper, CLAM, and cfPeak. Piranha v1.2.1 takes primary-mapped bam as input and calls peaks with the flag "-b 5." CLIPper v2.1.2 takes primary-mapped bam as input and calls peaks with default parameters and a modified transcriptomic gtf as reference. CLAM v1.2.0 takes multi-mapped bam as input and intrinsically reassigns multi-mapped reads and calls peaks with the flag "permutation_callpeak –extend 5." CfPeak v1.0.6 takes EM-reassigned bam as input and calls peaks with the flag "call_peaks_localmax –boundary background –mode local –max-iter 100 –permutate-pval 0.05 –poisson-pval 1 –bin-width 10 –min-peak-length 10 –max-peak-length 200 –thread 6 –decay 0.5 –min-cov 2"; it outputs a peak file where the columns represent "transcript, start, end, name, maximum depth, strand, maximum position, background depth, permutation *P*-value, Poisson *P*-value" from left to right. The CNN filtering outputs a peak file where the columns represent "transcript, start, end, name, probability to be false peak (0–1), strand." One could choose to directly skip traditional Poisson significance test and CNN filtering step to get more peaks of low abundance by setting flag "–poisson-pval 1" in peak calling script and "–threshold 1" in anomaly detection script. All other parameters in peak calling modules were set as default. To ensure consistency of different methods, we further applied peak length filtering for each peak file from four methods, and only peaks with lengths between 10 and 200 nt were kept.

Interval overlapping analysis

In annotation-peak overlapping analysis, length-matched background regions were generated by randomly shuffling peak regions of each sample across all 11 transcript species using bedtools v2.30.0 with flags "shuffle -noOverlapping." Peak regions and annotation records were intersected for each sample using bedtools with flags "intersect -wao -s." The overlapping score was defined as the product of intersection-peak ratio and intersection-annotation ratio, and an overlapping event was defined as having an overlapping score higher than 0.01. The ratio of annotation-overlapped peaks to all peaks was compared to that calculated from background regions; group comparison was performed by Wilcoxon rank sum test (one-tailed), and error bar means standard error of mean (SEM) calculated from multiple samples (Fig. 5C).

For the meta coverage plot and heatmap of annotation records (Fig. 5D), we first extended 50 nt from both boundaries in peak regions of each sample, and peaks beyond full-length boundary were removed, then the records of three annotations (RBP-binding, EV-sorting, and RGS sites) in the extended peak regions were fed into EnrichedHeatmap v1.16.0 R package (normalizeToMatrix function) to calculate the occurrences of annotation records of each sample, the value at each position was min–max scaled (two-rounds) across all positions in a region and smoothed (two-rounds) by loess model in R. Peak regions in heatmap were ranked by mean of scaled occurrence of annotation records.

Mature miRNA boundary positions (genomic coordinates) were downloaded from miRBase release 22, and only mature miRNA records of high abundance (median counts ≥ 5) in the reference plasma small RNA-seq dataset (GSE71008) were kept for overlapping analysis. Transcriptomic peak regions of the pooled sample from four methods

were converted into genomic coordinates and were overlapped with annotated miRNA regions using bedtools; overlapping records were filtered by removing those with overlapping bases less than 10 or overlapping score smaller than 0.01, and only one mature miRNA with the highest overlapping score was assigned for each peak if multiple overlapping records exist (Additional file 1: Fig. S3B).

In protection contribution analysis, RBP-, EV-, and RGS-protected peaks were defined as overlapped with annotation, and the overlapping score was higher than 0.01 (Fig. 5E).

In peak-peak overlapping analysis, peak-peak overlapping analysis (Additional file 1: Fig. S3C) was performed using Intervene [77] v0.6.5 using consensus peaks of the pooled sample.

Classification based on machine learning

We attempt to distinguish colorectal cancer (CRC) cancer patients and normal controls (NC) using plasma cfRNA peaks by the logistic regression (LR) models. First, we normalized the raw count matrices of consensus peaks in known and novel transcripts. We used bootstrap strategy to perform the model evaluation. One hundred times of uniform resampling with replacement of all samples was performed to generate 100 discovery sets, each with the same scale as the original samples. On average, in each turn of random resampling with replacement, 63.2% of original samples were included in the discovery sets, and the rest served as test sets for evaluation. The LR models (L2 regularization) were trained on each discovery set with hyperparameter tuning of threefold cross-validation, and classification probabilities were predicted on paired test sets by the LR model. Then, AUROC and AUPR metrics were estimated by the final combination of each sample's average predicted probabilities at every turn of resampling. The feature importance was estimated by peak coefficients computed by the LR model and averaged over all turns of bootstrap.

Identification of tissue dominantly contributed peaks

CfRNA fragments in cell-free plasma samples are mainly derived from blood cell (blood-derived) compared to tissue cell (tissue-derived) like colon. In cancer patients' plasma samples, the fraction of cancer tissue-derived cfRNA fragments in all tends to increase [21, 27, 78, 79]. Tissue dominantly contributed peaks (TDCPs) were defined as peak regions in which cfRNA fragments are most likely tissue-derived instead of blood-derived. For a specific cancer plasma cohort paired with normal controls, the TDCP candidates (part of which can be detected in plasma as TDCPs) were identified from the same cancer tissue and blood control sample through four criteria: (a) the 90th percentile abundance (CPM) of this gene in the control group is less than 0.1, (b) the ratio of control samples with detectable signal (count ≥ 1) is less than 10%, (c) the ratio of specific cancer samples with detectable signal is greater than 10%, (d) the gene has differentially higher abundance in the specified cancer group compared to the control group (P value < 0.1). The two-group differential abundance was calculated by the edgeR [80] v3.28.1 package (glmFit and glmLRT functions) in R.

Cancer peak-index

Take CRC, for example, the cancer peak-index (CPI) score was defined as the mean of the *z*-score standardized abundance (CPM) for all CRC TDCP candidates:

$$Y_{i,j} = \frac{CPM_{i,j} - mean(CPM.LAML_i)}{sd(CPM.ALL_i)}$$

$$CPI_j = mean(Y_{i,j})$$

where i denotes the index of peak in CRC TDCP candidates, j denotes the index of plasma sample, *CPM.LAML* is the vector of abundance of CRC TDCP candidates calculated from TCGA LAML small RNA-seq cohort, and *CPM.ALL* is the vector of abundance of CRC TDCP candidates calculated from both TCGA LAML and COAD small RNA-seq cohort. LAML (primary blood cell) cohort was treated as control group in TCGA data. AUROC for each dataset was calculated by pROC v1.17.0.1 R package [81] using CPI score as a metric. We used the same consensus peak set from plasma to count reads to keep the count matrix consistent across plasma and tissue datasets.

Additional software and algorithms used in this study are listed in Additional file 1: Supplementary Methods [42, 44, 54, 55, 74, 80, 82–112].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03590-x.

Additional file 1. Supplementary Figures S1 to S15 and Supplementary Tables S1–S3.

Additional file 2. Peer review history.

Acknowledgements

We thank all providers of public data used in this work for their sharing. In particular, we thank Dr. Victor Ambros at the University of Massachusetts Medical School for providing cell-free AGO2 RIP-seq dataset. We thank Shida Zhu, Fang Chen, and Wen-Jing Wang at BGI-Research (Shenzhen) for approving our request of cell-free total RNA-seq dataset. We also thank Dr. Junchao Shi at the Beijing Institute of Genomics, Chinese Academy of Sciences, and Dr. Tong Zhou at the University of Nevada for their detailed consultations of machine learning strategy and experimental validation in their work. Some parts of the figures were created with BioRender.com.

Review history

The review history is available as Additional file 2.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

Z.J.L. and P.B. conceived and designed the project. S.X. and Q.Z. performed the experiments. P.B. and T.W. developed the framework of the whole pipeline. P.B., T.W., and Y.T. processed the data and completed the bioinformatics analyses. T.W., P.B., and H.M. performed machine learning classification analyses. E.B-R. collected parts of the G4 structural annotations. H.R., Q.Z., and M.M. curated the sample and clinical information. All authors contributed to the manuscript writing.

Funding

This work is supported by the National Key Research and Development Program of China (2024YFC2510300, 2024YFC3405900, 2021YFC2501300), National Natural Science Foundation of China (82371855, 82341101, 32170671), CAMS Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-1-005), National High Level Hospital Clinical Research Funding (2022-PUMCH-B-013, D-009), Beijing Hospitals Authority Youth Programme (QML20211501), Innovation Foundation of Beijing Stomatological Hospital Capital Medical University (21-09-16), and Tsinghua University Initiative Scientific Research Program of Precision Medicine (2022ZLA003). This study was also supported by the BioComputing Platform of the Tsinghua University Branch of China National Center for Protein Sciences.

Data availability

CfPeak peak calling module and cfRNA peak analysis pipeline are publicly accessible in the GitHub repository under GNU General Public License v2.0 (https://github.com/lulab/cfPeak) [113] and ZENODO (DOI: 10.5281/zenodo.14773110) [114]. All FASTQ files generated in our work or other laboratories can be downloaded in the NCBI Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo), Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra),

and China National GeneBank Database (CNGBdb, https://db.cngb.org). Public small RNA-seq datasets of blood cell and tissue include GSE50676 [39] and The Cancer Genome Atlas (TCGA,https://portal.gdc.cancer.gov/) [111]. Public small RNA-seq datasets of cell-free specimens include GSE71008 [115], GSE94533 [116], GSE110381 [117], GSE94582 [118], GSE123972 [24], GSE126051 [12], PRJNA540919 [119], Geekiyanage et al. (2020-PNAS) [37], GSE148861 [10], PRJNA640428 [5], GSE221088 [76], GSE129255 [120], GSE112343 [121], GSE56686 [122], and CNP0003091 [123]. Small RNA-seq datasets generated in this study include GSE238204 (small cfRNA-seq) [124] and GSE278414 (EV total cfRNA-seq) [125].

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Beijing Stomatological Hospital (CMU, Beijing) (CMU-IRB-KJ-PJ-2022–15) and the Ethics Committee of Peking Union Medical College Hospital (PUMCH, Beijing) (JS-3386D). Informed consent was obtained for all patients prior to the enrolment of this study. All the experimental methods comply with the Helsinki Declaration.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]MOE Key Laboratory of Bioinformatics, State Key Lab of Green Biomanufacturing, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China. [2]Department of Rheumatology and Clinical Immunology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China. [3]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. [4]Department of Oral and Maxillofacial & Head and Neck Oncology, Beijing Stomatological Hospital, Capital Medical University, Beijing, China. [5]Institute for Precision Medicine, Tsinghua University, Beijing 100084, China. [6]Peking University–Tsinghua University–National Institute of Biological Sciences Joint Graduate Program, School of Life Sciences, Tsinghua University, Beijing, China. [7]Academy for Advanced Interdisciplinary Studies (AAIS)and, Sciences Joint Graduate Program (PTN), Peking University, Beijing, China. [8]Centre for Genomics and Oncological Research (GENYO), Avenida de La Ilustración 114, Granada 18016, Spain. [9]Department of Biochemistry and Molecular Biology II, Faculty of Pharmacy, University of Granada, Granada, Spain. [10]Instituto de Investigación Biosanitaria Ibs.GRANADA, Hospital Virgen de Las Nieves, Granada, Spain. [11]The Center for Regeneration Aging and Chronic Diseases, School of Basic Medical Sciences, Tsinghua University, Beijing, China.

Received: 18 August 2023 Accepted: 25 April 2025
Published online: 08 May 2025

References

1. Kalluri R, LeBleu VS. The biology, function, and biomedical applications of exosomes. Science. 2020;367:eaau6977.
2. Mateescu B, Kowal EJK, van Balkom BWM, Bartel S, Bhattacharyya SN, Buzás EI, et al. Obstacles and opportunities in the functional analysis of extracellular vesicle RNA – an ISEV position paper. J Extracell Vesicles. 2017;6:1286095.
3. Tosar JP, Witwer K, Cayota A. Revisiting extracellular RNA release, processing, and function. Trends Biochem Sci. 2021;46:438–45.
4. Isakova A, Fehlmann T, Keller A, Quake SR. A mouse tissue atlas of small noncoding RNA. Proc Natl Acad Sci. 2020;117:25634–45.
5. Yao J, Wu DC, Nottingham RM, Lambowitz AM. Identification of protein-protected mRNA fragments and structured excised intron RNAs in human plasma by TGIRT-seq peak calling. eLife. 2020;9:e60743.
6. Tosar JP, Segovia M, Castellano M, Gámbaro F, Akiyama Y, Fagúndez P, et al. Fragmentation of extracellular ribosomes and tRNAs shapes the extracellular RNAome. Nucleic Acids Res. 2020;48:12874–88.
7. Nechooshtan G, Yunusov D, Chang K, Gingeras TR. Processing by RNase 1 forms tRNA halves and distinct Y RNA fragments in the extracellular environment. Nucleic Acids Res. 2020;48:8035–49.
8. Shi J, Zhou T, Chen Q. Exploring the expanding universe of small RNAs. Nat Cell Biol. 2022;24:415–23.
9. Zhang Z, Zhang J, Diao L, Han L. Small non-coding RNAs in human cancer: function, clinical utility, and characterization. Oncogene. 2021;40:1570–7.
10. Gu W, Shi J, Liu H, Zhang X, Zhou JJ, Li M, et al. Peripheral blood non-canonical small non-coding RNAs as novel biomarkers in lung cancer. Mol Cancer. 2020;19:159.
11. Fish L, Zhang S, Yu JX, Culbertson B, Zhou AY, Goga A, et al. Cancer cells exploit an orphan RNA to drive metastatic progression. Nat Med. 2018;24:1743–51.
12. Giraldez MD, Spengler RM, Etheridge A, Goicochea AJ, Tuck M, Choi SW, et al. 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. von Felden J, Garcia-Lezana T, Dogra N, Gonzalez-Kozlova E, Ahsen ME, Craig A, et al. Unannotated small RNA clusters associated with circulating extracellular vesicles detect early stage liver cancer. Gut. 2021;71:1935–6.
14. Kuksa PP, Amlie-Wolf A, Katanić Ž, Valladares O, Wang L-S, Leung YY. DASHR 2.0: integrated database of human small non-coding RNA genes and mature products. Kelso J, editor. Bioinformatics. 2019;35:1033–9.

15. Fehlmann T, Backes C, Pirritano M, Laufer T, Galata V, Kern F, et al. The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. Nucleic Acids Res. 2019;47:4431–41.
16. Anfossi S, Babayan A, Pantel K, Calin GA. Clinical utility of circulating non-coding RNAs — an update. Nat Rev Clin Oncol. 2018;15:541–63.
17. Zhu Y, Wang S, Xi X, Zhang M, Liu X, Tang W, et al. Integrative analysis of long extracellular RNAs reveals a detection panel of noncoding RNAs for liver cancer. Theranostics. 2021;11:181–93.
18. Hulstaert E, Morlion A, Avila Cobos F, Verniers K, Nuytens J, Vanden Eynde E, et al. Charting extracellular transcriptomes in the Human Biofluid RNA Atlas. Cell Rep. 2020;33: 108552.
19. Munchel S, Rohrback S, Randise-Hinchliff C, Kinnings S, Deshmukh S, Alla N, et al. Circulating transcripts in maternal blood reflect a molecular signature of early-onset preeclampsia. Sci Transl Med. 2020;12:eaaz0131.
20. Pan W, Ngo TTM, Camunas-Soler J, Song C-X, Kowarsky M, Blumenfeld YJ, et al. Simultaneously monitoring immune response and microbial infections during pregnancy through plasma cfRNA sequencing. Clin Chem. 2017;63:1695–704.
21. Koh W, Pan W, Gawad C, Fan HC, Kerchner GA, Wyss-Coray T, et al. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. Proc Natl Acad Sci. 2014;111:7361–6.
22. Yan Z, Zhou Z, Wu Q, Chen ZB, Koo EH, Zhong S. Presymptomatic increase of an extracellular RNA in blood plasma associates with the development of Alzheimer's disease. Curr Biol. 2020;30:1771–82.
23. Toden S, Zhuang J, Acosta AD, Karns AP, Salathia NS, Brewer JB, et al. Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing. Sci Adv. 2020;6:eabb1654.
24. Tan C, Cao J, Chen L, Xi X, Wang S, Zhu Y, et al. Noncoding RNAs serve as diagnosis and prognosis biomarkers for hepatocellular carcinoma. Clin Chem. 2019;65:905–15.
25. Rozowsky J, Kitchen RR, Park JJ, Galeev TR, Diao J, Warrell J, et al. exceRpt: a comprehensive analytic platform for extracellular RNA profiling. Cell Syst. 2019;8:352-357.e3.
26. Allen RM, Zhao S, Ramirez Solano MA, Zhu W, Michell DL, Wang Y, et al. Bioinformatic analysis of endogenous and exogenous small RNAs on lipoproteins. J Extracell Vesicles. 2018;7:1506198.
27. Larson MH, Pan W, Kim HJ, Mauntz RE, Stuart SM, Pimentel M, et al. A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. Nat Commun. 2021;12:2357.
28. Ning C, Cai P, Liu X, Li G, Bao P, Yan L, et al. A comprehensive evaluation of full-spectrum cell-free RNAs highlights cell-free RNA fragments for early-stage hepatocellular carcinoma detection. eBioMedicine. 2023;93:104645.
29. Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol. 2013;20:1434–42.
30. Zheleznyakova GY, Piket E, Needhamsen M, Hagemann-Jensen M, Ekman D, Han Y, et al. Small noncoding RNA profiling across cellular and biofluid compartments and their implications for multiple sclerosis immunopathology. Proc Natl Acad Sci. 2021;118: e2011574118.
31. Hubé F, Ulveling D, Sureau A, Forveille S, Francastel C. Short intron-derived ncRNAs. Nucleic Acids Res. 2017;45:4768–81.
32. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, et al. Tiny RNAs associated with transcription start sites in animals. Nat Genet. 2009;41:572–8.
33. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7.
34. Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, et al. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell. 2010;141:956–69.
35. Zhang Z, Xing Y. CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Res. 2017;45:9260–71.
36. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, et al. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. Proc Natl Acad Sci. 2011;108:5003–8.
37. Geekiyanage H, Rayatpisheh S, Wohlschlegel JA, Brown R, Ambros V. Extracellular microRNAs in human circulation are associated with miRISC complexes that are accessible to anti-AGO2 antibody and can bind target mimic oligonucleotides. Proc Natl Acad Sci. 2020;117:24213–23.
38. Morrissey A, Shi J, James DQ, Mahony S. Accurate allocation of multimapped reads enables regulatory element analysis at repeats. Genome Res. 2024;34:937–51.
39. Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature. 2014;505:706–9.
40. Tosar JP, Gámbaro F, Darré L, Pantano S, Westhof E, Cayota A. Dimerization confers increased stability to nucleases in 5′ halves from glycine and glutamic acid tRNAs. Nucleic Acids Res. 2018;46:9081–93.
41. Max KEA, Bertram K, Akat KM, Bogardus KA, Li J, Morozov P, et al. Human plasma and serum extracellular small RNA reference profiles and their clinical utility. Proc Natl Acad Sci. 2018;115:E5334–43.
42. Gebert D, Hewel C, Rosenkranz D. unitas: the universal tool for annotation of small RNAs. BMC Genomics. 2017;18:644.
43. Boivin V, Reulet G, Boisvert O, Couture S, Elela SA, Scott MS. Reducing the structure bias of RNA-Seq reveals a large number of non-annotated non-coding RNA. Nucleic Acids Res. 2020;48:2271–86.
44. Li B, Liu S, Zheng W, Liu A, Yu P, Wu D, et al. RIP-PEN-seq identifies a class of kink-turn RNAs as splicing regulators. Nat Biotechnol. 2023;42:119–31.
45. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell. 2016;164:57–68.
46. Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun. 2019;10:4666.

47. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun. 2021;12:5060.
48. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. 2021;372:eaaw3616.
49. Hudecova I, Smith CG, Hänsel-Hertsch R, Chilamakuri C, Morris JA, Vijayaraghavan A, et al. Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. Genome Res. 2021;32:215–27.
50. Thierry AR. Circulating DNA fragmentomics and cancer screening. Cell Genomics. 2023;3: 100242.
51. Foda ZH, Annapragada AV, Boyapati K, Bruhm DC, Vulpescu NA, Medina JE, et al. Detecting liver cancer using cell-free DNA fragmentomes. Cancer Discov. 2023;13:616–31.
52. Markus H, Zhao J, Contente-Cuomo T, Stephens MD, Raupach E, Odenheimer-Bergman A, et al. Analysis of recurrently protected genomic regions in cell-free DNA found in urine. Sci Transl Med. 2021;13:eaaz3088.
53. LaPlante EL, Stürchler A, Fullem R, Chen D, Starner AC, Esquivel E, et al. exRNA-eCLIP intersection analysis reveals a map of extracellular RNA binding proteins and associated RNAs across major human biofluids and carriers. Cell Genomics. 2023;3: 100303.
54. Bourdon S, Herviou P, Dumas L, Destefanis E, Zen A, Cammas A, et al. QUADRatlas: the RNA G-quadruplex and RG4-binding proteins database. Nucleic Acids Res. 2022;51:240–7.
55. Belmonte-Reche E, Morales JC. G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. NAR Genomics Bioinforma. 2020;2:lqz005.
56. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017;8:1324.
57. Nazarova I, Slyusarenko M, Sidina E, Nikiforova N, Semiglazov V, Semiglazova T, et al. Evaluation of colon-specific plasma nanovesicles as new markers of colorectal cancer. Cancers. 2021;13:3905.
58. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature. 2018;563:579–83.
59. Keller A, Gröger L, Tschernig T, Solomon J, Laham O, Schaum N, et al. miRNATissueAtlas2: an update to the human miRNA tissue atlas. Nucleic Acids Res. 2022;50:D211–21.
60. Xie J, Lan T, Zheng D-L, Ding L-C, Lu Y-G. CDH4 inhibits ferroptosis in oral squamous cell carcinoma cells. BMC Oral Health. 2023;23:329.
61. Zhang C, Liu X, Jin S, Chen Y, Guo R. Ferroptosis in cancer therapy: a novel approach to reversing drug resistance. Mol Cancer. 2022;21:47.
62. Li B, Yang L, Peng X, Fan Q, Wei S, Yang S, et al. Emerging mechanisms and applications of ferroptosis in the treatment of resistant cancers. Biomed Pharmacother. 2020;130: 110710.
63. Hafner M, Katsantoni M, Köster T, Marks J, Mukherjee J, Staiger D, et al. CLIP and complementary methods. Nat Rev Methods Primer. 2021;1:1–23.
64. Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, et al. Site identification in high-throughput RNA–protein interaction data. Bioinformatics. 2012;28:3013–20.
65. Park S, Ahn SH, Cho ES, Cho YK, Jang E-S, Chi SW. CLIPick: a sensitive peak caller for expression-based deconvolution of HITS-CLIP signals. Nucleic Acids Res. 2018;46:11153–68.
66. Oh D, Strattan JS, Hur JK, Bento J, Urban AE, Song G, et al. CNN-Peaks: ChIP-Seq peak detection pipeline using convolutional neural networks that imitate human visual inspection. Sci Rep. 2020;10:7933.
67. Hentges LD, Sergeant MJ, Cole CB, Downes DJ, Hughes JR, Taylor S. LanceOtron: a deep learning peak caller for genome sequencing experiments. Bioinformatics. 2022;38:4255–63.
68. Bang I, Lee S-M, Park S, Park JY, Nong LK, Gao Y, et al. Deep-learning optimized DEOCSU suite provides an iterable pipeline for accurate ChIP-exo peak calling. Brief Bioinform. 2023;24:bbad024.
69. Shi J. PANDORA-seq expands the repertoire of regulatory small RNAs by overcoming RNA modifications. Nat Cell Biol. 2021;23:30.
70. Hu JF. Quantitative mapping of the cellular small RNA landscape with AQRNA-seq. Nat Biotechnol. 2021;39:18.
71. Shah A, Qian Y, Weyn-Vanhentenryck SM, Zhang C. CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. Bioinformatics. 2017;33:566–7.
72. Hoshino A, Kim HS, Bojmar L, Gyan KE, Cioffi M, Hernandez J, et al. Extracellular vesicle and particle biomarkers define multiple human cancers. Cell. 2020;182:1044-1061.e18.
73. Costa B, Li Calzi M, Castellano M, Blanco V, Cuevasanta E, Litvan I, et al. Nicked tRNAs are stable reservoirs of tRNA halves in cells and biofluids. Proc Natl Acad Sci U S A. 2023;120: e2216330120.
74. Zhao W, Zhang S, Zhu Y, Xi X, Bao P, Ma Z, et al. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Res. 2021;50:D287–94.
75. Chen S, Jin Y, Wang S, Xing S, Wu Y, Tao Y, et al. Cancer type classification using plasma cell-free RNAs derived from human and microbes. Lo YD, Zhong C, editors. eLife. 2022;11:e75181.
76. Wang H, Zhan Q, Ning M, Guo H, Wang Q, Zhao J, et al. Depletion-assisted multiplexed cell-free RNA sequencing reveals distinct human and microbial signatures in plasma versus extracellular vesicles. Clin Transl Med. 2024;14: e1760.
77. Khan A, Mathelier A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. BMC Bioinformatics. 2017;18:287.
78. Ibarra A, Zhuang J, Zhao Y, Salathia NS, Huang V, Acosta AD, et al. Non-invasive characterization of human bone marrow stimulation and reconstitution by cell-free messenger RNA sequencing. Nat Commun. 2020;11:400.
79. Vorperian SK, Moufarrej MN, Quake SR. Cell types of origin of the cell-free transcriptome. Nat Biotechnol. 2022;40:855–61.

80. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
81. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
82. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. The Innovation. 2021;2:100141.
83. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47:D155–62.
84. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023;2:e107.
85. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.
86. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
87. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res. 2008;36:D173–7.
88. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–73.
89. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47:199–208.
90. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.
91. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinforma Oxf Engl. 2014;30:1006–7.
92. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. 2017;27:491–9.
93. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.
94. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl. 2010;26:841–2.
95. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinforma Oxf Engl. 2012;28:593–4.
96. Langenberger D, Bermudez-Santana C, Hertel J, Hoffmann S, Khaitovich P, Stadler PF. Evidence for human microRNA-offset RNAs in small RNA sequencing data. Bioinformatics. 2009;25:2298–301.
97. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol. 2008;26:407–15.
98. Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. Science. 1999;283:1168–71.
99. Schattner P, Decatur WA, Davis CA, Ares M, Fournier MJ, Lowe TM. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. Nucleic Acids Res. 2004;32:4281–96.
100. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49:9077–96.
101. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA package 2.0. Algorithms Mol Biol. 2011;6:26.
102. Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, et al. RNA duplex map in living cells reveals higher-order transcriptome structure. Cell. 2016;165:1267–79.
103. Gendron P, Lemieux S, Major F. Quantitative analysis of nucleic acid three-dimensional structures. J Mol Biol. 2001;308:919–36.
104. Darty K, Denise A, Ponty Y. VARNA: interactive drawing and editing of the RNA secondary structure. Bioinforma Oxf Engl. 2009;25:1974–5.
105. Tsybulskyi V, Mounir M, Meyer IM. R-chie: a web server and R package for visualizing cis and trans RNA-RNA, RNA-DNA and DNA-DNA interactions. Nucleic Acids Res. 2020;48: e105.
106. McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, et al. The biochemical basis of microRNA targeting efficacy. Science. 2019;366:eaav1741.
107. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28–36.
108. Bailey TL, Grant CE. SEA: Simple Enrichment Analysis of motifs. bioRxiv; 2021. p. 2021.08.23.457422. Available from: https://www.biorxiv.org/content/https://doi.org/10.1101/2021.08.23.457422v1
109. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.
110. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38:675–8.
111. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.
112. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013;14:7.
113. Pengfei B, Taiwei W. cfPeak. Github. https://github.com/lulab/cfPeak (2023).
114. Pengfei B, Taiwei W. cfPeak-archive. ZENODO. https://zenodo.org/records/14773110 (2023).
115. Yuan T, Huang X, Woodcock M, Du M, Dittmar R, Wang Y, et al. Plasma extracellular RNA profiles in healthy and cancer patients. Sci Rep. 2016;6:19413.
116. Elias KM, Fendler W, Stawiski K, Fiascone SJ, Vitonis AF, Berkowitz RS, et al. Diagnostic potential for a serum miRNA neural network for detection of ovarian cancer. eLife. 2017;6:e28932.

117. Roberts BS, Hardigan AA, Moore DE, Ramaker RC, Jones AL, Fitz-Gerald MB, et al. Discovery and validation of circulating biomarkers of colorectal adenoma by high-depth small RNA sequencing. Clin Cancer Res Off J Am Assoc Cancer Res. 2018;24:2092–9.
118. Giraldez MD, Spengler RM, Etheridge A, Godoy PM, Barczak AJ, Srinivasan S, et al. Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. Nat Biotechnol. 2018;36:746–57.
119. Min L, Zhu S, Chen L, Liu X, Wei R, Zhao L, et al. Evaluation of circulating small extracellular vesicles derived miRNAs as biomarkers of early colon cancer: a comparison with plasma total miRNAs. J Extracell Vesicles. 2019;8:1643670.
120. Park IJ, Yu YS, Mustafa B, Park JY, Seo YB, Kim G-D, et al. A nine-gene signature for predicting the response to preoperative chemoradiotherapy in patients with locally advanced rectal cancer. Cancers. 2020;12:800.
121. Godoy PM, Bhakta NR, Barczak AJ, Cakmak H, Fisher S, MacKenzie TC, et al. Large differences in small RNA composition between human biofluids. Cell Rep. 2018;25:1346–58.
122. Selth LA, Roberts MJ, Chow CWK, Marshall VR, Doi SAR, Vincent AD, et al. Human seminal fluid as a source of prostate cancer-specific microRNA biomarkers. Endocr Relat Cancer. 2014;21:L17-21.
123. Liu Z, Wang T, Yang X, Zhou Q, Zhu S, Zeng J, et al. Polyadenylation ligation-mediated sequencing (PALM-Seq) characterizes cell-free coding and non-coding RNAs in human biofluids. Clin Transl Med. 2022;12: e987.
124. Pengfei Bao, Shaozhen X, Qing Z. Peak analysis of small cell-free RNA in oral cancer plasma finds recurrently protected narrow regions with clinical potential. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE238204 (2023).
125. Pengfei B, Qing Z, Shaozhen X, Chun N, Mengtao L, Zhi John L. EV-sorting small RNA in normal human plasma-EV compared with EV-depleted plasma. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE278414 (2024).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.