A comprehensive evaluation of full-spectrum cell-free RNAs highlights cell-free RNA fragments for early-stage hepatocellular carcinoma detection

Chun Ning,[a,b,h] Peng Cai,[c,h] Xiaofan Liu,[b,h] Guangtao Li,[d,h] Pengfei Bao,[b] Lu Yan,[b] Meng Ning,[e] Kaichen Tang,[a,b] Yi Luo,[d] Hua Guo,[d] Yunjiu Wang,[f] Zhuoran Wang,[g] Lu Chen,[d,***] Zhi John Lu,[b,**] and Jianhua Yin[c,*]

[a]Chinese Academy of Medical Sciences & Peking Union Medical College, No. 9 Dongdansantiao, Beijing, 100730, China
[b]MOE Key Laboratory of Bioinformatics, Centre for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing, 100084, China
[c]Department of Epidemiology, Naval Medical University, Key Laboratory of Biosafety Defense, Ministry of Education, Shanghai, 200433, China
[d]Department of Hepatobiliary Cancer, Liver Cancer Research Centre, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Centre for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Centre for Cancer, Tianjin, 300060, China
[e]Tianjin Third Central Hospital, 83 Jintang Road, Hedong District, Tianjin, 300170, China
[f]Department of Clinical Laboratory, Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, 200433, China
[g]Department of Surgery, Eastern Hepatobiliary Surgery Hospital, Navy Medical University, Shanghai, 200433, China

eBioMedicine
2023;93: 104645
Published Online 12 June 2023
https://doi.org/10.1016/j.ebiom.2023.104645

Summary

Background Various studies have reported cell-free RNAs (cfRNAs) as noninvasive biomarkers for detecting hepatocellular carcinoma (HCC). However, they have not been independently validated, and some results are contradictory. We provided a comprehensive evaluation of various types of cfRNA biomarkers and a full mining of the biomarker potential of new features of cfRNA.

Methods We first systematically reviewed reported cfRNA biomarkers and calculated dysregulated post-transcriptional events and cfRNA fragments. In 3 independent multicentre cohorts, we further selected 6 cfRNAs using RT-qPCR, built a panel called HCCMDP with AFP using machine learning, and internally and externally validated HCCMDP's performance.

Findings We identified 23 cfRNA biomarker candidates from a systematic review and analysis of 5 cfRNA-seq datasets. Notably, we defined the *cfRNA domain* to describe cfRNA fragments systematically. In the verification cohort (n = 183), cfRNA fragments were more likely to be verified, while circRNA and chimeric RNA candidates were neither abundant nor stable as qPCR-based biomarkers. In the algorithm development cohort (n = 287), we build and test the panel HCCMDP with 6 cfRNA markers and AFP. In the independent validation cohort (n = 171), HCCMDP can distinguish HCC patients from control groups (all: AUC = 0.925; CHB: AUC = 0.909; LC: AUC = 0.916), and performs well in distinguishing early-stage HCC patients (all: AUC = 0.936; CHB: AUC = 0.917; LC: AUC = 0.928).

Interpretation This study comprehensively evaluated full-spectrum cfRNA biomarker types for HCC detection, highlighted the cfRNA fragment as a promising biomarker type in HCC detection, and provided a panel HCCMDP.

Funding National Natural Science Foundation of China, and The National Key Basic Research Program (973 program).

*Corresponding author. Department of Epidemiology, Naval Medical University, Key Laboratory of Biosafety Defense, Ministry of Education, Xiangyin Road 800, Shanghai, 200433, China.
**Corresponding author. MOE Key Laboratory of Bioinformatics, Centre for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Shuangqing Road 30, Beijing, 100084, China.
***Corresponding author. Department of Hepatobiliary Cancer, Liver Cancer Research Centre, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Centre for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Centre for Cancer, Huanhuxi Road, Tianjin, 300060, China.
E-mail addresses: hawkyjh163@163.com (J. Yin), zhilu@tsinghua.edu.cn (Z.J. Lu), chenlu@tmu.edu.cn (L. Chen).
[h]These authors contributed equally.

Keywords: Liquid biopsy; Liver neoplasms; cfRNA biomarker; Early cancer detection

Research in context

Evidence before this study

Globally, liver cancer is the third major cause of cancer-related deaths. Current noninvasive tools for early-stage HCC detection like AFP are suboptimal, and new liquid biopsy markers like cell-free RNA (cfRNA) are emerging. Despite multiple studies having reported various cfRNA as biomarkers for HCC detection, the reproducibility and stability of reported biomarkers need to be proved in a unified way. Furthermore, adding new cfRNA features may improve the sensitivity of HCC detection.

Added value of this study

Our comprehensive evaluation of full-spectrum cell-free RNAs filled the gap created by previous cfRNA biomarker studies, most of which were biased to miRNAs. Furthermore, we revealed the cfRNA fragment as a promising biomarker type, which extended the dimension of cfRNA types for cfRNA panel development in future clinical trials.

Implications of all the available evidence

The panel HCCMDP, outperformed existing biomarkers like AFP in our phase I–II biomarker study, and has a promising prospect to be a noninvasive and economic HCC detection biomarker in future clinical practice. This meets the urgent demand of identifying early-stage HCC patients who can receive curative treatments. HCCMDP can fit into broad clinical application scenarios because of its economic features and capability to rapidly give results, especially when imaging infrastructure for HCC diagnosis is unavailable.

Introduction

Liver cancer is the third major cause of cancer-related deaths worldwide.[1] Hepatocellular Carcinoma (HCC) is the main category of primary liver cancer (accounts for 75%–85%). HCC patients diagnosed at early stages have significantly prolonged survival time,[2] making early detection crucial. Early detection of HCC is exceptionally urgent for the high-risk population, including chronic hepatitis B (CHB) and liver cirrhosis (LC) patients, who have annual HCC incidence rates of 0.3–0.6% and 2.2–3.7%, respectively.[3] Nowadays, noninvasive HCC screening methods mainly include ultrasound and blood tests.[2] However, they are suboptimal, which brings the need for better-performed biomarkers.

With improvements intechnology, many types of biomarkers in the bloodstream, including circulating tumour cells (CTCs), circulating tumour DNAs (ctDNAs), proteins, metabolites, tumour-educated platelets, and cell-free RNAs (cfRNAs)[4] have been shown to have diagnostic capabilities in HCC. Among them, cfRNA is perceived as a promising type because RNA dysregulation, including differential RNA expression and multiple RNA post-transcriptional events, can depict the dynamic changes in patients. Moreover, cfRNA can be detected through cost-effective RT-qPCR, thus offering an economic advantage in extensive clinical applications.

Multiple cfRNA subtypes in blood have been identified as HCC biomarkers. Zhou et al. identified a 7-miRNA panel in plasma[5] with an area under the receiver operating characteristic curve (AUC) of 0.888; long noncoding RNAs (lncRNAs) like HULC,[6] srpRNA,[7] and snoRNA,[8] and circRNA[8,9] were also reported to be biomarkers in HCC. However, the results are sometimes contradictory in different studies due to the variation in sample sizes, experimental methods, data analysis, and the risk of overfitting.[10–12] Therefore, the reproducibility and stability of reported biomarkers need to be demonstrated in a unified way.

Additionally, adding new RNA biomarker subtypes to a traditional cfRNA panel may complement and improve the panel's performance. Dysregulated RNA post-transcriptional events, including alternative splicing,[13] alternative polyadenylation,[14] and chimeric RNAs,[15] contribute to the complexity of the RNA landscape. They all play significant roles in cancer and have been mostly reported to be tissue-based diagnostic biomarkers in other cancers. cfRNA fragments have been identified in blood by different RNA library construction strategies. These fragments are derived from long RNAs like tRNA, srpRNA,[7] mRNA, and lncRNA.[16,17] The biomarker potential of cfRNA fragments has just started to be acknowledged, tRNA-derived small RNA (tsRNA) being a good example.[18] Besides, a recent study reported that exosomal RNA fragments in unannotated regions of the genome could well discriminate HCC from controls.[19] However, most of these new RNA-related features mentioned above are based on various NGS library construction strategies and analysis methods. Their biomarker potential needs a unified way.

In this work, we selected and verified 23 candidate biomarkers covering a full spectrum of cfRNA types, dysregulated RNA post-transcriptional events, and cfRNA fragments through a comprehensive analysis of literature and NGS data. Notably, we used a biomarker

feature named *cfRNA domain* to uniformly identify cfRNA fragments from traditional small cfRNA-seq datasets. We then selected, verified and validated these cfRNA candidates in 3 independent cohorts and revealed a low-cost and effective panel for the non-invasive detection of HCC.

Methods

Systematic review for cfRNA biomarkers

First, we searched Pubmed, Scopus, Embase, and the Cochrane library for articles reporting RNA biomarkers in plasma or serum or exosomes for HCC diagnosis according to the search terms listed below. RNA-related reports were published between August 1992 and March 7, 2023. Search terms are listed in Supplementary File S1.

Studies were excluded for the reasons listed below: 1) The study is not focused on identifying RNA as biomarkers (e.g., The study is about RNA regulation or RNA-encoded protein's function, e.g.[20]; 2) The samples used are tissue or PBMC or whole blood, rather than plasma or exosomes or serum, e.g.[21]; 3) The RNA is not used for diagnosis, but prognosis, e.g.,[22] etc. 4) The study has no control group, or the cohort has less than 10 people; 5) The study provides incomplete information about the reported RNA and their exact sequences cannot be accessed (after request to the authors); 6) The study type is a review, a comment, a systematic review or a meta-analysis. The studies' characteristics were aggregated by Endnote 20 and compared against the inclusion and exclusion criteria.

The whole procedure was conducted under PRISMA 2020's guidelines[23] and was independently checked by 2 people (K.T., C.N.). Assessment of quality and risk of bias was done using modified QUADAS2,[24] and only reports having both training and validation cohorts were evaluated as high quality. Basic information about the biomarkers and statistics that reflect the biomarkers' ability (i.e., AUC, sensitivity, specificity, expression trend, number of participants in each cohort and group, information about the RNA biomarker and specimen) were collected from all included reports independently by two people (K.T., C.N.). All results that were compatible with each outcome domain in each study were sought. Missing or unclear information was marked as "NA". Study investigators were contacted for confirming obscure information. Statistics such as sum, median, max and min were used to present the systematic review results.

Meta-analysis for miRNA biomarkers

First, the meta-analysis only included studies that reported the numbers of true positive, true negative, false positive and false negative of the biomarker alone, or other indexes that can be transformed to them rather than only reporting these indexes of the whole panel. We first checked the values of summary statistic (log (diagnostic odds ratio) and Higgins' $\mathbf{I^2}$) by performing univariate analysis.[25] Considering that the data heterogeneity in the accuracy of diagnostic tests is more pronounced, bivariate normal random effect model[26] was used to plot the summary ROC curve and calculate the Summary AUC (sAUC) and pooled sensitivity, false positive rate with their corresponding 95% confidence intervals (95% CI). The model's core idea is that it assumes the logit of the observed sensitivity ($\widehat{s}_i$) and false positive rate ($\widehat{f}_i$) of the i^{th} report follow a bivariate normal distribution,

$$\left(\text{logit}\left(\widehat{s}_i\right), \text{logit}\left(\widehat{f}_i\right)\right)^{\top} \sim N(\mu, \Sigma)$$

in which μ denotes the actual value of the biomarker's logit sensitivity and false positive rate, and the equation $\sum = \begin{pmatrix} \sigma_s^2 & \sigma_{sf} \\ \sigma_{sf} & \sigma_f^2 \end{pmatrix}$ describes between-study variance and covariance between logit sensitivity and false positive rate. Publication bias was assessed by visual inspection of the funnel plots. Sensitivity analysis was done by leaving each research out. Possible causes of heterogeneity among study results were analysed through subgroup analysis. All analyses were done using the R package "Mada", "Meta" and "Metafor".

Preprocessing of cfRNA-seq data

cfRNA-seq datasets were systematically searched in GEO using search terms, and 5 public datasets (GSE100207, GSE142987, GSE174302, GSE123972, and GSE104251) were downloaded. All cfRNA-seq datasets used are available at Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). The datasets used are further summarised in Supplementary Table S1.

Adaptors and low-quality sequences from raw data were trimmed off by cutadapt[27] version 2.3. In total RNA-seq data, reads shorter than 30 nt were excluded, and GC oligos introduced in reverse transcription were trimmed off. In small cfRNA-seq data, reads shorter than 16 nt were excluded. The remaining reads were sequentially mapped to ERCC's spike-in sequences, NCBI's UniVec sequences, and human rRNA sequences by bowtie2[28] (version 2.3.5, for small RNA libraries) or STAR[29] (version 2.5.3a_modified, for total RNA libraries).

Annotation of cfRNA-seq reads

For total cfRNA-seq data, STAR was used to map clean reads to the hg38 genome built with the GENCODE v27 annotation. STAR was also used to align reads which were unaligned to hg38 to circRNA junctions. Picard Tools MarkDuplicates (version 2.20.0) was used to remove duplicates in the aligned reads. Each aligned read pair is assigned to a specific RNA if at least one of the mates overlapped with the region corresponding to

this RNA. The aligned reads were sequentially assigned to spikein_long, univec, rRNA, lncRNA, mRNA, snoRNA, snRNA, srpRNA, tRNA, TUCP, Y RNA, and circRNA junctions using bedtools[30] according to the GENCODE v27 annotation.

For small cfRNA-seq data, bowtie2 was used to sequentially map clean reads to the longest isoform of each gene with the following sequence: spikein_small, univec, rRNA, miRNA, lncRNA, mRNA, piRNA, snoRNA, snRNA, srpRNA, tRNA, tucpRNA, and Y_RNA.

Annotation of human transcripts with uncertain coding potential (TUCP) was extracted from mitranscriptome.[31] Enhancers, promoters, and repeat regions were downloaded from the UCSC Genome Browser. Circular RNA annotation was downloaded from circBase.[32] Junction sequences were prepared by concatenating upstream 150 nt and downstream 150 nt sequences around the back spliced sites of circRNAs.

Differential expression analysis and statistics calculation in the analyses of cfRNA-seq data

EdgeR[33] was used to calculate differential expression and produce statistics, including log fold change ($|\log_2 FC|$) and FDR. Other calculated statistics for a specific gene included AUC, ratio (ratio_HD, ratio_HCC), TMM_mean, and Gini index[34] (Gini_HD and Gini_HCC). Firstly, to calculate AUC, the min-max method was used to normalise the features to 0–1. Then, the tendency of features was judged. If the mean of the feature in the cancer patients was greater than the mean in the healthy donors, it stayed the same. If it was the opposite trend, we took 1 to minus the feature. Finally, we calculated the AUC of the feature. The ratio is the non-zero proportion of a feature in HCC and healthy samples. TMM_mean is the average normalised expression value of each gene (or feature) in HCC and healthy samples, and the normalised method is TMM. The Gini index (ranging from zero to one) is a single value that measures data heterogeneity.[34] The higher the Gini index, the more heterogeneous the data is. We calculated the Gini index of each gene (or feature) for different cancer and healthy samples.

mRNA candidates were selected using datasets of GSE100207, GSE142987 and GSE174302. The cut-off was: (FDR<0.05) & ($|\log_2 FC| > 1$) & (AUC > 0.8) & (ratio_HD > 0.9|ratio_HCC > 0.9) & (TMM_mean_HD > 8|TMM_mean_HCC > 8) & (Gini_HD < 0.7|Gini_HCC < 0.7). (FDR < 0.05) & (|log2FC| > 1) means the RNA is significantly differentially expressed in HCC and HD. ratio >0.9 and TMM_mean >8 are used to avoid low abundance of gene expression. Gini <0.7 is used to avoid heterogeneity in the gene's expression in a population. The cut-off was formulated according to the distribution of these indexes in this RNA subtype because we want to pick out the best-performing candidates in this RNA subtype while keeping the number of candidates in each subtype balanced and the total number appropriate for qPCR validation. The distribution of indexes in the selection cutoff of mRNA is listed in Supplemental Figure S6 as an example. The cut-offs below were set in the same way.

Identification of cfRNA domains from long transcripts

We used a local maximum-based peak calling method to identify cfRNA domains using the small cfRNA-seq data (GSE123972 and GSE104251). The local maximum-based peak calling method is divided into three steps: determining effective bins, searching local maximum peaks and merging, and filtering confident peaks. In the first step, we divided a transcript into 20-nt bins. We then calculated the average read coverage over each bin and filtered out bins with average read coverage below 5 (minimal read coverage). Bins with average read coverage above the minimal read coverage were effective bins. In the second step, we firstly searched for the local maximum beginning at the start position of the first effective bin. Once a local maximum was found, peaks were extended from the position of the local maximum in both directions until the read coverage of a position dropped below the minimal read coverage or half the local maximum. This peak finding procedure was performed from 5′ to 3′ of transcripts to identify all peak candidates. Peaks shorter than 10-nt were removed, and adjacent peaks were merged. We performed steps 1 and 2 on each sample to find all peak candidates. In the third step, we filtered and adjusted these local maximum peaks using the recurrence of peaks among samples. Peak confidence is the minimum of the peak sub-region recurrence ratio. The peak sub-region recurrence ratio refers to the frequency of this region in all sample peaks. We filtered peaks with peak confidence below 10% and merged the adjacent confidence regions (region recurrence ratio above 10%) of these peaks to obtain the confident peaks. Finally, we define these confident peaks larger than 10-nt as cfRNA domains.

Differential expression analysis was done, and the same statistics were calculated as described in the upper part "Differential expression analysis and statistics calculation in the analyses of cfRNA-seq data". The cut-off for cfRNA domain candidates was: (FDR < 0.05) & ($|\log_2 FC| > 1$) & (AUC > 0.8) & (ratio_HD > 0.9|ratio_HCC > 0.9) & (TMM_mean_HD > 10|TMM_mean_HCC > 10) & (Gini_HD < 0.3|Gini_HCC < 0.3).

The secondary structure and RBP enrichment analysis of the cfRNA domain

We downloaded icSHAPE data from GSE74353,[35] and the in vivo matrix was chosen. Only mRNA-related transcripts shorter than 10 kb were used for simplicity. We averaged trimmed icSHAPE reactivity

values for each confident fragment along its peak. Background regions were generated by shuffling confident fragments within all mRNA transcripts in the filtered icSHAPE matrix, from which transcripts' lengths and total number were not changed. RBP binding sites were downloaded from the POSTAR3 human RBP database,[36] and only those with a length longer than 20 nt and autosomal location were selected. Background regions were generated by shuffling confident cfRNA fragments within the union transcript set from 8 RNA species in a manner similar to structure analysis. Known RBP sites in these domains are predicted using FIMO(v5.4.1),[37] with *P*-value <0.01 as the threshold using Ray2013_rbp_Homo_sapiens.meme RBP database.

Alternative splicing analysis

RMats[38] was used in identifying RNA alternative splicing events with the parameter setting of –cstat 0.0001 –libType fr-secondstrand. Reads were aligned to the "alternative-spliced regions" and the upstream and downstream "reference regions" generated by rMats, and normalised to TMM. Ratios of "alternative-spliced regions" divided by upstream or downstream "reference regions" were then calculated. Differential expression was calculated based on the ratio, and the cut-off for this step was:

(FDR < 0.05) & ($|\log_2 FC|>1$) & (AUC>0.8) & (ratio_HD > 0.9|ratio_HCC > 0.9) & (Gini_HD < 0.3|Gini_HCC < 0.3).

Alternative polyadenylation analysis

DaPars[39] was used to identify RNA alternative polyadenylation events with parameter setting of: FDR_cutoff = 0.05, PDUI_cut-off = 0.2, Fold_change_cutoff = 0.59 (PDUI: percentage of distal polyA site usage statistics). PDUI was used to conduct differential expression and calculate the statistics described above. Overall, the differential statistics for alternative polyadenylation candidates are lower than in other categories. To include this type as well, we adopted looser cut-offs (AUC > 0.65). The cut-off was: (FDR < 0.05) & ($|\log_2 FC| > 1$) & (AUC > 0.65) & (ratio_HD > 0.3|ratio_HCC > 0.3).

Chimeric RNA analysis

Reads that did not map to genome sequence and circRNAs were mapped to chimeric RNAs derived from ChimerDB 3.0,[40] GTEx,[41] and annotations generated from GSE142987 and GSE174302. The identified chimeric RNA's expression matrix went through differential expression analysis. The statistics were calculated as described in "Differential expression analysis and statistics calculation in the analyses of cfRNA-seq data". The cut-off was: (FDR < 0.05) & ($|\log_2 FC| > 1$) & (AUC > 0.8) & (ratio_HD > 0.5|ratio_HCC > 0.5).

False positives were removed by manual inspection on IGV, including mitochondrial genes or human HLA genes as one of the partners, those having incomplete poly-A trimming, and those having many reads that were not mapped uniquely across the fusion breakpoints.

Study cohorts

We retrospectively collected 3 independent cohorts: (1) a verification cohort containing 183 samples (83HCC, 12LC, 21CHB, and 67 HD; allocated about 30 HCC, 10 CHB, 10 LC, and 30 HD for each candidate) for candidate selection; (2) an algorithm development cohort containing 287 samples (109 HCC, 58 CHB, 54 LC, and 66 HD) for model training and bootstrap validation; (3) an independent validation cohort containing 171 samples (66 HCC, 40 CHB, 36 LC, and 29 HD) for external validation and reproducibility assessment. Samples were collected from Second Military Medical University, Tianjin Medical University Cancer Institute and Hospital, and Beijing Ditan Hospital from 2019 to 2023. Potentially eligible participants were identified by results from previous tests and medical records. Participants formed a consecutive series. Samples of HCC patients were collected before treatments. There are no adverse events from sampling. Clinical information is in Supplementary Tables S2, S3 and S4.

The inclusion and exclusion criteria for healthy donors were as follows: (1) No history of hepatitis, liver cancer, and family history of liver diseases; (2) Blood tests showed negative HBV DNA and normal levels of ALT, AST, serum HBV markers, and tumour biomarkers such as AFP and CEA; (3) B-ultrasound examination showed no intrahepatic space-occupying lesions; (4) No other malignant tumours, immune diseases, and other major diseases. HCC was diagnosed by pathological analysis or α-fetoprotein combined with computed tomography or magnetic resonance imaging. LC was diagnosed by liver biopsy or ultrasonography, supplemented with clinical complications indicating portal hypertension, and free of HCC according to medical records in at least the past 6 months. CHB was diagnosed according to seropositivity for HBsAg ≥6 months, apparent inflammatory activity, and absence of cirrhosis or HCC. Different sexes had equal possibility to be potentially eligible. Sexes were self-reported by study participants.

Ethics

The study protocol conformed to the 1975 Declaration of Helsinki and was approved by the ethics committees of Second Military Medical University, Tianjin Medical University Cancer Institute and Hospital, and Beijing Ditan Hospital (approval number: No.2019OE-023). All patients included in this study voluntarily signed informed consent.

Plasma preparation

Blood samples were processed within 2 h of collection and avoided enzyme pollution during processing. 1–5 mL of whole blood were collected through peripheral venipuncture in K2EDTA tubes and were gently reversed 8–10 times. Centrifugation was performed at 800 g at 4 °C for 10 min. The supernatant was transferred to a new tube and centrifuged again at 4 °C for 10 min at 3,000 g. The plasma was transferred into 0.5 mL/tube and stored in a –80 °C refrigerator.

RNA isolation and quantitative reverse transcription PCR for long and short RNAs

RNA was extracted from plasma using QIAzol®Lysis Reagent (QIAGEN, Cat. No. 79306) and ethanol precipitation method. Clinical information and the sample category were unavailable to the performers before finishing the experiment.

For long RNAs, RNA extracted was reverse transcribed into cDNA with TIANScript II First-Strand cDNA Synthesis Kit (TIANGEN, KR107-02) according to the manufacturer's instructions. For real-time quantitative PCR, the reactions were performed in a 20-μL final volume system with 2 × FastFire qPCR premix (SYBR, Lot#U9116) according to the manufacturer's instructions. The qPCR reaction procedure was set as follows: 95 °C for 1 min, then replicate 40 cycles of 95 °C for 5 s and 60 °C for 15 s, and then 65 °C for 5 s, heat 0.5 °C per cycle to 95 °C. All reactions were replicated two times and yielded the mean Cq value. Expression levels of *GAPDH* were detected as endogenous control measurements. Two no-template controls were set at reverse transcription (RT-NTC) and RT-qPCR (qPCR-NTC) to monitor these 2 processes. To control the intra-assay variability and inter-assay variability, we set quality control standard for the whole process as: 13 < Cq (*GAPDH*) < 36 & 13 < Cq (candidates) < 37 & Cq (RT-NTC) > 37 & Cq (qPCR-NTC) > 37 & numerical difference of Cq (candidate) between replicates <1. Raw data that did not meet this standard were considered to have problems in the experiment and discarded.

For short RNAs, RNA was reverse transcribed into cDNA with miRcute Plus miRNA First-Strand cDNA Synthesis Kit (TIANGEN, KR211). The reverse transcription reaction conditions were as follows: 42 °C for 60 min, followed by 95 °C for 3 min and then 4 °C for infinite hold. The miRcute Plus miRNA qPCR Detection Kit (TIANGEN, FP411) was used for real-time quantitative PCR. The qPCR reaction procedure was set as follows: 95 °C for 15 min, followed by 40 cycles of 94 °C for 20 s and 60 °C for 34 s, and then 65 °C for 5 s, heat 0.5 °C per cycle to 95 °C. All reactions were replicated two times and yielded the mean Cq value. The expression level of *miR-16* was detected as endogenous control. Two no-template controls were set at reverse transcription (RT-NTC) and RT-qPCR (qPCR-NTC) to monitor these 2 processes. To control the intra-assay variability and inter-assay variability, we set quality control standard for the whole process as: 13 < Cq (*miR-16*) < 32 & Cq (RT-NTC) > 37 & Cq (qPCR-NTC) > 37 & numerical difference of Cq (candidate) between replicates <1.

The expression levels of all candidates were normalised to those of the reference genes using the –ΔCq ($-(Cq_{candidate} - Cq_{reference})$) method. In the verification stage, at about 15 HCC vs. 15 HD, RNAs that did not show significance (Wilcoxon Rank-Sum test, *P*-value <0.2) between HCC and HD or whose Cq values fell out of 37 too often were pre-excluded.

Machine learning method for combining multiple RNAs and AFP as a panel

Random Forest was applied to the 6 cfRNAs and AFP to build the panel HCCMDP. 6 HCC samples in the algorithm development cohort were excluded due to experimental technical failure. The function we used for the random forest is sklearn.ensemble.RandomForestClassifier (maximum tree depth = 500, tree number = 1000). Missing data were treated as n.a. The model built from the algorithm development cohort was tested by 500-times bootstrap, and the distributions of AUC are presented. Further, the model was externally validated by an independent validation cohort. We did not use the optimism-corrected AUC because our training and validation data did not overlap. We calculated the Brier score, Hosmer–Lemeshow test, net reclassification indexes and the misclassification table to evaluate the model.

To evaluate cfRNA fragments, we trained another random forest model, only including three cfRNA fragments and AFP on the algorithm development cohort, and tested it on the independent validation cohort. The model probability cut-offs of HCCMDP (0.53), AFP (0.51) and 3 cfRNA fragment combination (0.50) was determined on the algorithm development cohort requiring the specificity of discriminating HCC from all controls ≥95%. 400 ng/mL was used as the cut-off when deciding AFP-negative patients.

Statistical analysis

Statistical significance for continuous variables was calculated by the Wilcoxon Rank Sum test except for differential expression analysis of NGS data, which used the exact test based on a negative binomial model in edgeR. Fisher's exact test was used for categorical variables. Unless otherwise mentioned, a *P*-value or false discovery rate (FDR) < 0.05 were considered statistically significant. Receiver operating characteristic (ROC) analysis was done by the python package sklearn.metrics. 95% confidence interval (CI) was calculated for AUC, sensitivity and specificity by bootstrap estimation in the R package pROC. We used all samples available at that time in the verification cohort and the algorithm

development cohort. Based on the results of the algorithm development cohort, the sample size of the independent validation cohort was estimated by MedCalc (Type II error = 0.05, Type I error = 0.01, hypothesised Area under the ROC curve = 0.8, ratio of sample sizes in negative/positive groups = 0.5). The minimal sample size is 39 for HCC and 20 for each negative control group. In boxplots, whiskers extend to the furthest data point within the range that is 1.5 times of the interquartile range. More extreme points are marked as outliers.

Other statistical details were described in relevant sections of Method.

Role of the funding source

The sponsors did not have any role in the study design, data collection, data analyses, interpretation, or writing of the manuscript.

Results

Overview of the analysis and validation of full-spectrum cell-free RNA biomarkers for HCC detection

We evaluated and validated full-spectrum cfRNA biomarkers in 4 stages: discovery, verification, algorithm development and independent validation stages. In the discovery stage, we systematically reviewed published cfRNA biomarkers and further meta-analysed miRNA candidates with homogeneous reports. We selected 14 published candidates based on reported statistics reflecting reproducibility (the number of independent studies), credibility (sample size), and diagnostic accuracy (reported AUC and specificity) (Fig. 1a).

Meanwhile, we also analysed small and total cfRNA-seq datasets from 3 perspectives: mRNA, cfRNA fragments, and dysregulated post-transcriptional events (alternative splicing (AS), alternative polyadenylation (APA), and chimeric RNAs). To fill the gaps in those biomarker types that previous studies have rarely reported, we mainly focused on selecting candidates of RNA biomarker features that were not picked out in the systematic review. We finally selected 9 markers based on calculated statistics reflecting abundance, differential expression, classification performance, heterogeneity (Gini-index[34] and non-zero ratio), and further screen conditions (Fig. 1b).

Together, the above 23 candidates were further verified, selected, and validated by RT-qPCR in multicentre cohorts during the verification, algorithm development and independent validation stages (Fig. 1c). Detailed results are described in the following sections.

Candidate cfRNA biomarkers selected from the systematic literature review for HCC detection

In the first part of the discovery stage, we systematically reviewed the literature on cfRNA biomarkers. A total of 3008 records were found by applying searching strategies in Pubmed, Embase, and the Cochrane library. Finally, 130 articles that contain 301 unique biomarkers on a total of 28,150 samples were included in the review (Fig. 2a).

To cover all reported cfRNA types in the selection and consider each type's research status, we set flexible standards for their reported statistics (Fig. 1a, Supplementary Tables S5 and S6). Furthermore, we meta-analysed six selected miRNAs and further picked out four based on meta-analysis statistics (Supplementary Table S7 and Supplementary Figure S1). Meta-analysis was not done on other types of cfRNAs since the statistics of single marker performance were unreported, which would bring nonnegligible heterogeneity when trying to meta-analyse them.

In summary, we selected 14 published candidates and summarised their reported performances (Fig. 2b). The 14 biomarkers include traditional biomarker types like miRNA, which has been researched in multiple studies with many samples. It also includes new types reported on small sample sizes, including lncRNA, circRNA, and cfRNA fragments (1 srpRNA fragment and 1 tsRNA).

Candidate biomarkers derived from total cfRNA-seq datasets of HCC

In the second part of the discovery stage, we comprehensively analysed total and small cfRNA-seq data sequenced from 258 samples to study types of cfRNA that were not picked out in the systematic review. From total cfRNA-seq data, we selected 3 differentially expressed mRNAs (Fig. 1b). We also thoroughly explored the biomarker potential of dysregulated post-transcriptional events in total cfRNA-seq. For alternative splicing and alternative polyadenylation, by applying cut-offs (Fig. 1b), we selected 1 and 2 candidates, respectively. For chimeric RNA analysis, after applying cut-offs on 107 differentially expressed chimeric RNAs, 13 chimeric RNAs were selected and went to primer design. However, no chimeric RNA candidates with specific primers could be verified by RT-qPCR and Sanger's sequencing, and they were aborted.

In total, we selected 3 mRNAs and 3 post-transcriptional events from the total cfRNA-seq data (Supplementary Figure S2).

cfRNA domain systematically describes cfRNA fragments captured in small cfRNA-seq datasets

We used *cfRNA domain* to describe the cfRNA fragments captured by small cfRNA-seq datasets. A cfRNA domain is identified as a small peak (>10 nt) enriched with small RNA reads inside a long (>50 nt) RNA (Fig. 3a–b). cfRNA domains exist in various types of RNAs and are mainly enriched in mRNAs and lncRNAs (Supplementary Figure S3). Remarkably, we found that the identified cfRNA domains enhanced classification

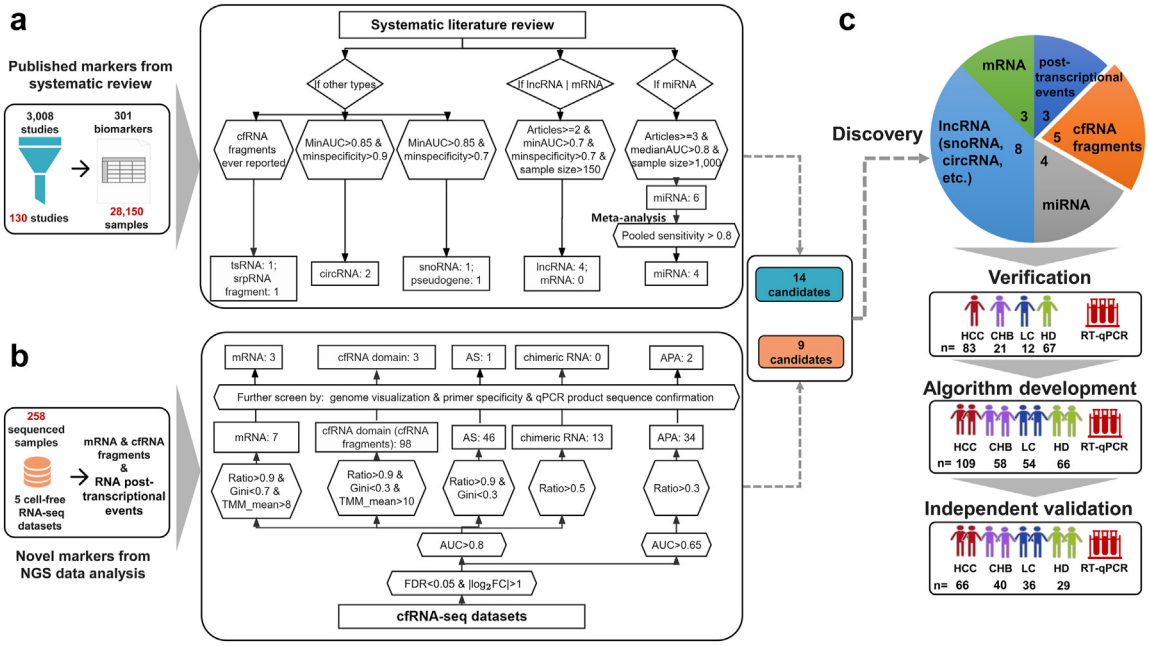


***Fig. 1:* Overview of the analysis and validation of full-spectrum cell-free RNA biomarkers for HCC detection.** Published cfRNA markers were collected from a systematic review and meta-analysis (a), while new markers were identified from a comprehensive NGS analysis of mRNA, cfRNA fragments, and RNA post-transcriptional events (b). Next, further filtering was done based on reported statistics (See Supplementary methods for a detailed explanation for statistics). Selected RNA candidates, summarised in the pie chart, were confirmed by electrophoresis and Sanger sequencing, and were selected, verified and validated in 3 independent cohorts further by RT-qPCR (c). HCC: Hepatocellular Carcinoma patients; CHB: Chronic hepatitis B patients; LC: Liver cirrhosis patients; HD: Healthy donors. Ratio = max (non-zero ratio of HD, non-zero ratio of HCC); Gini = max (Gini index of HD, Gini index of HCC); TMM_mean = max (TMM_mean of HD, TMM_mean of HCC). Detailed explanation of these indexes is in Method.

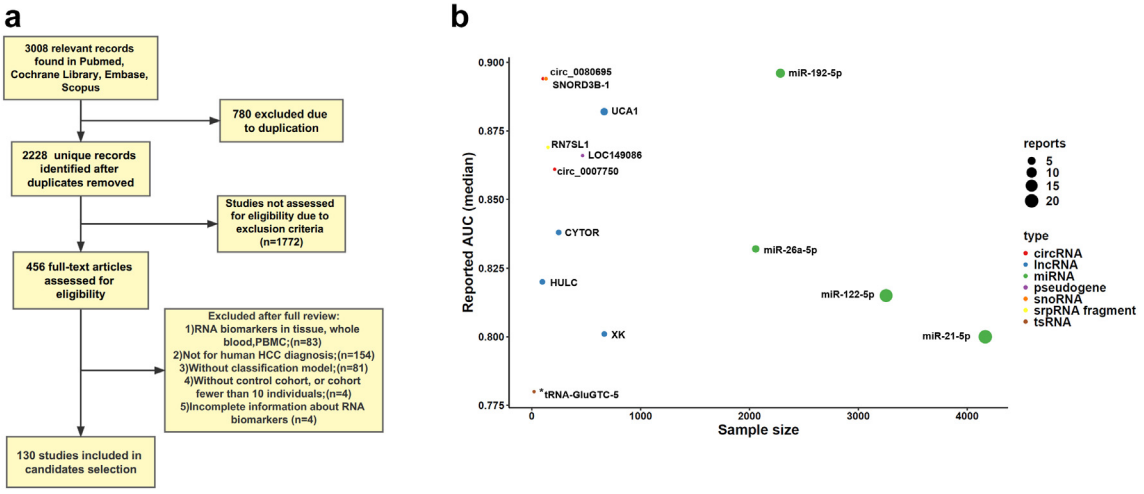
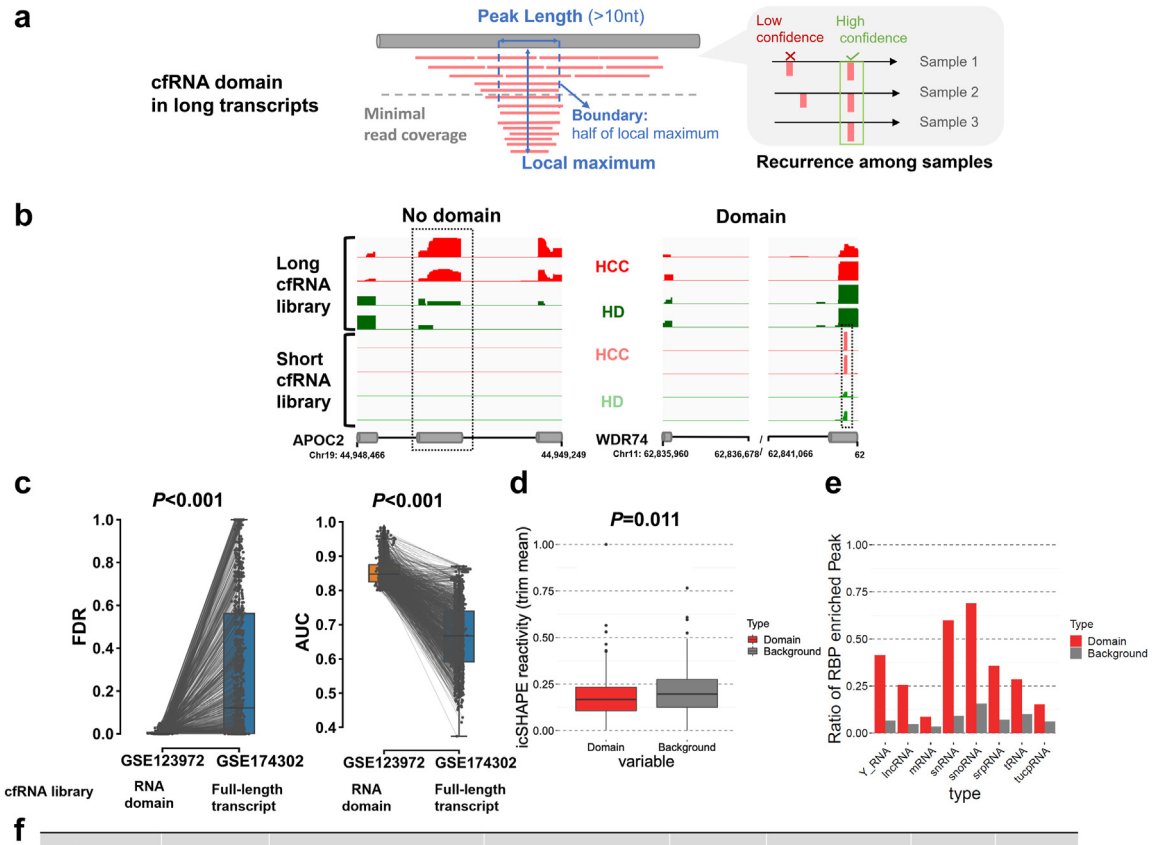


***Fig. 2:* Selected published cfRNA biomarkers from the systematic review.** (a) Flow chart of the systematic review process. (b) Summary information about published cfRNA biomarkers selected after the systematic review and meta-analysis. *: AUC is not reported for *tRNA-GluGTC-5*. The 2 partially overlapping dots (the orange dot's actual position and size are the same as the red dot) mean that *SNORD3B-1* has the same reported AUC (median) and sample size to *circ_0080695*.



f

cfRNA domain candidates	RBP	Structure	Function	Motif	Logo	Length	*P*-value
WDR74	SRSF2	-36.50 kcal/mol	Pre-mRNA splicing factor	AGGAGATG		8	<0.001
	MBNL1		Pre-mRNA alternative splicing regulation	AGCTTGC		7	<0.001
SNORD89	FXR1	-	Associate with polyribosomes	GATGACAA		8	<0.001
	TARDBP		Regulates the splicing of many non-coding RNAs	GGAATGAT		8	<0.001
RNU2-1	HNRNPCL1	-	May play a role in nucleosome assembly	ATTTTTG		7	<0.001
	HNRNPC		Binds pre-mRNA and nucleates the assembly of 40S hnRNP particles	ATTTTTG		7	<0.001

Fig. 3: **The cfRNA domain systematically depicted cfRNA fragments captured by small cfRNA-seq.** (a) The cfRNA domain defined by statistics (i.e., minimal read coverage, peak length, local maximum and its half value, and recurrence among samples). (b) Examples of an exon of a long cfRNA with no cfRNA domain and a cfRNA domain within a long cfRNA visualised by IGV. (c) FDR and AUC for the differentially expressed cfRNA domains and their full-length host transcripts (HCC vs. HD). The expression level was calculated from the small cfRNA-seq data (GSE123972) and the total cfRNA-seq data (GSE174302 & GSE142987) for each cfRNA domain and its full-length transcript, respectively. Wilcoxon rank-sum test was used. (d) Comparison of icSHAPE reactivities between cfRNA domains and shuffled background regions. RNAs with low icSHAPE reactivity are highly structured. Wilcoxon rank-sum test was used. (e) The proportion of cfRNA domain bound by at least one RNA-binding protein (RBP) in each RNA type compared to shuffled background regions. (f) The table summarises the predicted RNA-binding proteins (RBP), the predicted secondary structure and its minimum free energy, the functions of the RBPs, the corresponding RBP-binding motifs, the motif sequence logos, the lengths of the motifs, and the *P*-values of the RBP prediction of three cfRNA domain candidates. The green semicircles in the "Structure" column mean paired bases, and the y axis of the logos are "bits."

ability of their corresponding full-length transcripts (Fig. 3c). This improvement highlights the necessity of analysing cfRNA domains, which digs out the biomarker potential of those long RNAs insignificantly expressed when using the traditional full-length analysis strategy.

We further explored the underlying biological mechanisms of cfRNA domains and found that they tend to be structured and associated with RNA-binding proteins (RBPs). cfRNA domains were significantly more structured than the background (mean estimates with 95% CI in the domain group: 0.18 (0.17, 0.20); the background: 0.21 (0.19, 0.23), Fig. 3d), and the proportion of cfRNA domains bound by at least one RBP was higher than that found in the background for all types of RNAs (Fig. 3e).

Finally, using small cfRNA-seq data, we selected the 3 best-performing cfRNA domains as candidates for further verification. All 3 candidates were predicted to bind proteins, and *WDR74* has a predicted secondary structure (Fig. 3f). These traits may contribute to the stability of the candidates in blood.

cfRNA biomarkers in the verification stage

In an independent verification cohort, we verified the above 23 candidate biomarkers selected from the systematic review and NGS data by RT-qPCR (Fig. 4a). 1 out 4 miRNAs, 3 out of 5 cfRNA fragments, 1 out of 8 lncRNAs, neither the tsRNA nor mRNAs, and 1 alternative splicing event were significantly differentially expressed in HCC and HD (Wilcoxon rank-sum test, *P*-value <0.05. Fig. 4b–c). We also aborted all 2 reported circRNAs and a lncRNA (*XK*) selected from the literature because their Ct values fell out of our quality control range too often, meaning their amount was low (Supplementary Figure S4). Therefore, we deduce that some types of cfRNA biomarkers like circRNA, mRNA, and lncRNA may not be as robust as believed.

Next, we added ~10 chronic hepatitis B (CHB) patients and ~10 cirrhosis (LC) patients for the significant markers. *miR-21-5p and SNORD89* could distinguish HCC from LC; *RN7SL1* and *WDR74* could distinguish HCC from CHB (Fig. 4b–c). Finally, 1 lncRNA (*CYTOR*), 1 miRNA (*miR-21-5p*), 3 cfRNA fragments (*WDR74, SNORD89, RN7SL1*), and 1 alternative splicing candidate (*GGA2*) were selected into the validation stage. Half of the final 6 cfRNA markers belong to cfRNA fragments, indicating the importance of this type. All of these 6 selected markers have been reported to participate in tumour proliferation and metastasis (Supplementary Table S8).

Algorithm development and independent validation revealed HCCMDP as an effective panel

We chose the selected 6 cfRNAs and AFP to train a panel HCCMDP (HCC molecular detection panel) on the algorithm development cohort (n = 287) based on random forest. We used 500-times bootstrap to estimate AUC in order to assess the panel's performance. In the algorithm development, HCCMDP can discriminate all-stage HCC from all control groups at a mean AUC of 0.860, the CHB group at a mean AUC of 0.832, and LC group at a mean AUC of 0.819. The mean AUC values on early-stage HCC are 0.846 for all controls, 0.816 for CHB, and 0.801 for LC (Fig. 5a). HCCMDP demonstrated better performances than AFP both in differentiating HCC from all controls and high-risk populations (Fig. 5a, Supplementary Figure S5a. Brier scores and model calibration are reported in Supplementary Figure S5b and c).

We further externally validated the panel in an independent validation cohort (n = 171). HCCMDP remained good in distinguishing HCC and early-stage HCC from high-risk populations (HCC vs. CHB: AUC = 0.909 (95% CI [0.855, 0.963]); HCC vs. LC AUC = 0.916 (95% CI [0.863, 0.969]); early-stage HCC vs. CHB: AUC = 0.917 (95% CI [0.854, 0.980]); early-stage HCC vs. LC: AUC = 0.928 (95% CI [0.869, 0.987])) (Fig. 5b–c). Compared to AFP's performance, HCCMDP has a greatly improved sensitivity when discriminating high-risk populations, especially in discriminating early-stage HCC patients from them. Under the classifier probability cut-off of 0.53 optimised in the algorithm development cohort, HCCMDP achieved 84% sensitivity and 86% specificity in distinguishing LC and CHB from early-stage HCC, while AFP's sensitivity was 67%. (Other statistics and comparisons between HCCMDP and AFP were summarised in Table 1 and Supplementary Tables S9 and S10). To rule out potential confounders such as aetiology, we conducted a sub-analysis of the panel's performance in patients with HCC and hepatitis B cirrhosis, patients with hepatitis B cirrhosis only and CHB patients. The panel has good capability in distinguishing them (Supplementary Table S11).

To evaluate the biomarker value of cfRNA fragments, we trained another random forest model for the 3 cfRNA fragment combination. The model was built on the algorithm development cohort and validated in the independent validation cohort. The 3 cfRNA fragment combination alone can predict AFP-positive patients at 100% sensitivity and detect AFP-negative HCC patients at 58.3% sensitivity (Supplementary Table S10), demonstrating their complementary role to AFP in detecting HCC.

Discussion

Our work differs from previous cfRNA biomarker studies that only focused on 1 or several limited RNA types like lncRNA or miRNA. Our work combines systematic review with NGS data analysis, which reduces the limitation of many previous studies using only one cfRNA-seq dataset to produce panels. As the sample

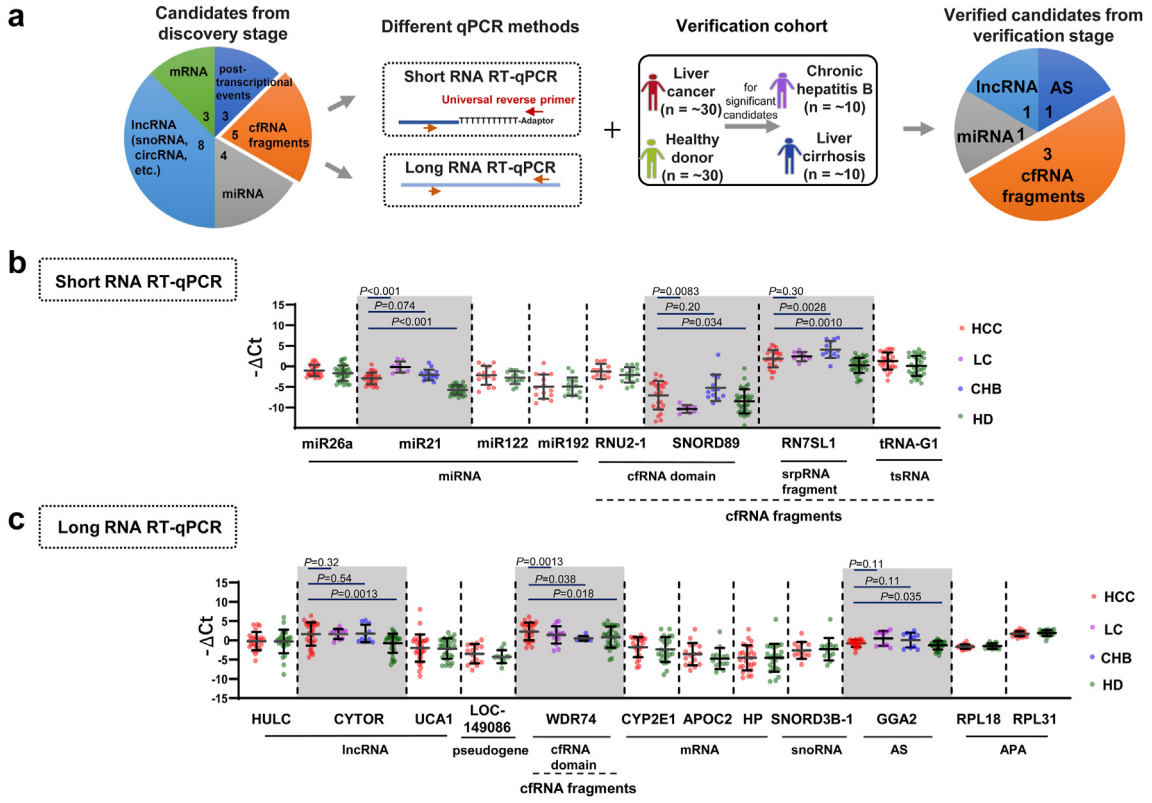


***Fig. 4:* Performance of selected candidate biomarkers in the verification cohort.** (a) The process of candidate selection in the verification stage. Candidates were allocated to different RT-qPCR strategies according to their lengths. (b) Validation of RNA candidates, each on about 30 HCC, 30 HD, 10 CHB, and 10 LC by short RNA RT-qPCR (normalised by *miR-16*). Candidates that could not be detected in multiple samples or had *P*-values >0.2 under Wilcoxon rank-sum test in distinguishing HCC and HD at about 15 HCC vs. 15 HD were aborted early. ns: not significant; HCC: liver cancer patients; HD: healthy donors; CHB: chronic hepatitis B patients; LC: liver cirrhosis patients. (c) Validation of long RNA candidates, each on about 30 HCC, 30 HD, 10 CHB, and 10 LC by long RNA RT-qPCR (normalised by *GAPDH*). Candidates that could not be detected in multiple samples or had *P*-values >0.2 under Wilcoxon rank-sum test in distinguishing HCC and HD at about 15 HCC vs. 15 HD were aborted early. The results of XK (a lncRNA) and 2 circRNAs are shown in Supplementary Figure S4.

sizes of cfRNA-seq datasets are usually not large enough to represent the population, this practice would cause poor generalization ability. This problem has been reflected in those candidates that are reported to have good performances but failed to be validated in our verification stage.

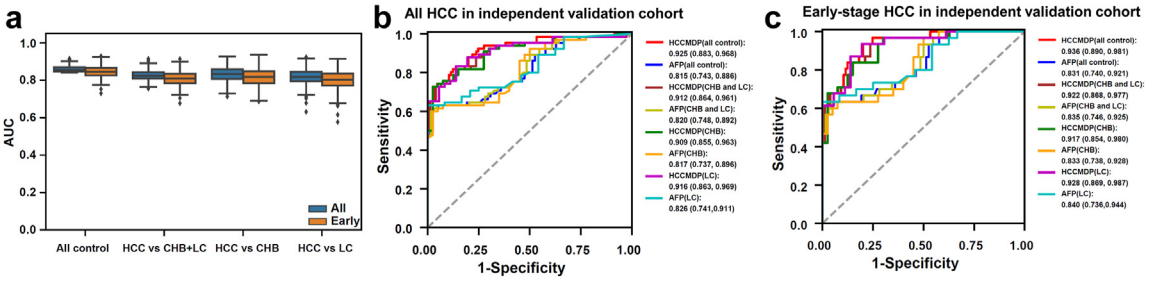


***Fig. 5:* Performances of the HCCMDP in algorithm development and independent validation cohorts.** (a) AUC distributions (displayed in boxplots) of HCCMDP in detecting HCC or early-stage HCC from all controls, CHB and LC from 500-times bootstrap in the algorithm development cohort. HCCMDP: Hepatocellular Carcinoma Molecular Detection Panel. (b) Comparisons of the receiver-operating characteristic curves between HCCMDP and AFP in discriminating HCC from all controls, CHB, and LC in the independent validation cohort. (c) Comparisons of the receiver-operating characteristic curves between HCCMDP and AFP in discriminating early-stage (Barcelona Clinic Liver Cancer (BCLC) stages 0/A) HCC from all controls, CHB, and LC in the independent validation cohort.

	HCCMDP			AFP		
	CHB + LC	CHB	LC	CHB + LC	CHB	LC
HCC vs						
Sensitivity, % (95% CI)	82 (73–91)	82 (73–91)	83 (74–91)	65 (52–75)	65 (52–74)	65 (54–75)
Specificity, % (95% CI)	86 (76–93)	85 (75–95)	86 (72–97)	80 (70–89)	73 (58–85)	92 (79–100)
Early-stage HCC vs						
Sensitivity, % (95% CI)	84 (71–97)	84 (71–97)	87 (74–97)	67 (50–83)	67 (50–83)	67 (50–83)
Specificity, % (95% CI)	86 (78–92)	85 (73–95)	86 (75–97)	80 (69–89)	73 (60–85)	92 (79–100)

AFP, a-fetoprotein; Early stages, Barcelona Clinic Liver Cancer stage 0/A; CI, Confidence interval.

Table 1: **HCCMDP's performances in distinguishing HCC from high-risk groups in the independent validation cohort.**

Here, we proposed the cfRNA domain to systematically describe cfRNA fragments captured in small cfRNA datasets. Together with previous findings,[16,17,19] our work shows the prevalence and abundance of this fragmented pattern of long RNAs in a cell-free environment. RNA fragments have been reported to have an important role in regulating cell viability, differentiation, and homeostasis through multiple mechanisms in cancer.[42] For example, exosomal RN7SL1's RBP-shielded region was reported to participate in enhancing tumour growth, metastasis, and therapy resistance.[43] In the extracellular space, EV was reported to selectively carry RNA fragments to regulate immune activation.[44] Therefore, the significant difference in the abundance of RNA fragments between HCC and HD patients may reflect their possible biological functions in tumours. Also, according to our analysis, cfRNA domains tend to have RNA binding proteins and secondary structures, which may enhance their stability in the blood. These factors above may be related to the finding that cfRNA fragments perform better in distinguishing HCC from non-HCC as compared to their parent cfRNAs. Furthermore, the cfRNA domain is constructed on small RNA seq data, the most prevalent library type in public RNA-seq datasets. Therefore, the cfRNA domain can be applied to other small RNA seq data to study cfRNA fragments in other diseases.

Compared with other biomarker-based tests like AFP,[45] AFP-L3, and GALAD score,[2] our panel demonstrated good sensitivity (82%) and specificity (86%) in HCC detection from high-risk populations, especially superior in early-stage HCC detection (sensitivity = 84%, specificity = 86%). Our panel combined the traditional protein marker with cfRNA markers, including an RNA alternative splicing event, which may enhance the robustness of the panel when facing the heterogeneity of patients in large cohorts. Another advantage of our panel is gaining accuracy while keeping low cost. Compared to other high-profile methods,[46,47] the cost of the RT-qPCR-based HCCMDP ($30–$50) is significantly cheaper and faster in giving results. This may help our panel fit into broader clinical application scenarios.

We used 400 ng/mL as the cutoff for AFP in our study mainly because the study cohorts are all Chinese and we used the cutoff recommended in the HCC guidelines in China.[48] Compared to the AASLD recommendation of 20 ng/mL,[49] the 400 ng/mL threshold has a higher specificity and AUC,[50] which is critical considering the large base of Chinese people at risk of liver cancer.

Several limitations exist in our study. We used all existing public cfRNA-seq datasets of HCC in GEO, but the total number of samples is still insufficient for calculating robust markers. The systematic review can compensate for this shortage to a certain degree. However, some eligible studies might have been missed despite the fact that we did thorough searches in the databases. Due to the low-throughput nature of RT-qPCR and insufficient number and blood sample volume from patients, especially in the verification cohort, the number of candidates we can validate by RT-qPCR is limited. Some candidates with good classification ability who do not fit our selection criteria may be missed. The SYBR Green approach is less expensive; however, some specific RNAs may be missed because of limitations in its sensitivity and specificity. Also, the samples in the independent validation cohort, especially the early-stage HCC patients, are limited, which causes relatively wide confidence intervals and may deviate our estimation of the performance of the panel. Limitations exist in our choice of control cohorts, in which we included the most prevalent etiologies that progress into HCC in China but did not include other liver diseases that can also move into HCC. As our study is retrospective, further large cohort validation in a prospective setting is required.

In conclusion, combining the systematic review and comprehensive analysis of cfRNA-seq, we identified cfRNA biomarkers of full-spectrum, revealed the biomarker potential of cfRNA fragments, and validated them on multicentre cohorts to develop and validate HCCMDP for HCC detection. Our work highlights the biomarker value of cfRNA fragments in plasma, delivers a promising panel, and offers guidance for biomarker selection in larger prospective cohorts.

Contributors

Concept, design and supervision: C.N., L.C., Z.J.L., and J.Y.

Material support: P.C., G.L., Y.W., Z.W., H.G, L.C., and J.Y.

Experimental research work: C.N., P.C., P.B., L.Y., M.N., K.T., and Y.L.

Drafting of the manuscript: C.N., L.C., Z.J.L., and J.Y.

Verifying underlying data and Statistical analysis: X.L., and C.N.

Obtained funding: Z.J.L., and J.Y.

Critical revision of the manuscript for important intellectual content: All authors.

All authors read and approved the final version of the manuscript.

Data sharing statement

All relevant data are within the paper and its supplementary files. The cfRNA datasets used in this study are available as described in the Method section. Patient-level data are available upon request. The source code is placed at https://github.com/xfliu1995/HCCMDP.

Declaration of interests

The authors declare no potential conflicts of interest.

Acknowledgements

The authors gratefully acknowledge the supports from National Natural Science Foundation of China [81972798 , 32170671, 81373067]; The National Key Basic Research Program (973 program) [2015CB554005]; Program of Shanghai Academic Research Leader [22XD1404800]; Tsinghua University Spring Breeze Fund [2021Z99CFY022]; Fok Ying Tong Education Foundation; Beijing Advanced Innovation Centre for Structural Biology; Open Research Fund Program of Beijing National Research Centre for Information Science and Technology; Bioinformatics Platform of National Centre for Protein Sciences (Beijing) (2021-NCPSB-005).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104645.

References

1 Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249.
2 Kanwal F, Singal AG. Surveillance for hepatocellular carcinoma: current best practice and future direction. *Gastroenterology*. 2019;157(1):54–64.
3 Fattovich G, Bortolotti F, Donato F. Natural history of chronic hepatitis B: special emphasis on disease progression and prognostic factors. *J Hepatol*. 2008;48(2):335–352.
4 Okajima W, Komatsu S, Ichikawa D, et al. Liquid biopsy in patients with hepatocellular carcinoma: circulating tumor cells and cell-free nucleic acids. *World J Gastroenterol*. 2017;23(31):5650–5668.
5 Zhou J, Yu L, Gao X, et al. Plasma microRNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma. *J Clin Oncol*. 2011;29(36):4781–4788.
6 Xie H, Ma H, Zhou D. Plasma HULC as a promising novel biomarker for the detection of hepatocellular carcinoma. *Biomed Res Int*. 2013;2013:136106.
7 Tan C, Cao J, Chen L, et al. Noncoding RNAs serve as diagnosis and prognosis biomarkers for hepatocellular carcinoma. *Clin Chem*. 2019;65(7):905–915.
8 Zhu Y, Wang S, Xi X, et al. Integrative analysis of long extracellular RNAs reveals a detection panel of noncoding RNAs for liver cancer. *Theranostics*. 2021;11(1):181–193.
9 Yu J, Ding WB, Wang MC, et al. Plasma circular RNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma: a large-scale, multicenter study. *Int J Cancer*. 2020;146(6):1754–1763.
10 Caviglia GP, Abate ML, Gaia S, et al. Risk of hepatocellular carcinoma in HBV cirrhotic patients assessed by the combination of miR-122, AFP and PIVKA-II. *Panminerva Med*. 2017;59(4):283–289.
11 Ding Y, Yan JL, Fang AN, Zhou WF, Huang L. Circulating miRNAs as novel diagnostic biomarkers in hepatocellular carcinoma detection: a meta-analysis based on 24 articles. *Oncotarget*. 2017;8(39):66402–66413.
12 Ma J, Li T, Han X, Yuan H. Knockdown of LncRNA ANRIL suppresses cell proliferation, metastasis, and invasion via regulating miR-122-5p expression in hepatocellular carcinoma. *J Cancer Res Clin Oncol*. 2018;144(2):205–214.
13 Bonnal SC, López-Oreja I, Valcárcel J. Roles and mechanisms of alternative splicing in cancer - implications for care. *Nat Rev Clin Oncol*. 2020;17(8):457–474.
14 Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol*. 2017;18(1):18–30.
15 Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer*. 2015;15(6):371–381.
16 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.
17 Giraldez MD, Spengler RM, Etheridge A, 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(11):e101695.
18 Zhu L, Li J, Gong Y, et al. Exosomal tRNA-derived small RNA as a promising biomarker for cancer diagnosis. *Mol Cancer*. 2019;18(1):74.
19 von Felden J, Garcia-Lezana T, Dogra N, et al. Unannotated small RNA clusters associated with circulating extracellular vesicles detect early stage liver cancer. *Gut*. 2022;71(10):2069–2080.
20 Bu FT, Zhu Y, Chen X, et al. Circular RNA circPSD3 alleviates hepatic fibrogenesis by regulating the miR-92b-3p/Smad7 axis. *Mol Ther Nucleic Acids*. 2021;23:847–862.
21 Abdelgawad IA, Radwan NH, Hassanein HR. KIAA0101 mRNA expression in the peripheral blood of hepatocellular carcinoma patients: association with some clinicopathological features. *Clin Biochem*. 2016;49(10):787–791.
22 Tao Q, Zhu K, Zhan Y, et al. Construction of a novel exosomes-related gene signature in hepatocellular carcinoma. *Front Cell Dev Biol*. 2022;10:997734.
23 Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*. 2009;339:b2700.
24 Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536.
25 Shim SR, Kim SJ, Lee J. Diagnostic test accuracy: application and practice using R software. *Epidemiol Health*. 2019;41:e2019007.
26 Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*. 2005;58(10):982–990.
27 Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17(1):3.
28 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–359.
29 Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
30 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842.
31 Iyer MK, Niknafs YS, Malik R, et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*. 2015;47(3):199–208.
32 Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. *RNA*. 2014;20(11):1666–1670.
33 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26(1):139–140.
34 O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB. GeneGini: assessment via the Gini coefficient of reference "housekeeping" genes and diverse human transporter expression profiles. *Cell Syst*. 2018;6(2):230–244.e1.
35 Lu Z, Zhang QC, Lee B, et al. RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell*. 2016;165(5):1267–1279.
36 Zhao W, Zhang S, Zhu Y, et al. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res*. 2022;50(D1):D287–D294.
37 Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017–1018.

38 Shen S, Park JW, Lu ZX, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*. 2014;111(51):E5593–E5601.
39 Xia Z, Donehower LA, Cooper TA, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun*. 2014;5:5274.
40 Lee M, Lee K, Yu N, et al. ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Res*. 2017;45(D1):D784–D789.
41 Singh S, Qin F, Kumar S, et al. The landscape of chimeric RNAs in non-diseased tissues and cells. *Nucleic Acids Res*. 2020;48(4):1764–1778.
42 Fu M, Gu J, Wang M, et al. Emerging roles of tRNA-derived fragments in cancer. *Mol Cancer*. 2023;22(1):30.
43 Nabet BY, Qiu Y, Shabason JE, et al. Exosome RNA unshielding couples stromal activation to pattern recognition receptor signaling in cancer. *Cell*. 2017;170(2):352–366.e13.
44 Chiou NT, Kageyama R, Ansel KM. Selective export into extracellular vesicles and function of tRNA fragments during T cell activation. *Cell Rep*. 2018;25(12):3356–3370.e4.
45 Hu X, Chen R, Wei Q, Xu X. The landscape of alpha fetoprotein in hepatocellular carcinoma: where are we? *Int J Biol Sci*. 2022;18(2):536–551.
46 Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*. 2018;359(6378):926–930.
47 Liu MC, Oxnard GR, Klein EA, et al. Sensitive and specific multicancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol*. 2020;31(6):745–759.
48 Zhou J, Sun H, Wang Z, et al. Guidelines for the diagnosis and treatment of hepatocellular carcinoma (2019 edition). *Liver Cancer*. 2020;9(6):682–720.
49 Heimbach JK, Kulik LM, Finn RS, et al. AASLD guidelines for the treatment of hepatocellular carcinoma. *Hepatology*. 2018;67(1):358–380.
50 Zhang J, Chen G, Zhang P, et al. The threshold of alpha-fetoprotein (AFP) for the diagnosis of hepatocellular carcinoma: a systematic review and meta-analysis. *PLoS One*. 2020;15(2):e0228857.