Circulating Small Extracellular Vesicle RNA Profiling for the Detection of T1a stage Colorectal Cancer and Precancerous Advanced Adenoma

Li Min, Fanqin Bu, Jingxin Meng, Xiang Liu, Qingdong Guo, Libo Zhao, Zhi Li, Xiangji Li, Shengtao Zhu*, Shutian Zhang**

Prof. L. Min, Dr. Fanqin Bu, Mr. Q. Guo, Prof. S. Zhu, Prof. S Zhang

Department of Gastroenterology, Beijing Friendship Hospital, Capital Medical University, National Clinical Research Center for Digestive Diseases, Beijing Digestive Disease Center, Beijing Key Laboratory for Precancerous Lesion of Digestive Disease, Beijing, 100050, P. R. China.

E-mail: zhangshutian@ccmu.edu.cn, shengtaozhu@126.com, minli@ccmu.edu.cn

Prof. L. Min, Prof. J. Meng

Key Laboratory of Bio-inspired Materials and Interfacial Science, Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Beijing, 100190, P. R. China.

Mr. X. Liu, Dr. L. Zhao, Mr. Z. Li

Echo Biotech Co., Ltd, Beijing, 102206, P. R. China.

Dr. Xiangji Li

Department of Retroperitoneal Tumor Surgery, International Hospital, Peking University, Beijing, 102206, P. R. China.

Keywords: extracellular vesicles; RNA profiling; early colorectal cancer; advanced adenoma

Abstract

It takes more than twenty years for normal colorectal mucosa to develop into metastatic carcinoma. The long time window provides a golden opportunity for early detection to terminate the malignant progression. Here we aim to enable liquid biopsy of T1a stage colorectal cancer (CRC) and precancerous advanced adenoma (AA) by profiling circulating small extracellular vesicle (sEV)-derived RNAs. We exhibited a full RNA landscape for the circulating sEVs isolated from 60 participants. A total of 58,333 annotated RNAs were detected from plasma sEVs, among which 1,615 and 888 sEV-RNAs were found differentially expressed in plasma from T1a stage CRC and AA compared to normal controls (NC). Then we further categorized these sEV-RNAs into 6 modules by a weighted gene coexpression network analysis and constructed a 60-gene t-SNE model consisting of the top 10 RNAs of each module that could well distinguish T1a stage CRC/AA from NC samples. Some sEV-RNAs were also identified as indicators of specific endoscopic and morphological features of different colorectal lesions. The top-ranked biomarkers were further verified by RT-qPCR, proving that these candidate sEV-RNAs successfully identified T1a stage CRC/AA from NC in another cohort of 124 participants. Finally, we adopted different algorithms to improve the performance of RT-qPCR-based models and successfully constructed an optimized classifier with 79.3% specificity and 99.0% sensitivity. In conclusion, circulating sEVs of T1a stage CRC and AA patients have distinct RNA profiles, which successfully enable the detection of both T1a stage CRC and AA via liquid biopsy.

Introduction

With 1,880,725 new cases and 915,880 deaths, colorectal cancer (CRC) ranks as the 3rd most commonly diagnosed and 2nd most lethal cancer worldwide in 2020.[1] Both the incidence and mortality of CRC in developing countries are gradually increasing due to rapid social-economic improvement.[2] Even though a remarkably high CRC incidence was observed in developed areas, such as Europe and North America, the mortality of CRC has dramatically decreased over the past decade owing to the continuous improvement of multidisciplinary treatment and the promotion of CRC screening programs.[3] Noticeably, endoscopic treatments enable radical resection of T1a stage CRC without open surgery, and the 5-year survival rate of those patients has been improved to over 90%.[4, 5] Advanced adenoma (AA) is the crucial precancerous lesion of CRC, which is also resectable under endoscopy.[6-9] About 40% of AA patients could progress to invasive CRC within 10 years, which provided a very wide window of opportunity for curative treatment.[10] Unfortunately, most AA patients were asymptomatic, and hard to be identified by opportunistic screening. Thus, early diagnosis of CRC and precancerous AA is of vital importance.

Endoscopy is the main approach for the screening of CRC and AA, whereas the invasiveness and complicated operation procedures largely restricted its application in asymptomatic populations. For most endoscopy screening programs, participants are simply pre-selected by their age and other epidemiological factors, resulting in low compliance and unsatisfactory cost-effectiveness.[11, 12] Fecal occult blood test (FOBT) and fecal immunohistochemistry test (FIT) are also popular in identifying individuals of high CRC/AA risk, whereas their sensitivities are far from satisfactory.[13] The performances of FOBT and FIT to detect AA are especially disappointing, which exhibited a sensitivity of 20% (95% CI 6.8% to 40.7%) and 32% (95% CI 14.9% to 53.5%), respectively.[14] Therefore, the invention of alternative technologies allowing early CRC and AA detection with minimal invasion is urgently needed.

Extracellular vesicles are various-sized membrane particles derived from host cells, which become a new source of biomarkers in liquid biopsy.[15, 16] Small extracellular vesicles (sEVs) of 40-100nm diameter isolated by 100K ultracentrifugation were mostly considered to be exosomes, which derived from endosomes and participated in cell-cell communication.[17] There are plenty of RNA species stuffed in sEVs, and one major function of sEVs is the delivery of functional donor cell RNAs to the recipient cell.[17, 18] Additionally, the phospholipid bilayer of sEVs could effectively protect enclosed RNA from the RNase in the environment,[19, 20] thus making sEV-RNA a relatively stable detection target. Consequently, for the diversity of enclosed RNAs and the properties of being protected from degradation by RNase,[21] the sEV-RNAs become an attractive treasury of biomarkers in cancer diagnosis.

Many studies have already reported the function of sEV-RNAs in the progress of many diseases and their potential application in diagnosis and prognosis.[22-25] However, even though it is generally accepted that a landscape profiling of the contents from plasma EVs would largely facilitate the progress of biomarker discovery in liquid-biopsy of cancers, systematic screening of plasma sEV-RNAs is extremely hampered by the highly instrument-dependent, time-consuming isolation of sEVs from plasma and the exquisite manipulating of tiny amount RNA.[26] Previously, with a modified differential centrifugation (DC) procedure,[27] we isolated sEVs from plasma and characterized those sEVs according to the MISEV2018 guideline.[28] Then we proved that sEVs encapsuled miRNAs outperformed their free-floating counterparts in diagnostic liquid-biopsy, which proposed a new promising biomarker category.[29] We also verified the abnormal abundance of several miRNAs in the circulating sEVs of CRC patients.[22, 29] However, even though the sEV miRNAs have already been carefully investigated, the panorama of circulating sEV-RNAs in CRC patients has not been revealed yet. Additionally, detecting AA by liquid biopsy could also be much more challenging than detecting CRC since AA always exhibited a smaller size and fewer vascular inside networks as compared to CRC.[30]

To strengthen the ability of sEV biomarkers in identifying both T1a stage CRC and AA, here we accomplished the first whole-transcriptomic profiling of circulating sEV-RNAs in a large cohort with 60 participants, including T1a stage CRC, AA, and normal controls (NC). The identification of CRC or AA-specific lncRNA and mRNA largely extended the sEVs biomarker repertoire, considering there were more than 50,000 lncRNA and mRNA species, nearly 25 times of known miRNA species.[31] Finally, with a comprehensive analysis of plasma sEVs, we obtained the T1a stage CRC and precancerous AA-specific sEV-RNA landscape, proposed a 10-gene signature, and verified its ability in identifying T1a stage CRC and AA in another cohort of 124 participants (**Figure 1**).

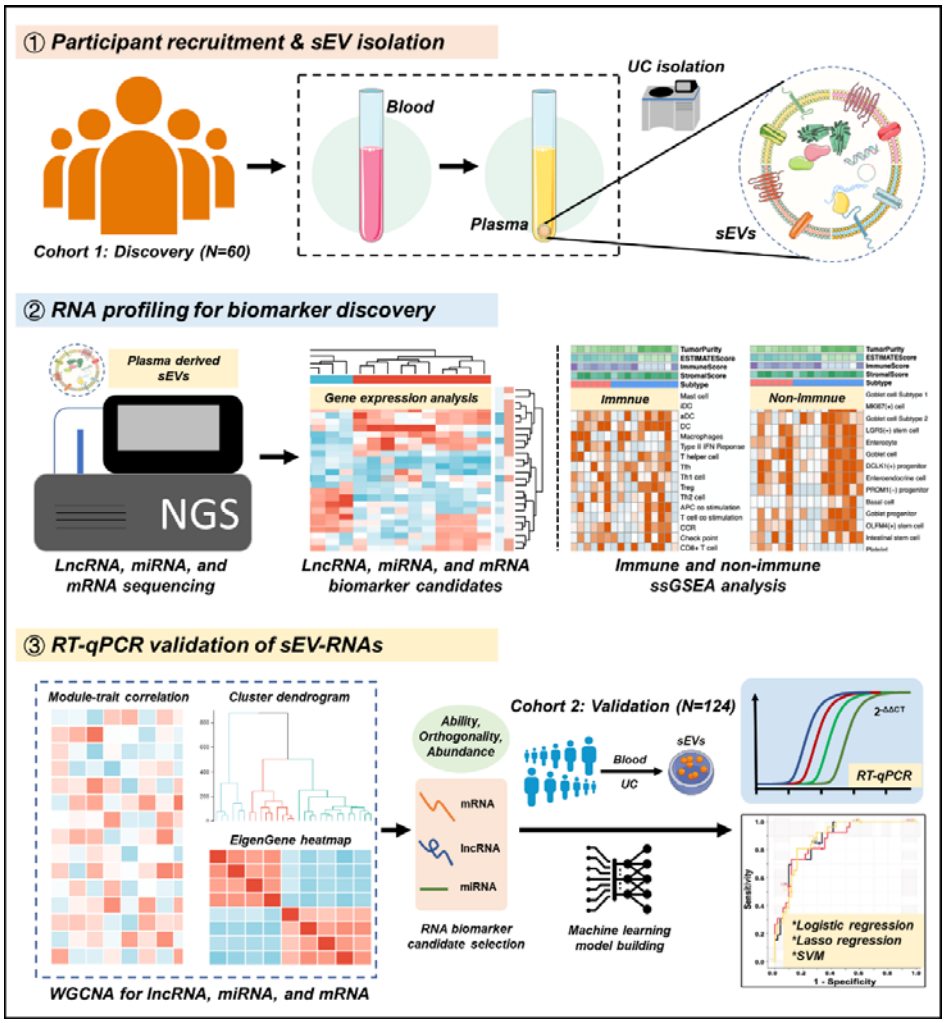

Figure 1. Schematic overview of the study design.

Results

Plasma-derived sEV characterization and RNA profiling

TEM imaging showed that the sEVs isolated from plasma exhibited cup-shaped, vesicle-like structures (**Figure 2**a). NTA analysis revealed a heterogeneous size distribution ranging between 75nm to 200nm (Figure 2b). Western blot analysis exhibited an enrichment of sEVs markers (CD63, TSG101, and Alix) and an absence of Calnexin, a negative marker of sEVs, indicating that the isolated fractions consisted mostly of sEVs (Figure 2c). The total RNA levels of the sEVs fractions were also assessed, and an EV-associated RNA concentration of 3.70±2.39ng per mL plasma was reported.

For RNA profiling, a total of 58,333 annotated genes, including 2,694 miRNAs, 24,927 mRNAs, and 30,712 lncRNAs, were detected. The numbers of detected miRNA species are almost equal among different groups, while there were slightly more mRNA and lncRNA species in T1a stage CRC than in NC/AA (Figure 2d). Differentially expressed genes (DEGs) among different groups were filtered by an FDR-adjusted ANOVA p-value. The top 100 miRNAs, mRNAs, and lncRNAs could roughly distinguish T1a stage CRC patients from AA and NC participants by unsupervised hierarchical clustering, respectively (Figure 2e). The clustering results using all those 300 RNAs exhibited clear discrimination between T1a stage CRC/AA patients and NC participants, and crude discrimination between T1a stage CRC and AA patients (Figure 2—figure supplement 1). Unsupervised t-SNE clustering by those 300 RNAs provided a very clear separation of T1a stage CRC, AA, and NC individuals (Figure 2f, Figure 2—figure supplement 2).

The DEGs under comparisons between any two groups were identified by the Mann-Whitney U tests, and the overlaps of those DEGs were shown in a Venn diagram (Figure 2g, Table S1-9). Noticeably, there were 888 DEGs between AA and NC, and 519 (58%) of them overlapped with DEGs between T1a stage CRC and NC, suggesting most sEV-RNAs in AA remained abnormally expressed while progressing to T1a stage CRC. KEGG pathway analysis

suggested that those DEGs were enriched in Pathways in cancer, MAPK signaling, Focal Adhesion, *etc.*, indicating a close relationship with cancer progression (Figure 2h). Additionally, the potential core regulatory networks between miRNAs and mRNAs in those DEGs were also exhibited (Figure 3i-k).

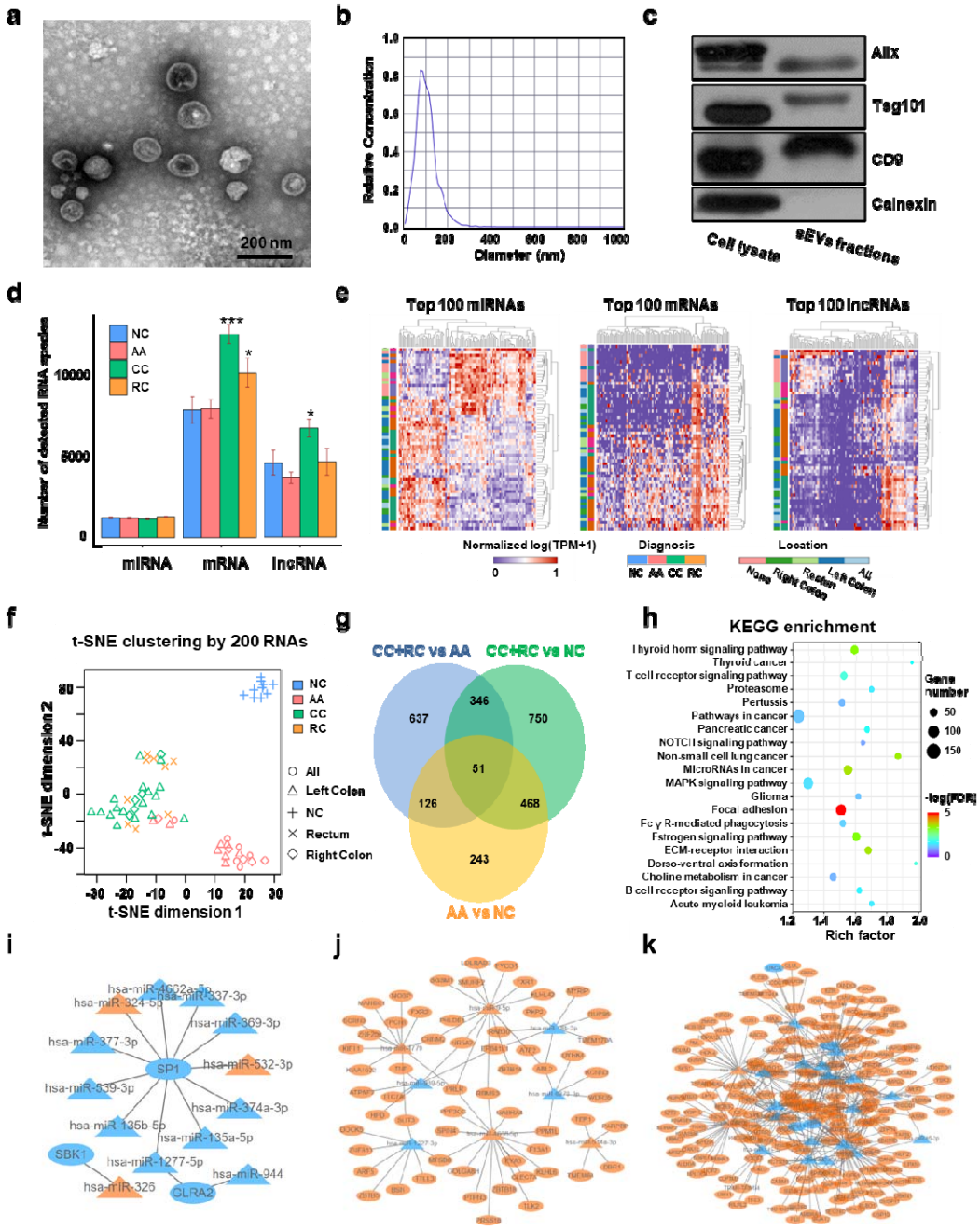

Figure 2. Transcriptome profiling of circulating sEVs. a) TEM images of circulating sEVs isolated from human plasma. b) NTA results of circulating sEVs enriched from plasma. c)

WB results of sEV positive (Alix, TSG101, CD9) and negative (Calnexin) markers. d) The numbers of detected RNA species in different groups. e) The hierarchical clustering results of top 100 miRNAs (left panel), mRNAs (middle panel), and lncRNAs (right panel). f) t-SNE clustering by those 300 RNAs. g) A Venn diagram showed DEGs shared between different comparisons (CRC vs NC, AA vs NC, CRC vs AA). h) KEGG enrichment of all those DEGs identified. i-k) potential core regulatory networks between miRNAs and mRNAs in DEGs identified in three (i), two (j), and one (k) of all comparisons.

Cell-specific features of the sEV-RNA profile indicated the different proportion of cells of sEV origin among different groups

We employed the ssGSEA algorithm to calculate the correlations between the sEV-RNA profile of each sample and the transcriptome characteristics of different cells to investigate the possible difference in the proportion of cells of sEV origin among different groups. The immune cell-specific features of each sample were shown in **Figure 3**a, and the stromal-related features were shown in Figure 3b. The complete landscape of cell-specific features of the sEV-RNA profile were shown in Figure 3—figure supplement 1a. Generally, plasma sEV-RNAs of different groups showed distinct cell-specific features. Especially, sEV-RNA of CC and RC showed higher expression levels of RNA markers of inflammation-promoting cells, MHC class I, HLA, TIL, and Paneth cells (Figure 3c). Correspondingly, RNA markers of DC cells, Mast cells, and Enterocyte cells were mostly enriched in sEV-RNAs of NC and AA compared to those of CC and RC (Figure 3d).

According to those cell-specific features, the tumor microenvironment-associated scores were calculated. The sEV-RNAs of CC, RC, and AA exhibited significantly higher Immune scores and significantly higher Estimate scores as compared to those of NC, whereas the NC group possessed the highest Stromal score (Figure 3e). Then we estimated the possible correlation among different cell-specific features, and a strong correlation among Inflammation promoting cells, MHC class I, HLA, TILs, Paneth cells, and Paneth-like cells

was identified (Figure 3f, Figure 3—figure supplement 1b). Taking all together, cell-specific features of the sEV-RNA profile indicated that there is a higher proportion of inflammatory cell-originated sEVs in the plasma of CC, RC, and AA patients as compared to NC participants.

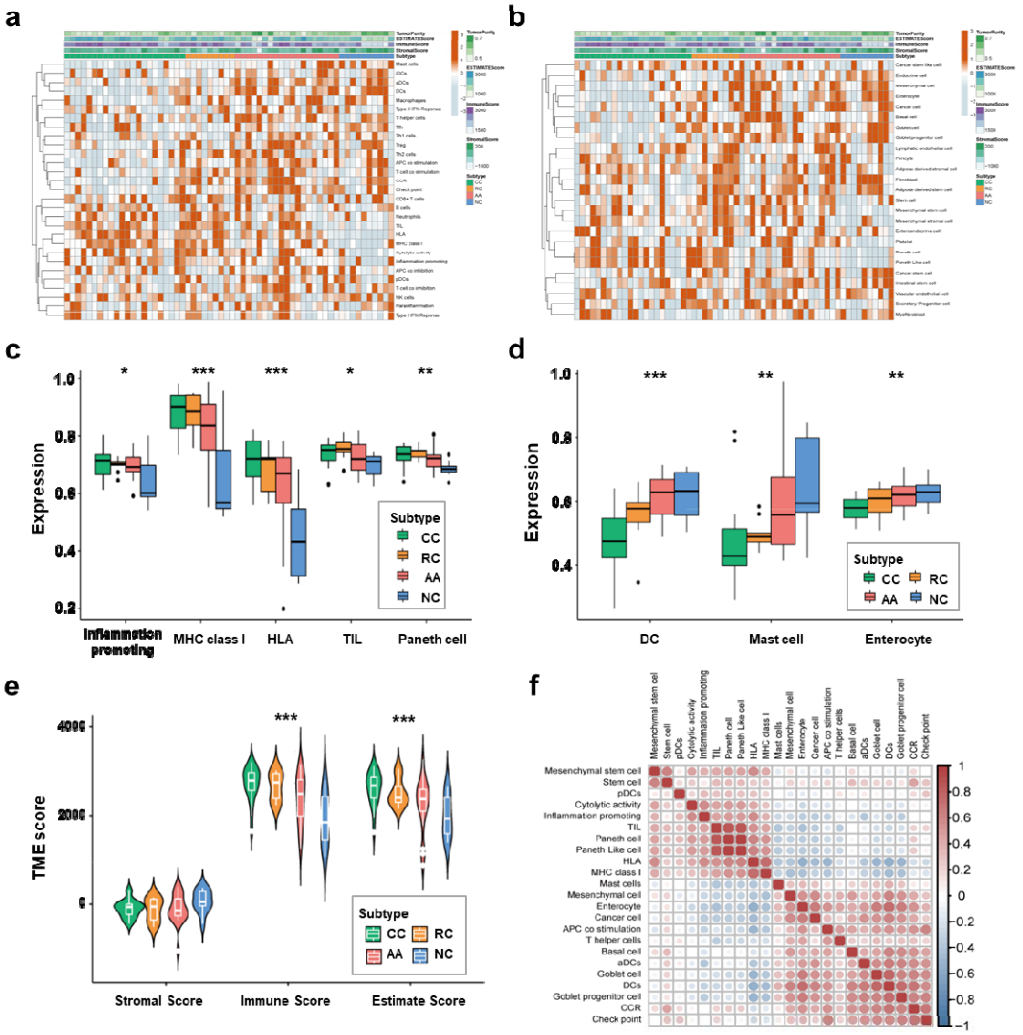

Figure 3. Cell-specific features of the sEV-RNA profile. a) The hierarchical clustering heatmap of immune cell-specific features of each sample. b) The hierarchical clustering heatmap of stromal-related features of each sample. c) Boxplot of cell-specific features overexpressed in CC and RC patients (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). d) Boxplot of cell-specific features overexpressed in NC participants (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

e) The violinplot of the microenvironmental scores in different subgroups (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). f) Correlation among cell-specific features differentially enriched among different groups.

sEV-derived DEGs could be divided into 6 WGCNA modules

We performed WGCNA to reveal the inner correlation between DEGs and clinical parameters. Six coexpression modules were constructed based on the expression levels of 1525 DEGs in each sample by WGCNA (**Figure 4**a). The associations among different modules were shown in a heatmap (all DEGs with kME>0.7 in each module were used for calculating the Pearson correlation), revealing relatively high correlations between the green and turquoise modules, and between the blue and brown modules (Figure 4b). Additionally, the number and percentage of different RNA species in each module were also displayed in Figure 4c and Figure 4d, while the distributions of DEGs with kME>0.7 were shown in Figure 4—figure supplement 1. Noticeably, most miRNAs were categorized into the brown module, suggesting that the abundances of those sEV-miRNAs could be affected by the same factors, which further enhanced the necessity of including mRNAs and lncRNAs as additional biomarkers to cover as many patients as possible.

The expression levels of the top 10 DEGs in each module were shown in a heatmap (Figure 4e). We further used those DEGs for t-SNE analysis to cluster different samples and obtained a nearly perfect separation of T1a stage CRC, AA, and NC individuals (Figure 4f). In light of this result, we then tried to reduce the number of DEGs used for t-SNE analysis by picking only the top 5 DEGs and the top 1 DEG from each module, respectively. As shown in Figure 4g, when enrolled more DEGs of each module, the overall distinguishing effect significantly increased with less sample mis-clustered. Noticeably, the top 1 DEG from each module could separate the NC samples from T1a stage CRC and AA samples very clearly, meanwhile, CRC samples could also be blurrily separated from AA samples with only 3 exceptions. (Figure 4h).

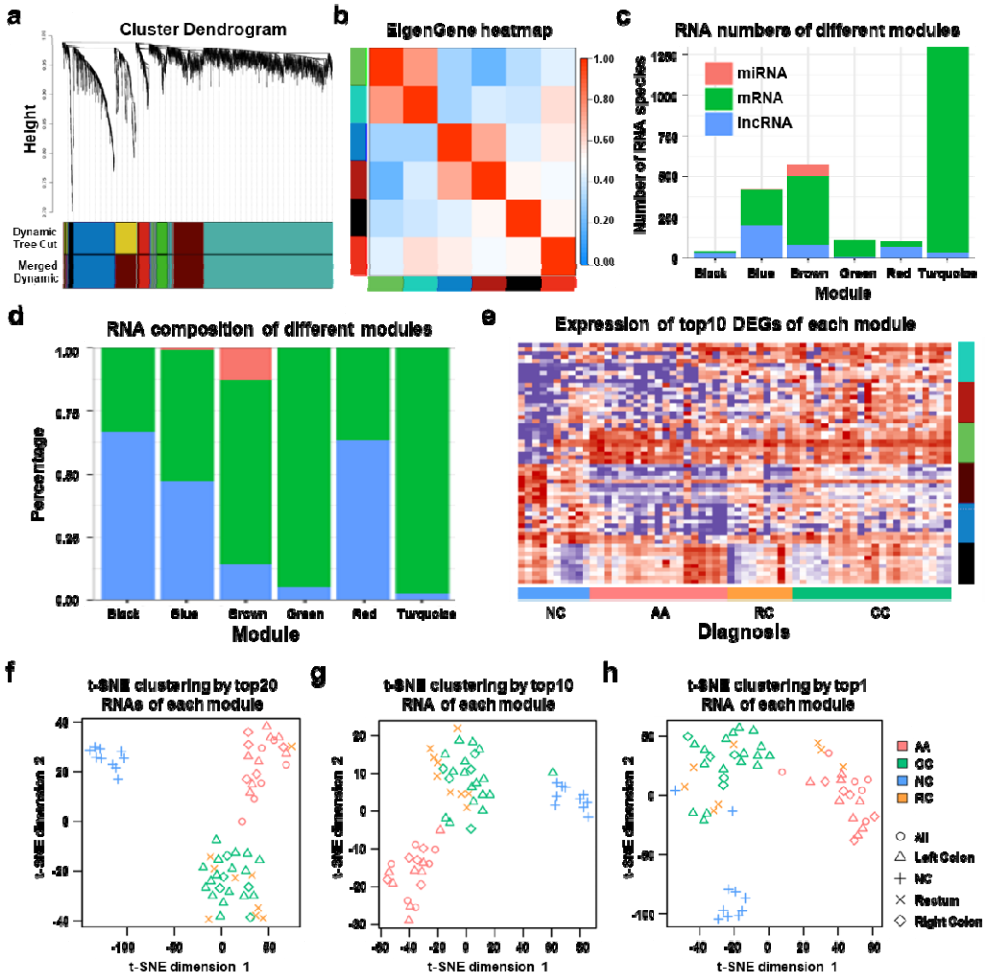

Figure 4. WGCNA analysis of sEV-RNAs. a) Gene coexpression module construction of all DEGs identified in sEV-RNAs. b) The heatmap exhibited Pearson correlations among different modules. c) Bar plot of module composition of different modules (all DEGs). d) Percentage bar plot of the RNA composition of different modules (all DEGs). e) A heatmap exhibited the expression levels of the top 10 DEGs in each module. f) t-SNE clustering by the top 10 DEGs in each module. g) t-SNE clustering by the top 5 DEGs in each module. h) t-SNE clustering by the top1 DEGs in each module.

Different modules showed different expression trends

We conducted a GSEA analysis to extract major trends in the DEG expression of different modules. In the comparison between T1a stage CRC and NC, red, green, and turquoise DEGs were enriched in CRC samples, while black, blue, and brown DEGs were enriched in NC

samples (Figure 5a). In the comparison between AA and NC, black, red, green, and turquoise DEGs were enriched in CRC samples, while blue and brown DEGs were enriched in NC samples (Figure 5b). In the comparison between T1a stage CRC and AA, black, brown, red, blue, and turquoise DEGs were enriched in CRC samples, while green DEGs were enriched in AA samples (Figure 5c).

The expression trends of the Top 10 DEGs of each module among NC, AA, and T1a stage CRC were also displayed (Figure 5d-i). Generally, green, red, turquoise and black DEGs are all overexpressed in AA as compared to NC. However, the expression level of green DEGs remained unchanged (Figure 5d), red/turquoise DEGs further increased to a much higher level (Figure 5ef), while black DEGs almost decreased to the baseline level in CRC (Figure 5g). Blue and brown DEGs showed similar expression trends, which decreased to the lowest level in AA, and were partially restored in T1a stage CRC samples (Figure 5hi).

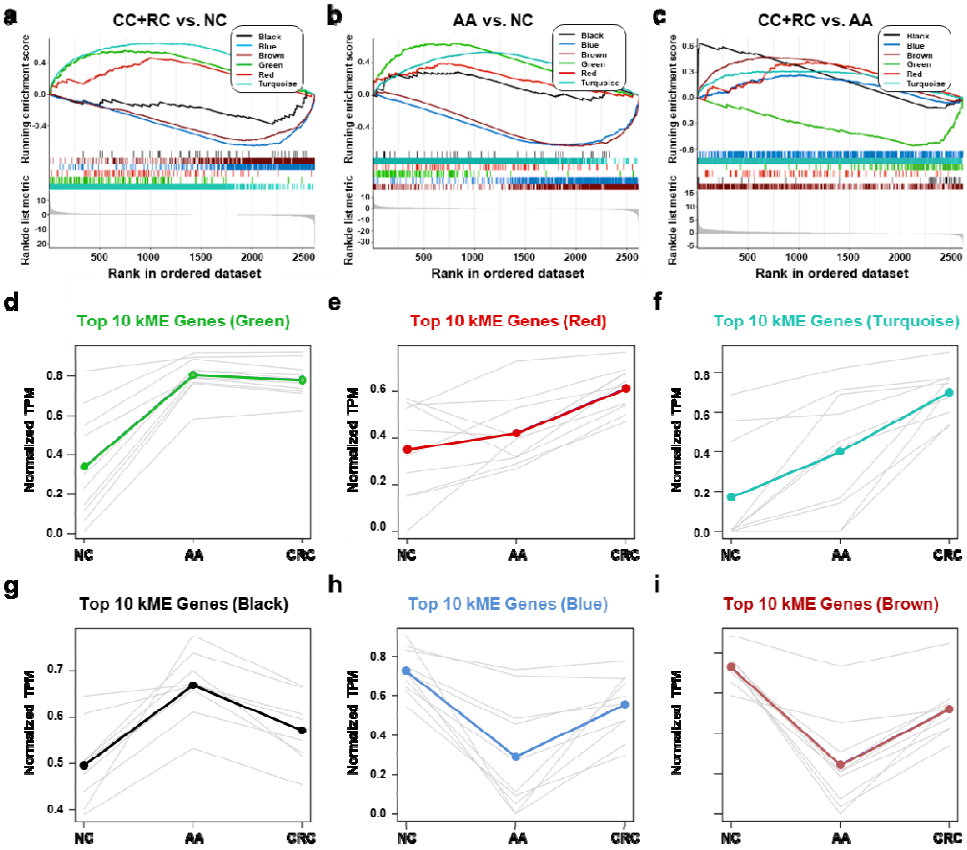

Figure 5. The expression trends of sEV-RNA modules. a-c) GSEA analysis of DEGs in different modules (a: CRC vs. NC; b: AA vs. NC; c: CRC vs. AA). d-i) The expression trends of the Top 10 DEGs of each module among NC, AA, and CRC (d: green module; e: red module; f: turquoise module; g: black module; h: blue module; i: brown module).

sEV-RNAs in different modules were correlated with different clinical factors

Detecting sEV-RNAs could also provide more data beyond diagnosis. Here we performed module-trait correlation analysis to reveal clinical factors associated with those sEV-RNAs. The red module was significantly associated with endoscopic classification (**Figure 6**a). The black module was significantly associated with the endoscopic classification and the LST morphology, while the green module was significantly associated with the LST morphology and the multiple-sited features (Figure 6a).

In brief, the DEG modules we identified in WGCNA showed a close association with morphological features of colorectal tumors, which was also well verified in RT-qPCR of representative sEV-RNAs. For example, flat lesions (subtype II) cases exhibited overexpressed MT-ND2 (blue module) levels in plasma sEVs as compared to protruding lesions (subtype Is/Ip/Isp) (Figure 6b, left panel), while LST tumors showed a relatively higher level of HIST2H2AA4 (green module) than non-LST tumors (Figure 6b, right panel).

Unsupervised hierarchical clustering suggested that sEV-RNAs of those three modules could not only separate T1a stage CRC/AA from NC but also roughly distinguish CRC and AA samples with different endoscopic classifications and other clinical features, such as LST and location (Figure 6c). The inner correlations among sEV-RNAs in the same module were also evaluated and exhibited by Circos plots. The top 2 sEV-RNAs of the green module exhibited a very strong correlation with each other (Figure 6d), while in other modules (red and black) the top 6 sEV-RNAs showed relative average correlation coefficients with each other (Figure 6ef).

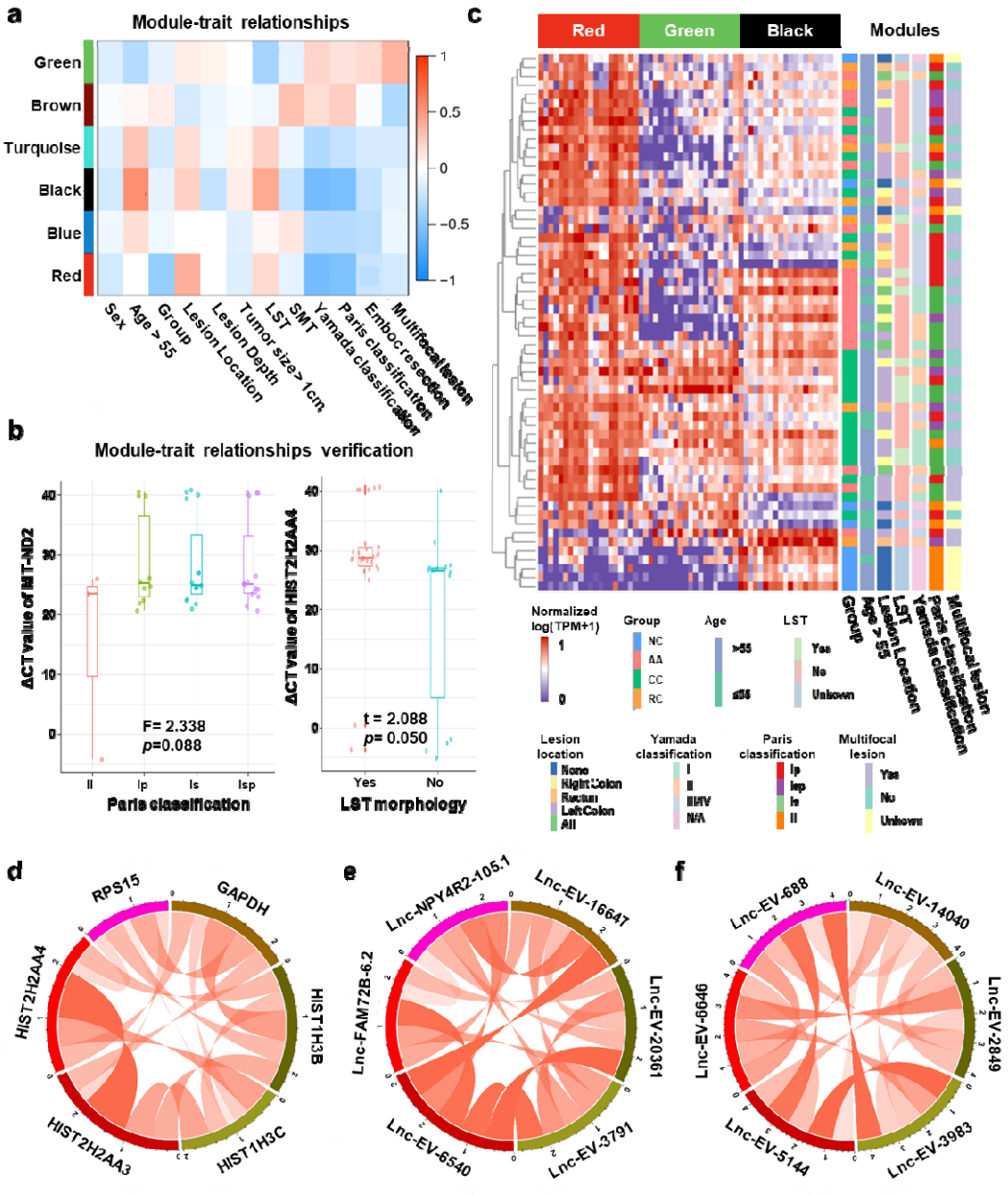

Figure 6. Module-trait correlation analysis of sEV-RNA modules. a) The heatmap exhibited the correlation between modules and clinical traits. b) The RT-qPCR validation of representive module-trait correlation (left panel: correlation between MT-ND2 and Paris classification; right panel: correlation between HIST2H2AA4 and LST morphology). c) The heatmap exhibited the sEV-RNA expression levels of red, black, and green modules. d-f) Circos plot showed the inner correlations among sEV-RNAs in the module green (d), red (e), and black (f).

Establishment of a sEV-RNA signature for AA and T1a stage CRC diagnosis

Even though the RNA signatures we established in Figure 4f-h showed promising performance in distinguishing T1a stage CRC, AA from NC, this next-generation sequencing (NGS)-based quantification system is not affordable in prospective large-scale screening of CRC. RT-qPCR is a cost-effective alternative to NGS in diagnostic tests, thus here we designed an RT-qPCR-based assay to diagnose CRC and AA patients.

We extracted all sEV-RNAs upregulated in CRC or AA using a 4-fold change cutoff (Table S10, S11). High-abundance sEV-RNAs were considered more appropriate for RT-qPCR quantification since they are much easier to detect in liquid biopsy-obtained trace samples. Therefore, we excluded all low-abundance sEV-RNAs (median TPM <50). A signature covering more different modules is preferred since it includes more orthogonal factors in prediction. Thus, the candidate sEV-RNAs were finally selected based on their fold change in CRC/AA, absolute abundance, and module attribution (**Table 1**).

Simultaneously detecting AA and CRC by a RT-qPCR-based sEV-RNA signature

The expression levels of those candidate sEV-RNAs were quantified by RT-qPCR in an independent cohort of 124 participants. Lasso regression was applied to select the most effective variables from all candidate sEV-RNAs to construct a multivariate CRC prediction model (Figure 7—figure supplement 1). The model with 8 sEV-RNAs exhibited an AUC of 0.76 (**Table 2**: upper panel; **Figure 7**a-d). Three different algorithms (logistic regression, Lasso regression, and SVM) were compared to further boost the performance of the 8-gene signature, and the highest AUC of 0.80 was achieved in the SVM model (Figure 7a-d). In predicting Stage I CRC patients, these sEV-RNAs panels also exhibit a promising predictive performance (Figure 7—figure supplement 2). Furthrtmore, we assessed the discriminative effects of CRC on NC, taking into account different age groups, genders, tumor sizes, and

tumor anatomical locations. To minimize the potential overfitting effect due to the reduction in sample size after partitioning, we implemented a 10-fold cross-validation for each panel and these sEV-RNAs panels exhibit promising performance in different clinical parameters (Figure 7—figure supplement 3).

Similarly, a model with 9 sEV-RNAs exhibited an AUC of 0.88 to predict AA from NC (Table 2: lower panel; Figure 7e-h). The Lasso model showed the highest AUC of 0.88 among all three algorithms, signifying an excellent efficacy in AA detection among noncancerous individuals (Figure 7e-h).

Candidate	*AA vs NC*	*CRC vs NC*	*CRC vs AA*	*Module attribution*	*RNA Type*	*Amount*	*Finally seclected*
miR-3615	+*	+	-**	*brown*	*miRNA*	*High*	*Yes*
miR-330-5p	+	+	-	*brown*	*miRNA*	*Low*	*No*
miR-425-5p	+	+	-	*NA*	*miRNA*	*High*	*Yes*
miR-106b-3p	+	+	-	*NA*	*miRNA*	*High*	*Yes*
miR-589-5p	+	+	-	*NA*	*miRNA*	*Low*	*No*
miR-181a-2-3p	+	+	-	*brown*	*miRNA*	*Low*	*No*
Let-7f-5p	-	-	+	*NA*	*miRNA*	*High*	*Yes*
Let-7e-5p	-	-	+	*NA*	*miRNA*	*High*	*No*
miR-320a/b-3p	-	-	+	*brown*	*miRNA*	*High*	*Yes*
miR-664a-5p	-	-	+	*brown*	*miRNA*	*Low*	*No*
YBX3	-	+	+	*turquoise*	*mRNA*	*Low*	*No*
C19orf43	-	+	+	*turquoise*	*mRNA*	*Medium*	*Yes*
TOP1	+	+	-	*turquoise*	*mRNA*	*Medium*	*Yes*
PPDPF	+	+	-	*brown*	*mRNA*	*Medium*	*Yes*
MT-ND2	+	+	-	*blue*	*mRNA*	*High*	*Yes*
HIST2H2AA4	+	+	-	*green*	*mRNA*	*Medium*	*Yes*
RPL10	+	+	-	*green*	*mRNA*	*High*	*No*
RPS29	+	+	-	*blue*	*mRNA*	*High*	*No*
IST1	-	+	+	*black*	*mRNA*	*Low*	*No*
CSE1L	-	+	-	*red*	*mRNA*	*Low*	*No*
lnc-MSI1-2:1	+	+	-	*brown*	*lncRNA*	*High*	*Yes*
lnc-FCGR1B-16:1	-	+	+	*red*	*lncRNA*	*Low*	*No*
lnc-NPY4R2-105:1	-	+	+	*red*	*lncRNA*	*Medium*	*No*

lnc-MKRN2-42:1	-	+	+	*turquoise*	*lncRNA*	*High*	*Yes*
LNC_EV_9572(Chr8: 34358093-34456247)	+	+	-	*black*	*lncRNA*	*High*	*Yes*
LNC_EV_21004(Chr21: 8212554-8440060)	+	+	-	*brown*	*lncRNA*	*High*	*No*
LNC_EV_15260(Chr14: 49555875-49923916)	+	+	-	*turquoise*	*lncRNA*	*High*	*No*

Table 1. Selection of plasma sEVs-RNA candidates for AA and T1a stage CRC diagnosis

**+: significant difference found in this comparision. ** -: no significant difference found in this comparision.*

Model type	*Signature*	*RNAs*	*Lambda*	*AUC*
CRC prediction	*5-RNA signature*	*Let-7f-5p, C19orf43, TOP1, PPDPF, lnc-MKRN2-42:1*	*0.05*	*0.73*
	6-RNA signature	*Let-7f-5p, C19orf43, TOP1, PPDPF, lnc-MKRN2-42:1, LNC-EV-9572*	*0.035*	*0.73*
	7-RNA signature	*Let-7f-5p, C19orf43, TOP1, PPDPF, lnc-MKRN2-42:1, LNC-EV-9572, HIST2H2AA4*	*0.03*	*0.74*
	8-RNA signature	*Let-7f-5p, C19orf43, TOP1, PPDPF, lnc-MKRN2-42:1, LNC-EV-9572, HIST2H2AA4, miR-320a-3p*	*0.02*	*0.76*
AA prediction	*6-RNA signature*	*miR-425-5p, Let-7f-5p, C19orf43, TOP1, PPDPF, LNC-EV-9572*	*0.05*	*0.83*
	7-RNA signature	*miR-425-5p, Let-7f-5p, C19orf43, TOP1, PPDPF, LNC-EV-9572, lnc-MKRN2-42:1*	*0.04*	*0.84*
	8-RNA signature	*miR-425-5p, Let-7f-5p, C19orf43, TOP1, PPDPF, LNC-EV-9572, lnc-MKRN2-42:1, HIST2H2AA4*	*0.1*	*0.87*
	9-RNA signature	*miR-425-5p, Let-7f-5p, C19orf43, TOP1, PPDPF, LNC-EV-9572, lnc-MKRN2-42:1, HIST2H2AA4, MT-ND2*	*0.05*	*0.88*

Table 2. CRC and AA prediction models established by Lasso regression

Additionally, we adopted quadratic discriminant analysis (QDA) to demonstrate the possibility of the plasma sEV-RNA signature for direct sample classification. An overall accuracy of 78% (ranging from 63% to 100%) was obtained for direct sample classification (Figure 7ij). Generally, the individuals classified into AA/CC/RC are considered as high-risk and should be advised to further endoscopic examination, and our QDA classifier provided a specificity of 79.25%, and a sensitivity of 99.0% (with only one CC sample missed) in identifying those high-risk individuals. Together, our RT-qPCR-based plasma sEVs-RNA

signature could be a powerful and better alternative to FIT and FOBT tests in CRC and precancerous AA screening programs.

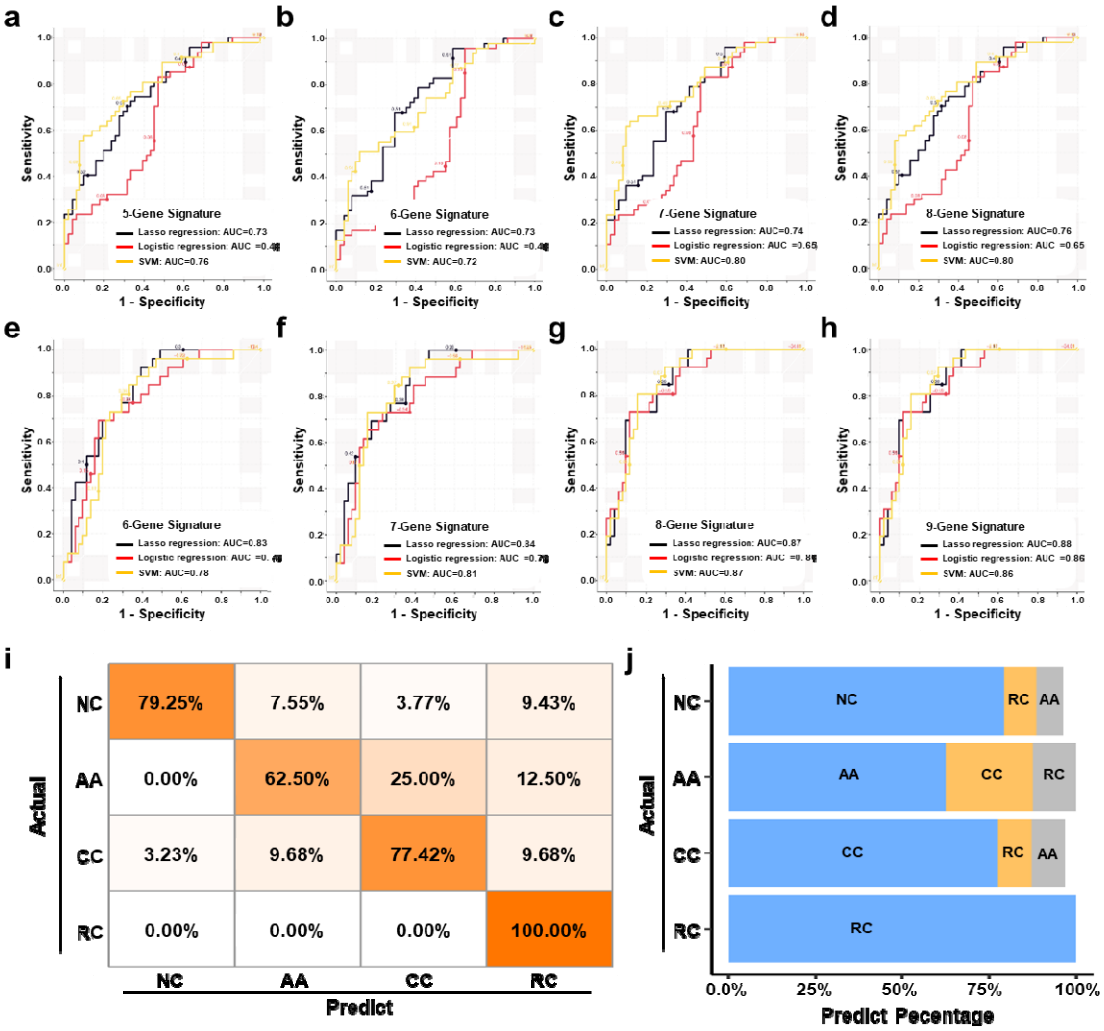

Figure 7. The plasma sEVs-RNA signature to detect early CRC and AA. a-d) The ROC analysis of different sEV-RNA signatures in the prediction of CRC patients by different algorithms (a: 5-gene panel; b: 6-gene panel; c: 7-gene panel; d: 8-gene panel). E-h) The ROC analysis of different sEV-RNA signatures in the prediction of AA patients by different algorithms (a: 6-gene panel; b: 7-gene panel; c: 8-gene panel; d: 9-gene panel). i) The QDA results of all 13 sEV-RNAs in classifying all samples. J) Statistical summary of QDA performance in each sample group.

Discussion

Previously, we proved that plasma sEV-encapsulated miRNAs outperformed their free-floating counterparts in the diagnostic liquid biopsy of cancer.[29] However, whole-transcriptomic profiling of plasma sEVs is much more challenging and highly attractive, as there are many more long transcript species than miRNAs in plasma sEVs.[21] Additionally, mRNAs and lncRNAs encapsulated in sEVs manifested very important functions during carcinogenesis and cancer progression. For example, lncARSR in sEVs promoted sunitinib resistance in renal cancer in a ceRNA-dependent manner.[32] Similarly, sEV-delivered HChrR6 mRNA showed a therapeutic effect in the treatment of human HER2(+) breast cancer.[33] Thus, whole-transcriptomic profiling of circulating sEVs would not only facilitate the minimally invasive diagnosis of T1a stage CRC but also promote our understanding of the microenvironment change during CRC carcinogenesis.

Here we identified 2,694 miRNAs (1,838 known, 856 novel), 24,927 mRNAs (18,947 known, 5,980 novel), and 30,712 lncRNAs (12,928 known, 17,784 novel) in human circulating sEVs. A total of 2,621 RNA species showed a primary potential in distinguishing colorectal tumors from others, which was a great treasury of biomarker candidates in liquid biopsy. The profiling of circulating sEV-RNAs for early detection of CRC and AA we provided here, which is, to our knowledge, not only the first published whole-transcriptomic profile (including miRNA, mRNA, and lncRNA) dataset from human plasma EVs, but also the first attempt to simultaneously diagnose early CRC and precancerous AA by a plasma sEV-RNA signature. It is also worth noting that Zheng *et al*. reported a data-independent acquisition (DIA)-mass spectrometry (MS) based protein profiling of circulating sEVs.[34] The plasma sEV-protein signature exhibited very good performance, especially for CRC patients with liver metastases. Thus, here we also suggested that a comprehensive biomarker panel consisting of both plasma sEV-RNAs and proteins could be very promising in identifying both early and advanced CRC patients in the future.

Our results proved that sEV-RNAs could well separate T1a stage CRC, AA, and NC from each other, indicating a significant overall change of plasma sEV-RNA profile during the whole process of CRC carcinogenesis. We also found that those sEV-RNAs were enriched in pathways associated with cancer, MAPK signaling, Focal adhesion, *etc.*, suggesting those sEV-encapsulated RNAs could be potentially effective. Noticeably, those sEV-RNAs were not necessarily originated from cancer cells, since target-selecting procedures, such as anti-EpCAM-based immuno-capture,[23, 35, 36] were not conducted before sEVs isolation. Nevertheless, biomarker studies that enrolled all possible biomarkers regardless of their origination could provide a larger candidate pool, a higher abundance and detectability, and a simplified procedure for easier detection of given targets. Detailed single sEV analysis or other methodology revealing sEV-RNA heterogeneity[37, 38] could further improve the performance of sEV-based liquid biopsy and reveal the underlying molecular mechanisms.

Here we provided a 6-RNA signature based on the RNA-seq data, which exhibited a good effect in distinguishing T1a stage CRC, AA from NC. The simplicity of this signature is quite inspiring but not surprising, considering it took advantage of the discrimination efficiency of sEV-RNAs and the orthogonality between WGCNA modules.[39] Some but not all of the key sEV-RNAs we identified have already been reported involved in carcinogenesis. For example, miR-425-5p was reported in promoting CRC via BRAF/RAS/MAPK Pathways.[40] Hypoxia-induced let-7f-5p regulated osteosarcoma cell proliferation and invasion by inhibiting the Wnt signaling.[41] Additionally, patients with higher PPDPF expression exhibited worse prognosis than others,[42] while high HIST1H2BK levels predicted poor prognosis in glioma patients.[43] Further investigation on the functions of other key sEV-RNAs was needed to provide a full landscape of the molecular mechanism of their roles during carcinogenesis.

To facilitate the clinical application of sEV-RNAs in CRC screening, an RT-qPCR-based assay was designed. RNA candidates were selected from the sEV-RNAs identified by RNA-seq under a tradeoff between expected performance and RT-qPCR detectability, and sEV-

RNAs with higher abundance distributed in different modules were preferred. Although we got a good performance for this sEV-RNA signature, other gene selection strategies were also worth trying and improved results could also be possibly achieved by a more meticulous pipeline considering other factors such as sample-specific effects.[44] To maximize the performance of given sEV-RNA signatures, we compared the traditional logistic regression model with two popular machine learning algorithms, *ie.* SVM and Lasso regression. The SVM model and Lasso model showed higher AUC in identifying T1a stage CRC and AA, respectively. QDA analysis was also tested for a direct classification of different samples, which exhibited a specificity of 79.25%, and a sensitivity of 99.0%. Those efforts also highlighted the power of modeling algorithm selection in biomarker discovery and utilization.

Detecting sEV-RNAs could also provide more data beyond diagnosis. The modules we identified in WGCNA showed a very close association with morphological features of colorectal tumors, which was also verified in RT-qPCR analysis of representative sEV-RNAs. For example, LST tumors showed a relatively higher level of plasma sEV-derived HIST2H2AA4 than non-LST tumors, while flat lesions exhibited overexpressed plasma sEV-derived MT-ND2 levels than protruding lesions. Considering the higher HIST2H2AA4 mRNAs and MT-ND2 mRNAs in tissues could reflect a higher proliferation rate and higher energy metabolism,[43, 45] their elevated levels in sEVs could be a potential indicator of invasiveness. Though those interpretations are hypothetical, the fact that one could get biological information on cancerous and precancerous tumors from plasma sEVs is still inspiring. Hence we believed that in-depth data mining of sEV-RNAs acquired by liquid biopsy would be helpful in different scenarios of healthcare decision-making.

Conclusions

We demonstrated the clinical value of circulating sEV-RNA profiling in CRC biomarker discovery and established a high-accuracy, low-cost RNA signature that could detect both T1a

stage CRC and AA from other individuals. We also suggested that T1a stage CRC and precancerous AA patients retained a specific plasma sEV-RNA profile, which would provide insight into an in-depth understanding of the mechanism of CRC carcinogenesis.

Methods

Patients' information and plasma collection

For biomarker discovery, 31 early (T1a stage) CRC patients, including 22 colon cancer (CC) patients and 9 rectum cancer (RC) patients who received endoscopic resection at Department of Gastroenterology, Beijing Friendship Hospital between January 2018 and December 2019 were enrolled along with 19 AA patients (characterized by high-grade dysplasia or adenomas ≥10mm or ≥25% villous component) and 10 non-cancerous (NC) outpatients with other gastrointestinal symptoms. Additional 124 participants (47 CRC, 24 AA, and 53 NC) were enrolled for validation. All clinical information was summarized in Table S12. This study was approved by the ethics committee of Beijing Friendship Hospital, and written informed consent was obtained from each participant.

A total of 6mL blood from each individual was collected in EDTA tubes in the morning before any food/water intake. After two-step centrifugation at 1,300×g, 15min (isolate plasma from whole blood) and 3,000×g for 15min (platelet removal) at 4°C, the supernatant of the processed plasma was aspirated and stored at -80 °C before use.

sEV isolation by a modified DC procedure

A total of 3mL plasma was collected from each participant. The DC isolation procedure was performed according to our previous report.[29] Briefly, the plasma was centrifugated at 3,000×g for 15min and 13,000×g for 30min, to remove cell debris and large EVs respectively, and the final supernatant was 1:7 diluted by PBS, filtered by a 0.22μm sieve, ultracentrifuged using a P50AT2-986 -rotor (CP100NX; Hitachi, Brea, CA) at 150,000×g, 4°C for 4h to collect the sEVs. The pellet was washing in PBS and centrifuged again at 150,000×g, 4°C for

2h. Then the sEVs enriched fraction was re-suspended in 100μL PBS. The full description of methodologies was also submitted to EV-TRACK (ID: EV190033). [46]

sEV characterization

The protein of sEVs was quantified by a BCA Assay Kit (Thermo Scientific, Product No. 23225) according to the manufacturer's protocol. A total of 10μg protein was loaded to each lane for Western blot analysis. The detailed procedure was referred to our previous publication,[29] using CD63 antibody (sc-5275, Santa Cruz, CA, USA), TSG101 antibody (sc-13611, Santa Cruz, CA, USA), Alix antibody (sc-53540, Santa Cruz, CA, USA) and calnexin antibody (10427-2-AP, Promega, Madison, WI).

Nanoparticle tracking analysis (NTA) and transmission electron microscopy (TEM) were used for morphological characterization, which was conducted according to our previous publication (Supplementary methods file).[29]

RNA sequencing, primary data processing, and analysis

ExoRNA was isolated by the miRNeasy® Mini kit (No. 217004, Qiagen, Hilden, Germany). RNA degradation and possible DNA contamination were monitored on a 1.5% agarose gel. Two different RNA sequencing libraries (long RNA, short RNA) were constructed, respectively. Detailed procedures of RNA library preparation and sequencing were shown in the Supplementary methods file. Furthermore, detailed information on the identification, quantification, and differential expression analysis of miRNAs, mRNAs, and lncRNAs was also shown in the Supplementary methods file. The RNA sequencing data have been deposited at the Sequence Read Archive (SRA) database of NCBI under the accession number PRJNA639943.

Single sample gene set enrichment analysis (ssGSEA)

"GSVA" R package (version 1.42.0) was employed to analyze the cell-specific features of the sEV RNA profiles. One-way ANOVA test was employed to identify the differentially

enriched cell-specific features among different groups. “ESTIMATE” R package (version 1.0.13) was used to estimate the stromal score, immune score and microenvironmental score.

Weighted gene coexpression network analysis (WGCNA)

The “WGCNA” R package (version 1.61) was used to construct a co-expression network for all differentially expressed genes (DEGs). All samples were used to calculate Pearson’s correlation matrices. The weighted adjacency matrix was created with the formula amn = |cmn|β (amn: adjacency between gene m and gene n; cmn: Pearson’s correlation between gene m and gene n; β: soft-power threshold). Furthermore, the weighted adjacency matrix was transformed into a topological overlap measure (TOM) matrix to estimate its connectivity property in the network. Average linkage hierarchical clustering was used to construct a clustering dendrogram of the TOM matrix. The minimal gene module size was set to 30 to obtain appropriate modules, and the threshold to merge similar modules was set to 0.1.

GeneSet Enrichment Analysis (GSEA) and t-SNE clustering

GSEA is supported by the “DOSE” R package (version 3.8.0). Each comparison (CRC vs NC, AA vs NC, CRC vs AA) was used as a phenotype label, while the gene list of each module obtained by WGCNA was adopted as a gene set. A signal-to-noise metric was used for ranking genes and all other parameters are all set as default.

The “Rtsne” R package (version 0.15) was used for t-distributed stochastic neighbor embedding (t-SNE) clustering. RNAseq data from all 60 samples were analyzed using t-SNE-based clustering of 300 DEGs, 162 DEGs, 18 DEGs, respectively.

Quantification of RNA expression with qPCR

Synthetic *Caenorhabditis elegans* cel-39-3p was spiked into each sEVs sample as an external calibration before RNA extraction. Total RNA was isolated and purified from sEVs factions by a miRNeasy® Mini kit (cat. 217004, Qiagen, Hilden, Germany). The total RNA was then reversely transcribed to synthesize cDNA using the PrimeScript™ RT reagent Kit (TAKARA, RR037A). The abundance of target gene expression was detected by the

TaqMan® probe using real-time qPCR. 2 μL of cDNA was used as the template for each PCR reaction. The sequences of primers and probes were shown as Table S13. All samples were normalized by the initial biofluid input volume used for RNA extraction and calibrated by the amount of spike-in cel-39-3p.

Statistical analysis

Statistical analysis were all performed by R 3.8.2 (www.r-project.org). $P<0.05$ was considered significant, and FDR adjusted p-value was calculated for multiple comparisons. All tests were two-tailed. The efficiency of candidate RNA models was assessed by calculating the area under the ROC curve (AUC). Pheatmap, VennDiagram, and ggplot2 were used for data visualization.

Acknowledgements

We thank Dr. Da Qin and Ms. Rui Wei (Capital Medical University) for their discussion and valuable advice on this study. We thank Ms. Yuan Wang, Ms. Jing Chen, Dr. Guanyi Kong (Echo Biotech Co., Ltd) for the meticulous work in project coordination. We thank Ms. Yun Zhang and the Clinical Data and Biobank Resource of Beijing Friendship Hospital for their help in sample preservation. This work was supported by grants from the Beijing Nova Program of Science and Technology (Z191100001119128); Beijing Municipal Science and Technology Project (Z191100006619081); National Natural Science Foundation of China (82073390); Beijing Municipal Administration of Hospitals' Youth Programme (QML20180108) and The Digestive Medical Coordinated Development Center of Beijing Municipal Administration of Hospitals (XXZ02, XXZ01). The study sponsors had no role in the design and preparation of this manuscript.

Author contributions

Study concept and design: LM and Shutian Zhang.

Data acquisition and data cleaning: LM and FB.

Data analysis and interpretation: FB, XL, LBZ and ZL.

Drafting of the manuscript: LM, QG, and XJL.

Supervision: LM, Shengtao Zhu and Shutian Zhang.

Proofreading and revision: LM, JM and XJL.

Declaration of interests

The authors declare no conflict of interest.

Reference

1. Sung H, *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin.* **71**, 209-249 (2021).

2. Chen W, *et al.* Cancer statistics in China, 2015. *CA Cancer J. Clin.* **66**, 115-132 (2016).

3. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. *CA Cancer J. Clin.* **67**, 7-30 (2017).

4. Li H, *et al.* Cytoplasmic Asporin promotes cell migration by regulating TGF-β/Smad2/3 pathway and indicates a poor prognosis in colorectal cancer. *Cell Death Dis.* **10**, (2019).

5. Cao B, Min L, Zhu S, Shi H, Zhang S. Long-term oncological outcomes of local excision versus radical resection for early colorectal cancer in young patients without preoperative chemoradiotherapy: a population-based propensity matching study. *Cancer Med.* **7**, 2415-2422 (2018).

6. Hsu WF, *et al.* Higher risk of advanced histology in adenoma less than 10 mm in fecal immunochemical test screening: Implication for management. *J. Gastroenterol. Hepatol.* . **35**, 1738-1745 (2020).

7. Good NM, *et al.* Ideal colonoscopic surveillance intervals to reduce incidence of advanced adenoma and colorectal cancer. *J. Gastroenterol. Hepatol.* . **30**, 1147-1154 (2015).

8. Tehranian S, Klinge M, Saul M, Morris M, Diergaarde B, Schoen RE. Prevalence of colorectal cancer and advanced adenoma in patients with acute diverticulitis: implications for follow-up colonoscopy. *Gastrointest Endosc.* **91**, 634-640 (2020).

9. Shaukat A, Gravely AA, Kim AS, Rank J, Church TR, Allen JI. Rates of Detection of Adenoma, Sessile Serrated Adenoma, and Advanced Adenoma Are Stable Over Time and Modifiable. *Gastroenterology*. **156**, 816-817 (2019).

10. Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. *Gut.* **56**, 1585-1589 (2007).

11. Force USPST, *et al.* Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA.* **315**, 2564-2575 (2016).

12. Sung JJ, *et al.* An updated Asia Pacific Consensus Recommendations on colorectal cancer screening. *Gut.* **64**, 121-132 (2015).

13. Haug U, Kuntz KM, Knudsen AB, Hundt S, Brenner H. Sensitivity of immunochemical faecal occult blood testing for detecting left- vs right-sided colorectal neoplasia. *Br. J. Cancer.* **104**, 1779-1785 (2011).

14. Graser A, *et al*. Comparison of CT colonography, colonoscopy, sigmoidoscopy and faecal occult blood tests for the detection of advanced adenoma in an average risk population. *Gut.* **58**, 241-248 (2009).

15. Shah R, Patel T, Freedman JE. Circulating Extracellular Vesicles in Human Disease. *N. Engl. J. Med.* **379**, 958-966 (2018).

16. Min L, *et al.* Advanced Nanotechnologies for Extracellular Vesicle-Based Liquid Biopsy. *Adv. Sci. (Weinh).* **8**, e2102789 (2021).

17. An T, *et al.* Exosomes serve as tumour markers for personalized diagnostics owing to their important role in cancer metastasis. *J. Extracell. Vesicles.* **4**, 27522 (2015).

18. Thind A, Wilson C. Exosomal miRNAs as cancer biomarkers and therapeutic targets. *J. Extracell. Vesicles*. **5**, 31292 (2016).

19. Thery C. Cancer: Diagnosis by extracellular vesicles. *Nature*. **523**, 161-162 (2015).

20. Vlaeminck-Guillem V. Extracellular Vesicles in Prostate Cancer Carcinogenesis, Diagnosis, and Management. *Front. Oncol.* **8**, 222 (2018).

21. Jeppesen DK, *et al.* Reassessment of Exosome Composition. *Cell.* **177**, 428-445.e418 (2019).

22. Min L, *et al.* Loss of Circulating Exosomal miR-92b is a Novel Biomarker of Colorectal Cancer at Early Stage. *Int. J. Med. Sci.* **16**, 1231-1237 (2019).

23. Dorayappan KDP, *et al.* A Microfluidic Chip Enables Isolation of Exosomes and Establishment of Their Protein Profiles and Associated Signaling Pathways in Ovarian Cancer. *Cancer Res.* **79**, 3503-3513 (2019).

24. Ko J, *et al.* Diagnosis of traumatic brain injury using miRNA signatures in nanomagnetically isolated brain-derived extracellular vesicles. *Lab Chip.* **18**, 3617-3630 (2018).

25. Cheng H, *et al.* Detecting miRNA biomarkers from extracellular vesicles for cardiovascular disease with a microfluidic system. *Lab Chip.* **18**, 2917-2925 (2018).

26. Srinivasan S, *et al.* Small RNA Sequencing across Diverse Biofluids Identifies Optimal Methods for exRNA Isolation. *Cell.* **177**, 446-462 e416 (2019).

27. Wei R, *et al.* Combination of Size-Exclusion Chromatography and Ultracentrifugation Improves the Proteomic Profiling of Plasma-Derived Small Extracellular Vesicles. *Biol. Proced. Online.* **22**, 12 (2020).

28. Thery C, *et al.* Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines. *J Extracell Vesicles.* **7**, 1535750 (2018).

29. Min 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.* **8**, 1643670 (2019).

30. Hong W, Dong L, Stock S, Basharat Z, Zippi M, Zhou M. Prevalence and characteristics of colonic adenoma in mainland China. *Cancer Manag. Res.* **10**, 2743-2755 (2018).

31. Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? *Front. Genet.* **6**, 2 (2015).

32. Qu L, *et al.* Exosome-Transmitted lncARSR Promotes Sunitinib Resistance in Renal Cancer by Acting as a Competing Endogenous RNA. *Cancer Cell.* **29**, 653-668 (2016).

33. Forterre AV, *et al.* Extracellular Vesicle-Mediated In Vitro Transcribed mRNA Delivery for Treatment of HER2(+) Breast Cancer Xenografts in Mice by Prodrug CB1954 without General Toxicity. *Mol. Cancer Ther.* **19**, 858-867 (2020).

34. Zheng X, *et al.* A circulating extracellular vesicles-based novel screening tool for colorectal cancer revealed by shotgun and data-independent acquisition mass spectrometry. *J. Extracell. Vesicles.* **9**, 1750202 (2020).

35. Ostenfeld MS, *et al.* miRNA profiling of circulating EpCAM+extracellular vesicles: promising biomarkers of colorectal cancer. *J. Extracell. Vesicles.* **5**, 31488 (2016).

36. Zhang P, Zhou X, Zeng Y. Multiplexed immunophenotyping of circulating exosomes on nano-engineered ExoProfile chip towards early diagnosis of cancer. *Chem. Sci.* **10**, 5495-5504 (2019).

37. Colombo M, *et al.* Analysis of ESCRT functions in exosome biogenesis, composition and secretion highlights the heterogeneity of extracellular vesicles. *J. Cell Sci.* **126**, 5553-5565 (2013).

38. Lee K, *et al.* Multiplexed Profiling of Single Extracellular Vesicles. *ACS Nano.* **12**, 494-503 (2018).

39. Qin D, Wei R, Liu S, Zhu S, Zhang S, Min L. A Circulating miRNA-Based Scoring System Established by WGCNA to Predict Colon Cancer. *Anal. Cell. Pathol.* **2019**, 1-7 (2019).

40. Angius A, *et al.* MicroRNA-425-5p Expression Affects BRAF/RAS/MAPK Pathways In Colorectal Cancers. *Int. J. Med. Sci.* **16**, 1480-1491 (2019).

41. Chen G, Gu H, Fang T, Zhou K, Xu J, Yin X. Hypoxia-induced let-7f-5p/TARBP2 feedback loop regulates osteosarcoma cell proliferation and invasion by inhibiting the Wnt signaling pathway. *Aging.* **12**, 6891-6903 (2020).

42. Mao Z, Li X, Ma X, Wang X, Zhang J, Fan X. Pancreatic progenitor cell differentiation and proliferation factor predicts poor prognosis in heptaocellular carcinoma. *Medicine.* **98**, e14552 (2019).

43. Liu W, *et al.* High Levels of HIST1H2BK in Low-Grade Glioma Predicts Poor Prognosis: A Study Using CGGA and TCGA Data. *Front. Oncol.* **10**, 627 (2020).

44. Yu S, *et al.* Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. *Gut.* **69**, 540-550 (2020).

45. Sun W, Zhou S, Chang SS, McFate T, Verma A, Califano JA. Mitochondrial mutations contribute to HIF1alpha accumulation via increased reactive oxygen species and up-regulated pyruvate dehydrogenease kinase 2 in head and neck squamous cell carcinoma. *Clin. Cancer Res.* **15**, 476-484 (2009).

46. Consortium E-T, *et al.* EV-TRACK: transparent reporting and centralizing knowledge in extracellular vesicle research. *Nat. Methods.* **14**, 228-232 (2017).

Additional files

Supplementary methods

PART1. Morphological characterization of sEVs

PART2. RNA sequencing, primary data processing and analysis

Supplementary figures

Figure 2—figure supplement 1.

The hierarchical clustering results of Top 100 miRNAs/mRNAs/lncRNAs.

Figure 2—figure supplement 2.

Unsupervised t-SNE clustering by those 200 RNAs with 9 repeats.

Figure 3—figure supplement 1.

(A) The hierarchical clustering heatmap of different cell features in all sEV samples.

(B) Correlation among all cell-specific features.

Figure 4—figure supplement 1.

(A) Percentage barplot of the RNA composition of different modules (only DEGs with kME>0.7).

(B) Barplot of module composition of different modules (only DEGs with kME>0.7).

Figure 7—figure supplement 1.

(A) Performance of Lasso regression in variable selection to identify CRC.

(B) Performance of Lasso regression in variable selection to identify AA.

Figure 7—figure supplement 2.

The ROC analysis of different sEV-RNA signatures in the prediction of Stage I CRC patients by different algorithms (a: 6-gene panel; b: 7-gene panel; c: 8-gene panel; d: 9-gene panel).

Figure 7—figure supplement 3.

The ROC analysis of different sEV-RNA signatures for predicting CRC patients using the Lasso regression algorithm in different clinical parameters (ab: age; cd: gender; ef: tumor size; gh: anatomical position).

Supplementary tables

Suppl Table 1-CRC vs NC mRNA.

Suppl Table 2-CRC vs NC miRNA.

Suppl Table 3-CRC vs NC lncRNA.

Suppl Table 4-AA vs NC mRNA.

Suppl Table 5-AA vs NC miRNA.

Suppl Table 6-AA vs NC lncRNA.

Suppl Table 7-CRC vs AA mRNA.

Suppl Table 8-CRC vs AA miRNA.

Suppl Table 9-CRC vs AA lncRNA.

Suppl Table 10-AA vs NC up-regulated Median_50 Log2FC_2 with Module sorted by FDR.

Suppl Table 11-CRC vs NC up-regulated Median_50 Log2FC_2 with Module sorted by FDR.

Suppl Table 12-Participants' characteristics for the training and validation cohorts.

Suppl Table 13-Transcripts and sequence of their primers and probes.