Molecular Cancer

RESEARCH

Open Access

Identification of blood-derived exosomal tumor RNA signatures as noninvasive diagnostic biomarkers for multi-cancer: a multi-phase, multi-center study

Fubo Wang[1,2,3,4,5*†], Chengbang Wang[1,2,29†], Shaohua Chen[4,5†], Chunmeng Wei[1,2†], Jin Ji[6,7†], Yan Liu[8,9†], Leifeng Liang[10], Yifeng Chen[11], Xing Li[12], Lin Zhao[6], Xiaolei Shi[6], Yu Fang[6], Weimin Lu[13†], Tianman Li[14], Zhe Liu[15], Wenhao Lu[1,16], Tingting Li[8,9], Xiangui Hu[17], Mugan Li[18], Fuchen Liu[19], Xing He[20], Jiannan Wen[21], Zuheng Wang[2], Wenxuan Zhou[19], Zehui Chen[22], Yonggang Hong[23], Shaohua Zhang[23], Xiao Li[3], Rongbin Zhou[1,16], Linjian Mo[2], Duobing Zhang[13,24], Tianyu Li[2], Qingyun Zhang[4], Li Wang[25†], Xuedong Wei[26†], Bo Yang[6†], Shenglin Huang[27†], Huiyong Zhang[1†], Guijian Pang[11†], Liu Ouyang[17,28*], Zhenguang Wang[19*], Jiwen Cheng[2*], Bin Xu[29*] and Zengnan Mo[1,2*]

Abstract

Background Cancer remains a leading global cause of mortality, making early detection crucial for improving survival outcomes. The study aims to develop a machine learning-enabled blood-derived exosomal RNA profiling platform for multi-cancer detection and localization.

Methods In this multi-phase, multi-center study, we analyzed RNA from exosomes derived from peripheral blood plasma in 818 participants across eight cancer types during the discovery phase. Machine learning techniques were applied to identify potential pan-cancer biomarkers. During the screening and model validation phases, the sample size was progressively expanded to 1,385 participants in two steps, while the candidate biomarkers were refined into a set of 12 exosomal tumor RNA signatures (ETR.sig). In the subsequent model construction phase, diagnostic

†Fubo Wang, Chengbang Wang, Shaohua Chen, Chunmeng Wei, Jin Ji and Yan Liu contributed equally to this work.

†Weimin Lu, Li Wang, Xuedong Wei, Bo Yang, Shenglin Huang, Huiyong Zhang and Guijian Pang were Senior authors.

*Correspondence:
Fubo Wang
wangfubo@gxmu.edu.cn
Liu Ouyang
aqqwbjsqtb@163.com
Zhenguang Wang
wangzhenguang82@163.com
Jiwen Cheng
chengjiwen@stu.gxmu.edu.cn
Bin Xu
Zengnan Mo
mozengnan@gxmu.edu.cn
Full list of author information is available at the end of the article

models were developed using the expanded cohort and ETR.sig. Statistical analyses included the calculation of receiver operating characteristic (ROC) curves and AUC values to assess the models' ability to distinguish cancer cases from controls and determine tumor origins. To further validate and explore the biological relevance of the identified biomarkers, we integrated tissue RNA-seq, single-cell data, and clinical information.

Results Machine learning analysis initially identified 33 candidate biomarkers, which were narrowed down to 20 ETR.sig in the screening phase and 12 ETR.sig in the validation phase. In the model construction phase, a diagnostic model based on ETR.sig, built using the Random Forest (RF) algorithm, showed excellent performance with an AUC of 0.915 for distinguishing pan-cancer from controls. The multi-class classification model also demonstrated strong classification power, with macro-average and micro-average AUCs of 0.983 and 0.985, respectively, for differentiating between eight cancer types. Additionally, tumor origin classification using the RF-based diagnostic models achieved high AUC values: BRCA 0.976, COAD 0.98, KIRC 0.947, LIHC 0.967, LUAD 0.853, OV 0.972, PAAD 0.977, and PRAD 0.898. Integration of tissue RNA-seq, single-cell data, and clinical information revealed key associations between ETR.sig-related genes and tumor development.

Conclusions The study demonstrates the robust potential of exosomal RNA as a minimally invasive biomarker resource for cancer detection. The developed ETR.sig platform offers a promising tool for precision oncology and broad-spectrum cancer screening, integrating advanced computational models with nanoscale vesicle biology for accurate and rapid diagnosis.

Keywords Multi-cancer, Exosome, Exosomal tumor RNA signatures, Biomarker, Cancer detection, Cancer localization

Introduction

Cancer, a pervasive global health concern, stands as one of the prominent causes of death worldwide [1]. A comprehensive global study in 2020 highlighted breast, lung, and prostate cancers as the top three in terms of incidence, while lung cancer, liver cancer, and stomach cancer were the leading causes of mortality [1]. A U.S. survey projects over two million new cancer cases in 2024, resulting in an estimated 600,000 deaths [2]. Anticipated global cancer cases are projected to surge to around 28 million by 2040 [1]. underscoring the urgent need for holistic strategies in cancer prevention, diagnosis, and treatment.

Epidemiological studies indicate that patients diagnosed with cancer in early stages exhibit superior clinical outcomes compared to those diagnosed in advanced stages [2]. Consequently, there is a compelling need for early screening and diagnosis, propelling clinical scientists to relentlessly explore the domain of dependable cancer biomarkers [3, 4]. Despite progress in the scientific community, the current landscape of cancer diagnosis still relies heavily on combinations of laboratory tests and imaging techniques [5]. This dual-pronged approach, while providing valuable insights, not only imposes a financial burden on patients but also falls short of supplanting the irreplaceable role of pathological examinations in diagnostic modalities.

Extracellular vesicles, including exosomes, can serve as alternatives or supplements to other forms of liquid biopsy for enhancing diagnostic performance [6, 7]. Mechanistically, viable cells continuously release exosomes [8], which contain DNA, RNA, and protein components, thus providing clinically relevant diagnostic information. A pivotal study in 2007 first demonstrated the functional role of exosome mRNAs, emphasizing exosomes as crucial mediators of intercellular RNA transfer essential for cell-to-cell communication [9]. Subsequent research has further affirmed the presence of various RNA subtypes, including small RNAs, microRNAs, and long noncoding RNA (lncRNA), within exosomes [10–12]. Recent studies are increasingly utilizing comprehensive transcriptomic approaches, such as RNA sequencing (RNA-seq), to investigate the RNA content within exosomes in physiological and pathological processes, as observed in various samples, including cerebrospinal fluid (CSF) [13], plasma [14, 15], and urine [16, 17]. Over the years, we have been directed toward exploring RNA-seq to identify clinical biomarkers for cancer diagnosis using blood-derived exosomes. To date, studies have been conducted in prostate cancer and renal clear cell carcinoma [18–21]. However, there is still a lack of reliable pan-cancer biomarkers and corresponding markers for individual cancers [22]. Furthermore, there is a dearth of systematic algorithms and validation models available for assessing the diagnostic efficacy of the selected biomarkers.

In this study, we conducted a multiphase, multicenter investigation of the transcriptome profiles of blood-derived exosomes across eight cancer types: breast cancer (BRCA), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), ovarian cancer (OV), pancreatic adenocarcinoma (PAAD), and prostate adenocarcinoma (PRAD), along with a

healthy control group. During the discovery phase, we identified 33 candidate biomarkers relevant to multi-cancer detection using RNA sequencing (RNA-seq). In the screening phase, samples from nine centers were analyzed using TaqMan real-time quantitative PCR (qPCR). Among 245 participants, 13 biomarkers were excluded due to insufficient detection reliability. In the validation phase, the cohort was expanded to 1,385 participants, leading to the refinement and confirmation of 12 key blood-derived exosomal tumor RNA signatures (ETR.sig). In the model construction phase, we developed a multi-cancer detection and tumor-specific classification model based on ETR.sig. Additionally, we explored the tissue and cellular origins of the key genes in ETR.sig using data from TCGA cohorts and single-cell RNA sequencing (scRNA-seq) of LUAD and KIRC patients. This portion of the study aimed to investigate exosome-mediated RNA communication between blood and primary tumor sites, with the goal of enhancing early cancer detection and localization.

Methods

Study design

This study comprised four distinct phases: discovery, screening, validation and model construction. In the discovery phase, we investigated the expression profiles of exosomal RNAs (exoRNAs) in blood samples from patients diagnosed with BRCA, COAD, LIHC, OV, PAAD, LUAD, KIRC, and PRAD, as well as healthy controls. Exosomal RNA sequencing (exoRNA-seq) data were utilized for biomarker identification, leading to the selection of 33 candidate biomarkers. During the screening phase, we refined these biomarkers using TaqMan qPCR analysis on samples obtained from nine independent centers. Among the 245 participants, 13 biomarkers were excluded due to insufficient detection reliability, ensuring that only robust candidates were retained for further validation. In the validation phase, the cohort size was expanded to 1,385 participants (validation cohort) to enhance the robustness of biomarker selection. To refine the biomarker set, we applied exclusion criteria to eliminate: (1) genes with an undetectable rate exceeding 30% across samples, and (2) genes exhibiting expression patterns inconsistent with RNA-seq results. This rigorous selection process led to the identification of a refined panel of 12 key biomarkers, designated as ETR.sig. In the model construction phase, these biomarkers were then incorporated into both training and validation groups (validation cohort: 1,385 participants) for the development and evaluation of a multi-cancer diagnostic model (Study design: Fig. S1). Additionally, we explored the correlation between candidate biomarkers and clinical information using data from The Cancer Genome Atlas (TCGA). To further elucidate the potential cellular origins of these biomarkers, we analyzed single-cell transcriptomic data in the final stage of our study.

Multi-cancer blood-derived exoRNA-seq dataset description

We compiled a dataset comprising 818 profiles of blood-derived exoRNA-seq data across eight distinct cancer types, drawn from the exoRbase database and our in-house cohorts. Specifically, exoRbase provided samples of BRCA ($n=140$), COAD ($n=35$), LIHC ($n=112$), OV ($n=30$), and PAAD ($n=164$) patients. Additionally, exoRNA-seq data for LUAD ($n=83$), KIRC ($n=29$) [21], and PRAD ($n=31$) [18] were obtained from our previous studies available in CNGBdb under accession numbers CNP0005119, CNP0002099, and CNP0000926, respectively. The dataset encompassed a total of 194 healthy control samples (exoRbase = 118; CNP0005119 = 31; CNP0002099 = 28; CNP0000926 = 17), ensuring comprehensive and diverse representation across various cancer types.

Search for proto-oncogenes and tumor suppressor gene sets

All tumor suppressor genes were retrieved from Tumor Suppressor Gene Database (https://bioinfo.uth.edu/TSGene/download.cgi?csrt=13868158509745841593). Canonical drivers and candidate driver genes were obtained from the NCG database (http://network-cancer-genes.org/download.php, 2022.04) (Supplementary Table S1).

Processing pipeline for raw RNA-seq data

Quality control of the raw fastq data was performed using trim-galore (version 0.6.7). Subsequently, alignment was performed using STAR (version 2.7.9a) with the reference genome hg38 obtained from https://www.gencodegenes.org/human/. Sorting was carried out using SAMtools (version 1.6), and final gene expression quantification was performed using featureCounts (version 2.0.1).

Differential gene expression analysis

Identification of DEGs in the bulk RNA-seq data of blood-derived exosomes was performed, setting the criteria at $|\log2FC|>1$ and a P value <0.05. Subsequently, with the applied thresholds of a P value <0.05 and $|\log2FC|>1$, DEGs within the TCGA bulk RNA-seq cohort were determined utilizing the DESeq2 (version 1.32.0), limma (version 3.52.4), and edgeR (version 3.38.4) packages. Visualization of the differential expression analysis results was performed using ggplot2 (version 3.4.3) through bar charts, and differences in gene expression according to TaqMan qPCR data were visualized through violin plots.

Gene enrichment analysis

We performed functional enrichment analysis of DEGs using the KEGG database through the clusterProfiler package (version 4.8.3), with the results filtered by a significance threshold of $P < 0.05$. Additionally, we calculated the proportion of DEGs enriched in the first-level pathways of KEGG (metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases), as well as the intersection of pathways enriched by DEGs. Pie charts and bubble charts were visualized using ggplot2 (version 3.4.3).

Intersection and selection of key genes in blood-derived exosomal RNA sequencing (Exorna-seq)

We utilized the UpSetR package (version 1.4.0) for the intersection analysis and visualization of DEGs in blood-derived exoRNA-seq data across various tumors. To select a gene set suitable for multi-cancer diagnosis, we employed the following feature selection methods: we integrated multi-cancer exoRNA-seq data and conducted differential expression analysis for tumor vs. control tissues, filtering DEGs with a P value < 0.01 and a fold change (FC) ≥ 1.4 or ≤ -1.6. Subsequently, single-factor logistic regression analysis was applied to filter the genes for constructing diagnostic models, retaining those with P values < 0.05 (using the "glm" function from the R package stats, version 4.3.1). Additionally, LASSO analysis, with 1000 repeated tenfold cross-validations (CVs), was employed to refine the selection of the diagnostic gene set (based on the "cv.glmnet" function from the R package glmnet, version 4.1–8).

Moreover, we employed machine learning techniques to select the optimal gene set obtained from LASSO analysis. A stratified random sampling method was applied to partition the exoRNA-seq cohort into training and validation sets at a 7:3 ratio. The training set was utilized for constructing the random forest (RF) classification model, and the validation set was used to assess diagnostic performance. Model effectiveness was evaluated by calculating the area under the curve (AUC) values from receiver operating characteristic (ROC) curves. Finally, we identified the optimal gene set by comparing the AUC values of the models in the validation set. This entire procedure was implemented using the tidymodels (version 1.0.0) and pROC (version 1.18.0) R packages.

Participants in the Screening, Validation and Model Construction Phases

This study obtained approval from the Clinical Research Ethics Committees of Guangxi Medical University (Approval Numbers: GXMU2022-0154). Clinical samples were collected from the First Affiliated Hospital of Guangxi Medical University, Guangxi Medical University Cancer Hospital, Shanghai Changhai Hospital, the First People's Hospital of Yulin, Suzhou Municipal Hospital, the First Affiliated Hospital of Soochow University, Shanghai General Hospital, Eastern Hepatobiliary Surgery Hospital and Shanghai Ninth People's Hospital. Written informed consent was obtained from the participants before sampling. From June 2022 to September 2023, a total of 1385 participants were sequentially enrolled, including individuals with BRCA ($n = 128$), COAD ($n = 120$), KIRC ($n = 141$), LIHC ($n = 120$), LUAD ($n = 121$), OV ($n = 115$), PAAD ($n = 111$), PRAD ($n = 143$), and healthy controls ($n = 386$). Tumor cases were confirmed by surgical pathology and independently examined by two pathologists. The inclusion criteria for the patients were as follows: (1) age > 18 years; (2) definitive pathological diagnosis; (3) no prior anticancer treatment before blood sample collection; (4) no history of other cancers; (5) signed informed consent; and (6) No tumor metastasis, with T-stage less than T4. The inclusion criteria for the healthy controls were as follows: (1) age > 18 years and (2) underwent a health check-up and considered asymptomatic and healthy. The exclusion criteria were as follows: (1) previously diagnosed with tumors, (2) received ablation treatment, and (3) had other malignant tumors.

Sample collection and processing

The patients with masses in the breast, colon, kidney, liver, lung, ovary, pancreas, or prostate signed informed consent forms on the first day of admission, and their fasting peripheral blood was collected on the second morning. The healthy controls provided signed informed consent on the day of the health check-up, and fasting peripheral blood was collected before the physical examination. The samples were stored at 4 °C and transported to the laboratory on ice. All blood samples were then centrifuged at 1600 × g for 15 min to separate the plasma, which was collected in 1.5-ml centrifuge tubes and numbered. The plasma samples were immediately stored at − 80 °C until further processing.

Isolation and RNA extraction of blood-derived exosomes

Blood-derived exosomes were isolated, and RNA was extracted using exoRNeasy Midi/Maxi Kit. The specific steps were as follows. Buffer XBP was added to the sample at a 1:1 ratio. The mixture was gently mixed by inverting 5 times to ensure thorough mixing. The mixture was subsequently added to the exoEasy spin column and centrifuged at 500 × g for 1 min. (Note: If there was residual liquid on the membrane, the mixture was

further centrifuged at 5000×g for 1 min to ensure complete passage of all the liquid through the membrane.) Then, 3.5 ml of Buffer XWP was added, and the column was centrifuged at 5000×g for 1 min to wash the column and remove residual buffer. The filtrate and the bottom collection tube were discarded, and the spin column was transferred to a new collection tube. Then, 700 µl of QIAzol was added to the membrane of the column, which was centrifuged at 5000×g for 5 min to collect the lysate, which was then transferred completely to a 2 ml tube. The lysate in the tube was gently mixed and incubated at room temperature (15–25 °C) for 5 min. Then, 90 µl of chloroform was added, the tube was tightly capped, and the solution was mixed vigorously for 15 s. The tube was incubated at room temperature for 2–3 min. The tube was then placed in a precooled centrifuge at 4 °C and centrifuged at 12,000×g for 15 min. The supernatant was transferred to a new 2.0 mL tube (care was taken to avoid aspirating the organic phase). An equal volume of ethanol was added, and the mixture was inverted several times. Then, 700 µL of the mixture was added to the RNeasy MinElute spin column in a 2 mL collection tube and incubated for 2 min, after which the tube was capped. The column was centrifuged at 12,000×g for 1 min at room temperature, after which the filtrate was discarded. This step was repeated until all the mixture was used. Then, 700 µL of Buffer RWT was added to the RNeasy MinElute spin column, the tube was capped, the column was centrifuged at 12,000×g for 1 min at room temperature, and the filtrate was discarded. Five hundred microliters of Buffer RPE was added to the RNeasy MinElute spin column, the tube was capped; the column was centrifuged at 12,000×g for 1 min at room temperature, and the filtrate was discarded. Five hundred microliters of Buffer RPE was added to the RNeasy MinElute spin column, the tube was capped, the column was centrifuged at 12,000×g for 2 min at room temperature, and the filtrate was discarded. The RNeasy MinElute spin column was placed in a new 2.0 mL collection tube and centrifuged at 12,000×g for 5 min at room temperature, after which the filtrate and the bottom 2.0 mL collection tube were discarded. The RNeasy MinElute spin column was placed in a new 1.5 mL tube, and 14 µL of preheated RNase-free water was added to the center of the silica membrane. The tube was capped, and the column was centrifuged at 12,000×g for 1 min at room temperature. The RNeasy MinElute spin column was discarded, and the purified RNA solution was collected in a 1.5 mL tube.

Quality control of exosome isolation and verification

For isolated exosomes, we employed distinct techniques for identification, including nanoparticle tracking analysis (NTA) (Fig. S2A), transmission electron microscopy (TEM) (Fig. S2B), and western blotting (WB) (Fig. S2C) techniques. The detailed experimental methods can be found in our previous study [21].

Reverse transcription of blood-derived exosomal RNA

cDNA was generated from blood-derived exosomal RNA using Takara RR047A Kit through a two-step reverse transcription process, which included gDNA digestion and cDNA chain synthesis. The reaction components and procedures for each step were as follows. Step 1: gDNA digestion in a 13 µl reaction mixture, 2.0 µl of 5×gDNA Eraser Buffer, 1.0 µl of gDNA Eraser, and 10 µl of total RNA were gently mixed. The reaction was carried out in a PCR instrument with a program set at 42 °C for 2 min, followed by a hold at 4 °C. Step 2: Reverse transcription was performed. In a 20 µl reaction mixture, 13.0 µl of the reaction mixture from Step 1, 1.0 µl of PrimeScript RT Enzyme Mix I, 1.0 µl of RT Primer Mix, 4.0 µl of 5×PrimeScript Buffer 2 (for Real Time), and 1.0 µl of RNase-free dH2O were gently mixed. The reaction was carried out in a PCR instrument with a program set at 37 °C for 15 min, followed by a 5-s incubation at 85 °C and incubation at 4 °C.

Primer design and synthesis of detection targets

Target gene sequences were retrieved using the Ensembl database. The NCBI Primer designing tool website (https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi? LINK_LOC=BlastHome) was utilized to design qPCR primers (Supplementary Table S2). The primer design criteria included a melting temperature (Tm) ranging from 55–65 °C, a GC content ranging from 40%−60%, minimal mismatches, and good specificity. The primers used for synthesis were obtained from Sangon Biotech (Shanghai) Co., Ltd.

Primer validation

RNA was extracted from cancer/tumor adjacent tissues or cell lines and reverse transcribed into cDNA templates. Exosomal RNA from the samples to be validated was also prepared and reverse transcribed into cDNA. The melting temperature (Tm) values for each primer pair were analyzed, and 3–4 gradient annealing temperature values were set.

Using the cancer/tumor adjacent tissue or cell line cDNA as templates and diluted primers as amplification primers, PCR amplification was performed with the TB Green II enzyme using different gradient annealing temperatures (35 cycles). The PCR program was as follows: initial denaturation at 95 °C for 30 s; 35 cycles of denaturation at 95 °C for 5 s, annealing at 58 °C or 59 °C or 60 °C for 30 s, and extension at 72 °C for 15 s; and a final extension at 72 °C for 5 min. The reaction mixture consisted

of 2 µl of cDNA, 5 µl of 2×TB Green, 0.4 µM/0.4 µM Primer F/R (10 µM), and 2.2 µl of ddH2O.

The PCR products were mixed with DNA loading buffer, and their sizes were verified by 1.5% agarose gel electrophoresis to confirm the expected amplification product size and primer specificity and to determine the optimal annealing temperature. Primers with a single band of amplified product and the expected fragment size were selected. Using the cancer/tumor adjacent tissue or cell line cDNA as templates and blood-derived exosome cDNA as a template, TB Green® Premix Ex Taq™ dye-based qPCR was used to determine the melting curve of the primers. The PCR program was as follows: initial denaturation at 95 °C for 30 s; 40 cycles of denaturation at 95 °C for 5 s; and annealing at the optimal annealing temperature for 34 s. A melting curve was generated by increasing the temperature to 95 °C for 15 s, annealing at the optimal annealing temperature for 1 min, and then increasing the temperature to 95 °C for 15 s. Primers with normal melting curves were selected as candidates. The reaction system consisted of 2 µl cDNA, 5 µl 2×TB Green, 0.4 µM/0.4 µM Primer F/R (10 µM), 0.2 µl ROX Reference Dye II (50X), and 2.2 µl ddH2O. The primer design for key genes is detailed in Table S17.

Probe design and verification

When designing the predicted product sequence based on the reference primers, primers were positioned near the primer end without overlapping with the primer, and the Taqman probe was ensured to have a suitable Tm value and GC content while adhering to the design principles. DNAMAN was used to detect mismatches in the designed probes and primer mismatches. For combination detection, attention was given to the mismatch between two pairs of primers and two probes, and severe mismatches were avoided. Based on the results of the dye-based qPCR experiment, genes with similar Ct values were grouped, the fluorescent groups that did not interfere with each other were modified, and the probes were synthesized by Sangon Biotech (Shanghai) Co., Ltd. After centrifuging the Taqman probe, the sample was dissolved in DEPC water, and a 100 µM stock solution was prepared. The following system was used (reference system—50 µl of probe and primer as an example: 1. One gene per well: 50 µl of probe and primer, 45 µl of DEPC water, 2 µl of F primer, 2 µl of R primer, and 1 µl of probe; 2. Two genes per well (50 µl of probe and primer, 40 µl of DEPC water, 4 µl of F primer, 4 µl of R primer, 2 µl of probe) to prepare primer–probe mixtures. Using cancer-adjacent tissue cDNA as a template, the reaction system was prepared with ABI TaqPath™ ProAmp™ Master Mix and the primer–probe mix for qPCR amplification verification. The reaction system and program were as follows: Reaction system (2×Taqpath 5 µl, cDNA 2 µl, primer–probe mix for a single gene 2.5 µl/for two genes 3 µl, supplemented with H2O to 10 µl), Reaction program (50 °C 5 min; 95 °C 5 min; 10 cycles: 95 °C 15 s, 60 °C (varied with annealing temperature) 1 min; 95 °C 5 min; 40 cycles: 95 °C 15 s, 60 °C (varied with annealing temperature) 1 min). According to the following criteria, we determined whether the probe was suitable for subsequent detection: stable amplification in each duplicate well for single detection; a change in Ct value before and after combination not greater than 0.5; and no amplification or a significantly lower amplification efficiency compared to that of positive detection wells for both single and combined detection in negative controls. Probes that meet the criteria were used for subsequent sample detection.

Sample detection

ABI TaqPath™ ProAmp™ Master Mix reagent was used for multiple qPCR detection via the Taqman probe method. The preparation of the primer–probe mixture and reaction system, as well as the reaction program, were the same as those described in Sect. "Differential Gene Expression Analysis". After detection, the Ct values of the samples were recorded. The copy numbers of the target genes in the samples were calculated using the standard curve method. The specific steps were as follows: Synthetic standards for each gene were prepared and diluted at 104, 105, 106, 107, 108, and 109-fold dilutions. Using the diluted standards as templates, amplification was performed using the primers and probes validated in Sects. "Processing Pipeline for Raw RNA-seq Data" and "Differential Gene Expression Analysis", and the Ct values were recorded. A standard curve was established by plotting the copy numbers of the standards against their corresponding Ct values, and the equation was derived (see Supplementary Table S3 for standard curve statistics generated using the Applied Biosystems™ 7500 Real-Time PCR System; see Supplementary Table S4 for standard curve statistics generated using the Applied Biosystems™ StepOnePlus Real-Time PCR System). The Ct values obtained from the sample detection were used to calculate the copy number of each sample via the standard curve equation.

TaqMan qPCR model training and parameter tuning

A stratified random sampling method was applied to partition the blood-derived exosomal TaqMan qPCR data cohort into training and validation sets at an 8:2 ratio. Data normalization was carried out using the "log2" and "scale" functions in the base package of R (version 4.3.1); batch effect removal was implemented through the "ComBat" function in the sva package (version 3.48.0).

Employing the caret package (version 6.0–94), we utilized nine common machine learning (ML) algorithms, including Support Vector Machine ("svmRadialWeights," "svmRadial"), Naïve Bayes ("NB"), Random Forest ("RF"), k-nearest neighbors ("knn"), AdaBoost Classification Trees ("AdaBoost"), Boosted Logistic Regressions ("LogiBoost"), Linear LASSO Ridge ("Glmnet"), and Gradient Boosting Machine ("Gbm") [23]. A multi-cancer classification model was trained using ETR.sig, and hyperparameter tuning was performed with tenfold CV to optimize model performance. To ensure robustness, the optimization process was repeated 10 times with different random seeds for each resampling.

Additionally, we endeavored to develop a multi-class classification model distinguishing among eight tumor types using machine learning algorithms, including multilayer perceptron (mlp), support vector machines with linear kernels (svmLinear), RF, knn, and a classification and regression tree (rpart). The sample grouping and parameter tuning methods used were consistent with the binary variable methods mentioned above.

TaqMan qPCR model validation

We employed a diverse set of algorithms to train multi-cancer tumor diagnostic and multi-class classification models on the training set. These models were subsequently applied to the validation cohort, and the results were systematically compared. The model demonstrating superior performance was selected as the ultimate multi-cancer diagnostic model.

To refine the gene features tailored for the diagnosis of individual tumor types, we iteratively subjected each tumor type to single-factor logistic regression and LASSO analysis. The gene features showing the most promising outcomes were then utilized in machine learning methods to individually construct tumor diagnostic models for eight categories of blood-derived exosome TaqMan qPCR data. The diagnostic efficacy of these models for each specific tumor type was thoroughly evaluated. The pROC package (version 1.18.0) and ggplot2 (version 3.4.3) were used for visualizing ROC curves, providing a comprehensive illustration of the diagnostic performance of the model.

Collection of key gene tissue traceability datasets

To explore the possible tissue and cellular origins of key genes, TCGA bulk RNA-seq and scRNA-seq data were collected. TCGA bulk RNA-seq data were obtained for BRCA (tumor = 747, control = 273), COAD (tumor = 320, control = 348), KIRC (tumor = 446, control = 100), LIHC (tumor = 366, control = 160), LUAD (tumor = 351, control = 347), OV (tumor = 381, control = 88), PAAD (tumor = 56, control = 171), and PRAD (tumor = 421, control = 152) patients. scRNA-seq datasets covered two cancer types. For KIRC, we included control samples ($n = 9$) from the GEO database with accession numbers GSE131685, GSE152938, and GSE156632. Additionally, peripheral blood mononuclear cell (PBMC) samples from KIRC patients ($n = 7$) and control volunteers ($n = 10$) were obtained from the Gene Expression Omnibus (GEO) database under accession numbers GSE121636, GSE139555, GSE139324, and GSE155698. Brain metastasis samples ($n = 1$) from KIRC patients were retrieved from the GEO database (accession number GSE186344) [24–31]. For LUAD, the datasets included 25 cancer samples and 15 control samples from the GEO database with accession IDs GSE146100, GSE131907, and GSE123902. Similarly, seven lymph node metastasis samples and ten control lymph node samples were obtained from the GEO database under accession number GSE131907. Brain metastasis samples ($n = 13$) under accession IDs GSE123902, GSE131907, adrenal metastasis samples ($n = 1$) under accession ID GSE123902, bone metastasis samples ($n = 1$) under accession ID GSE123902, and pleural effusion samples ($n = 5$) under accession IDs GSE123902, GSE131907 [27, 32, 33]. This comprehensive compilation of scRNA-seq datasets offers a thorough representation of various cancer states and metastatic conditions.

Single-cell data processing

The obtained fastq files were processed through Cell Ranger (version 6.1.2, 10×Genomics) with default parameters, and alignment was carried out against the 10×human transcriptome GRCh38-2020 (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest). Subsequent analysis of the single-cell data was performed using Seurat (version 4.4.0). Distinct strategies for eliminating low-quality cells have been implemented tailored to specific types of tumors. For various tumor types, distinct methods were employed to eliminate low-quality cells. Specifically, in the case of KIRC, mitochondrial genes (≤30%), UMIs, and gene counts (<25,000 and ranging from 400 to 5,000) were applied. For LUAD, mitochondrial genes (≤20%) and gene counts (ranging from 200 to 10,000) were used. Normalization and dimensionality reduction procedures were executed through the SCTransform, RunPCA, and RunUMAP functions [34]. Identification of cellular identities was performed with the SingleR package (version 2.2.0). The FindAllMarkers function of the Seurat package was subsequently used to discern marker genes specific to each cell subpopulation, ultimately through use of previously published cell markers (http://xteam.xbio.top/CellMarker/). In the process of identifying tumor cells, the unique molecular identifier (UMI) count matrix

was utilized as input to infer chromosomal copy number alteration (CNA) profiles. This analysis was conducted using the CopyKAT (version 1.1.0) [35].

Clinical correlation and survival analysis

We used GEPIA2.0 (http://gepia2.cancer-pku.cn/#index, accessed on July 7, 2023), a data visualization platform for the TCGA database, to assess the impact of ETR.sig on overall survival (OS) across 15 cancer datasets (BLCA, BRCA, CESC, COAD, ESCA, HNSC, KIRC, KIRP, LIHC, LUSC, OV, PAAD, PEAD, SKCM, and STAD). K–M survival curves were generated, and correlations between candidate biomarker genes and clinical indicators were examined. Selection of the prognostic gene set for tumor evaluation involved the application of the "Surv" function from the survival package (version 3.5–7) for Cox proportional hazards regression (Cox) analysis and single-factor logistic regression analysis. Genes with a significance level of $P < 0.05$ in these analyses were subjected to further analysis. Gene set scores were computed using the "gsva" function of the GSVA package (version 1.50.0). Tumor patients were stratified into high- and low-score groups based on the median score, and the "survdiff" function from the survival package (version 3.5–7) was used to assess the correlation between high- and low-score groups and tumor prognosis. A significance level of $P < 0.05$ was considered indicative of a significant difference in survival between high- and low-score patients. Survival analysis was performed using the "ggsurvplot" function of the Survminer package (version 0.4.9).

Statistics

Measurement data were tested for normality and variance homogeneity, and the independent samples that met the normal distribution were tested by a t-test in a group design (the t-test was used for variance nonhomogeneity). The Mann–Whitney U test was used if they did not meet the normal distribution. Paired data were tested for normality of the mean difference using the paired t-test for normal distribution and the Wilcoxon signed-rank test for nonnormality. All statistical analyses were conducted using the R language (version 4.3.1), with a significance level set at $P < 0.05$. Principal component analysis (PCA) dimensionality reduction was performed using the "SamplePCA" function fdrom the ClassDiscovery package (version 3.4.0). scRNA-seq data were scored with the AUCell package (version 1.24.0), while bulk RNA-seq gene set scoring was performed with the GSVA package (version 1.50.0). Additional information on the statistical tools, methods, and thresholds used is provided in the Methods section.

Results

Patient characteristics

Recent investigations have highlighted the potential of blood-derived exosomes as promising candidates for liquid biopsy-based diagnostic approaches for cancer [12, 36]. In this context, we conducted a study that included eight distinct cancer types and a healthy control group (Study design: Fig. S1). This study included a discovery cohort analyzed through blood-derived exoRNA-seq, and the sample statistics are detailed in Supplementary Table S5. Additionally, screening cohort and model construction cohort (training and validation groups) were included, which encompassed nine medical centers in total, each accompanied by demographic and clinical characteristics, as outlined in Supplementary Table S6. The participants in these groups had an average age of 60.2 ± 10.4 years (cancer group: 59.9 ± 10.3 years, healthy control: 61.1 ± 10.5 years), with a sex distribution of 807 males and 578 females (cancer group: male 539/female 460, healthy control: male 268/female 118). The distribution of tumor stages was 543 cases of T1, 324 cases of T2, and 132 cases of T3 (54.4%, 32.4%, and 13.2%, respectively). Importantly, there were no significant differences in sex or age distribution between the cancer group and the healthy control group (Supplementary Table S6). These balanced demographic and clinical characteristics help ensure that any observed differences in exoRNA profiles are likely attributed to disease status rather than confounding factors such as age or sex.

Exploring blood-derived exoRNA-seq profiles for multi-cancer biomarker identification

In the discovery phase, we compiled a dataset comprising 818 blood-derived exoRNA-seq profiles across eight distinct cancer types (Fig. 1A) from the exoRbase database [37] and our in-house cohorts, with sample details provided in the Methods section. Analyses of the datasets revealed the distribution of cancer and control samples of each tumor type, and PCA suggested limited ability to distinguish between groups (Fig. 1B), which may be partially affected by heterogeneity and confounding factors in blood. Furthermore, we identified differentially expressed genes (DEGs) between cancer and control samples within eight distinct tumor types utilizing their respective transcriptomic expression profiles from exoRNA-seq data (Supplementary Table S7). Bar plots in Fig. 1C enumerate the quantities of up- and downregulated DEGs from the analysis, along with pie plots categorizing these DEGs based on KEGG pathways. Notably, PRAD exhibited the highest number of DEGs ($n = 2365$) in comparative analysis

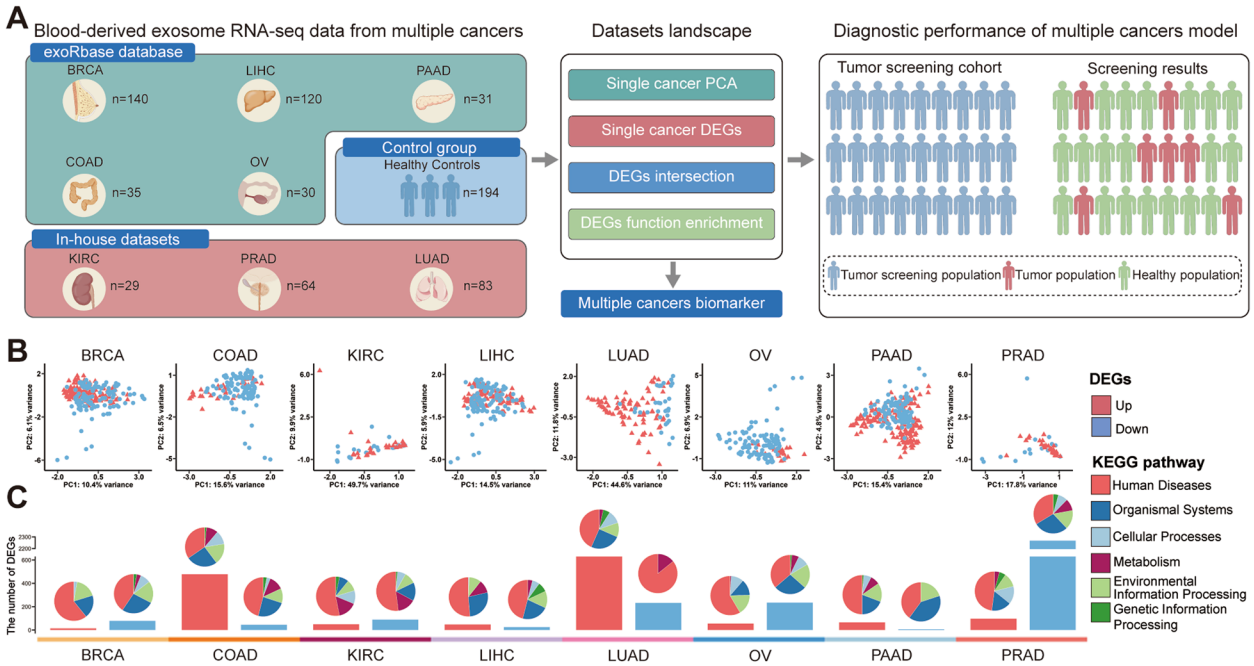

Fig. 1 Exploring blood-derived exoRNA-seq profiles. **A** Schematic diagram of identification of multi-cancer biomarkers using exoRNA-seq data. **B** PCA illustrating the distribution characteristics of blood-derived exoRNA-seq samples from the cancer and control groups across eight cancer types. Cancer samples are denoted in red, while control samples are represented in blue. **C** Bar plots depicting the exact number of DEGs between cancer and control samples in eight categories of blood-derived exoRNA-seq. The red bars indicate the number of upregulated DEGs in cancer samples relative to control samples, the blue bars represent downregulated DEGs, and the accompanying pie plots above the bars represent the KEGG pathways enriched by each group of DEGs

between samples, while LIHC displayed the fewest DEGs ($n=73$). KEGG enrichment analysis of these DEGs highlighted significant associations with "human diseases," underscoring the potential relevance of these genes in cancer biology.

Next, we systematically conducted intersection analyses of the upregulated (Fig. 2A) and downregulated (Fig. 2B) DEGs obtained from blood-derived exoRNA-seq across the eight tumor types. However, no genes were consistently identified as either up- or downregulated across all tumor types, suggesting a high degree of specificity in the transcriptomic profiles of different cancers. Further KEGG enrichment analyses of the upregulated (Fig. 2C, Supplementary Table S8) and downregulated DEGs (Fig. 2D, Supplementary Table S9) revealed significant enrichment in pathways such as “cytokine-cytokine receptor interaction”, “PI3K-Akt signaling pathway”, and “ECM-receptor interaction”, which are well-established as being involved in cancer-related processes. This result emphasizes the essential stage of exosome-mediated RNA communication in tumorigenesis and guides our approach to the development of exosome-based molecular diagnostic avenues, aligning with previous research findings [38].

Refining candidate multi-cancer diagnostic genes using exoRNA-seq profiles

To identify multi-cancer diagnostic genes and establish a comprehensive multi-cancer diagnostic model, we integrated the above data to further identify DEGs that distinguish multi-cancer patients from healthy controls (Supplementary Table S10). Initially, we screened DEGs through strict quality control criteria and identified 44 genes significantly associated with cancer compared to control samples via univariate logistic regression analysis (Supplementary Table S11). Next, these 44 genes were selected for LASSO analysis, which was iterated 1000 times with tenfold cross-validation, ultimately yielding six sets of candidate gene panels (the number of genes ranged from 30 to 40; Supplementary Table S12).

To evaluate the diagnostic potential of the candidate gene panels, we employed a random forest (RF) machine learning algorithm and trained these candidate gene panels using exoRNA-seq data. The data were divided into training ($n=569$) and validation ($n=249$) sets at a 7:3 ratio for model training and diagnostic performance evaluation. To enhance model robustness, a fivefold repetition method was applied three times for parameter tuning. The results revealed that the area under the curve (AUC)

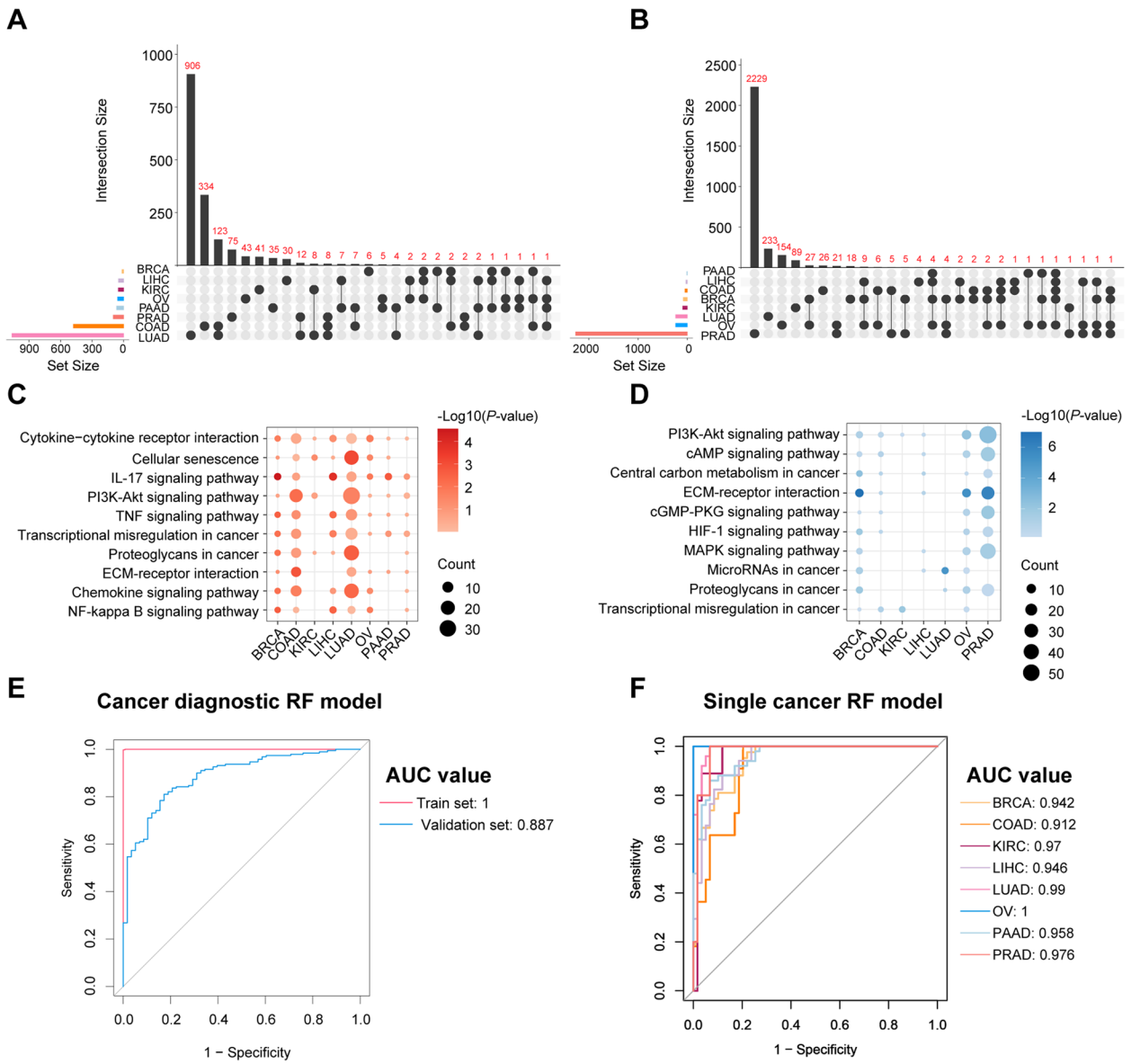

Fig. 2 Identification of multi-cancer biomarkers through blood-derived exoRNA-seq. **A**, **B** UpSet plots displaying the intersection analysis of up- and downregulated DEGs in the exoRNA-seq data across eight cancer types. **C**, **D** Bubble plots showing the top ten KEGG pathways enriched in up- and downregulated DEGs in the exoRNA-seq data across eight cancer types. **E** ROC curves evaluating the diagnostic performance of 33 key genes in distinguishing multi-cancer samples from control samples. **F** ROC curves were used to assess the diagnostic efficacy of 33 key genes in discriminating single cancer types from the respective control group

values for the six candidate gene panels in the training set were all optimal at 1 and consistently exceeded 0.85 in the validation set (Supplementary Table S12).

In particular, the diagnostic gene panel containing 33 key genes (ADAMTS5, AGO2, ALB, ANKRD36B, CKS2, CXCR1, CYSTM1, DNHD1, DOK4, DYNLL1, FCER1G, GPX4, KCNH2, KRT18, LCN2, MALAT1, MAN1A2, NKTR, NT5DC2, PPDPF, PROK2, RAB32, S100A8, S100A9, SLC24A4, SLC9A3R2, SMARCA5, SOCS3, TEAD4, TGFB3, TTN, UBE2Q2, and VCAN) demonstrated robust performance, achieving an AUC of 0.887 in the validation set (Fig. 2E). To mitigate sampling errors, all samples were randomly grouped ten times at ratios ranging from 0.1 to 0.9, and the diagnostic performance of the constructed model was validated (Fig. S3A).

Further evaluation of the model's performance in single cancer types revealed AUC values consistently exceeding 0.95 (Fig. S3B), suggesting strong diagnostic potential

across individual cancer types. To develop cancer-specific diagnostic models, we constructed independent diagnostic models for each of the eight cancer types using the 33 key genes. The same univariate logistic regression and LASSO analysis selection strategy was applied. Receiver operating characteristic (ROC) curves for the diagnostic models of diverse cancer types are shown in Fig. 2F, with AUC values as follows: BRCA (0.942), COAD (0.912), KIRC (0.97), LIHC (0.946), LUAD (0.99), OV (1), PAAD (0.958), and PRAD (0.976). The genes included for each tumor type are listed in Supplementary Table S13.

Screening and validation of 33 key blood-derived exoRNAs via Taqman qPCR assays

Nucleic acid sequencing technology, renowned for its high throughput and sensitivity, often comes with drawbacks like high costs and long detection cycles. In contrast, qPCR offers a simpler, more sensitive, and efficient alternative. In the screening and validation phase, to validate the expression profiles of the diagnostic model constructed, we used TaqMan qPCR assay, ensuring its practicality and reliability in real-world clinical applications. The experimental workflow, as outlined in Fig. 3A and Fig. S1.

Initially, a small sample set (screening cohort: $N=245$) was used to assess the expression of 33 key genes (with experimentally validated probes). We excluded 13 genes that could not be effectively detected (most samples had CT = 40). Subsequently, we expanded the screening cohort to a total of 1,385 samples, which served as the validation cohort for detecting the remaining 20 candidate RNAs. During this process, biomarkers were further filtered based on the following criteria: 1. Genes with CT values equal to 40 in more than 30% of the samples; 2. Genes that showed differential expression in the multi-cancer versus healthy control comparison, where the direction of change (up or down-regulation) was opposite to the exoRbase RNA-seq DEGs results (Fig. S1). Ultimately, 12 key RNAs were retained, namely ALB, FCER1G, KRT18, LCN2, PPDPF, SLC9A3R2, AGO2, CKS2, MALAT1, RAB32, S100A9, and UBE2Q2. These genes were defined as exosomal tumor RNA signatures (ETR.sig).

Differential expression analysis revealed significant differences in the expression patterns of ALB, RAB32, KRT18, LCN2, and UBE2Q2 within ETR.sig between the cancer and control groups in the multi-cancer dataset (Fig. 3B, Supplementary Table S14). Also, differences in expression of ETR.sig among diverse cancer origins varied significantly compared to that in the control group (Supplementary Table S14). These findings demonstrate the potential of ETR.sig for multi-cancer detection and origin determination, highlighting its diagnostic potential in clinical settings.

Detecting and locating multiple human cancers by ETR.Sig with high accuracy

In the model construction phase, the multi-cancer dataset (validation cohort) was randomly divided into training ($n=1109$) and validation ($n=276$) sets at an 8:2 ratio to validate the diagnostic performance of ETR.sig in diagnosing multi-cancer. Utilizing the validation set, nine different machine learning algorithms were employed to train the diagnostic model, and each diagnostic model underwent ten-fold cross-validation with five repetitions for parameter optimization. A diagnostic model constructed using different algorithms consistently demonstrated strong diagnostic performance across cancers, with the RF algorithm exhibiting superior performance and achieving an AUC value of 0.915 in the validation set (Fig. 4A, Supplementary Table S15).

After discriminating between cancer patients and control samples, the crucial step in clinical practice is to assist physicians in precisely determining the specific cancer origin. Subsequently, the dataset was similarly partitioned into training and validation sets at an 8:2 ratio. We created a sophisticated multi-class classification model to differentiate between eight cancer types. This involved employing five machine learning algorithms and conducting five repeated ten-fold cross-validation steps to optimize parameters for each model. Following training, we individually assessed their classification effectiveness (Supplementary Table S16). The results revealed that the multi-class classification model constructed by RF demonstrated optimal performance with an AUC of 0.983, accuracy of 0.848, and kappa of 0.827. Subsequently, we utilized the micro-/macro-average and one-versus-all methods to assess the performance of the RF model in predicting the validation set (Fig. 4B). The results indicated that the multiclass diagnostic model constructed by RF effectively differentiated between the eight cancer types, with all AUC values exceeding 0.95.

Additionally, we applied the RF machine learning algorithm to enhance the diagnostic credibility for a single type of cancer. Following univariate logistic regression and LASSO analysis selection of ETR.sig for each cancer (genes selected for each cancer type are detailed in Supplementary Table S17), we independently trained diagnostic models for classification of cancer and control samples. The AUC values for individual tumor diagnosis were as follows: BRCA 0.976, COAD 0.98, KIRC 0.947, LIHC 0.967, LUAD 0.853, OV 0.972, PAAD 0.977, and PRAD 0.898 (Fig. 4C).

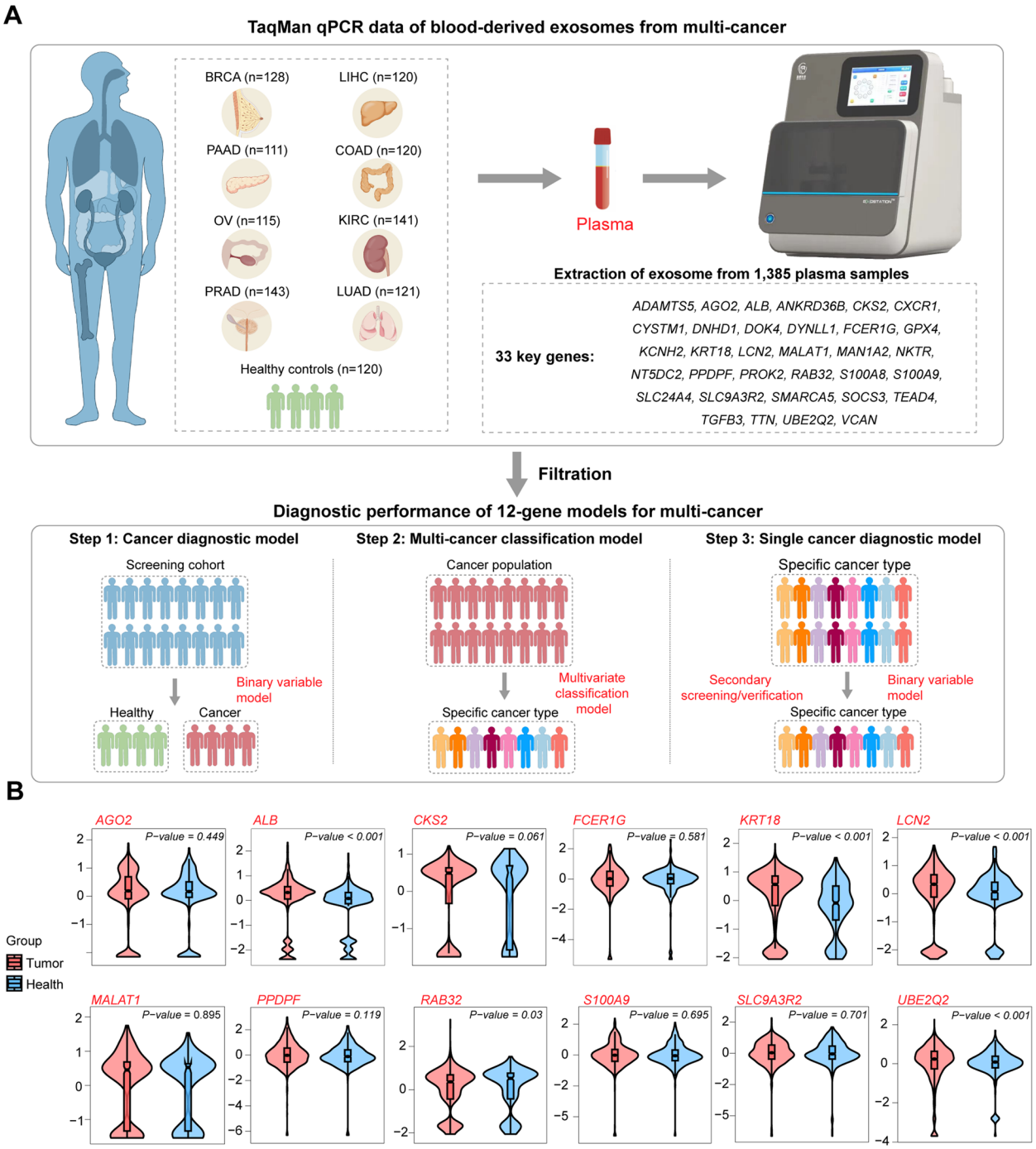

Fig. 3 Detection of key genes in ETR.sig based on TaqMan qPCR technology. **A** Schematic diagram of cancer types, experimental methods, and analysis procedures involved in constructing an TaqMan qPCR-based blood-derived exoRNA diagnostic model. **B** Violin plots comparing the standardized expression levels of key genes in ETR.sig between the multi-cancer group and healthy control group, with red indicating the multi-cancer group and blue the healthy control group

Tissue origins and cellular tracing of key genes in ETR.Sig

Release of exosomes plays a crucial role in the TME and cancer progression [39]. Understanding the tissue and cellular origins of the key genes within ETR.sig is essential for elucidating their roles in cancer development. To explore these origins, we conducted a comprehensive analysis of the gene expression profiles of the key genes in ETR.sig using RNA-seq data from the TCGA

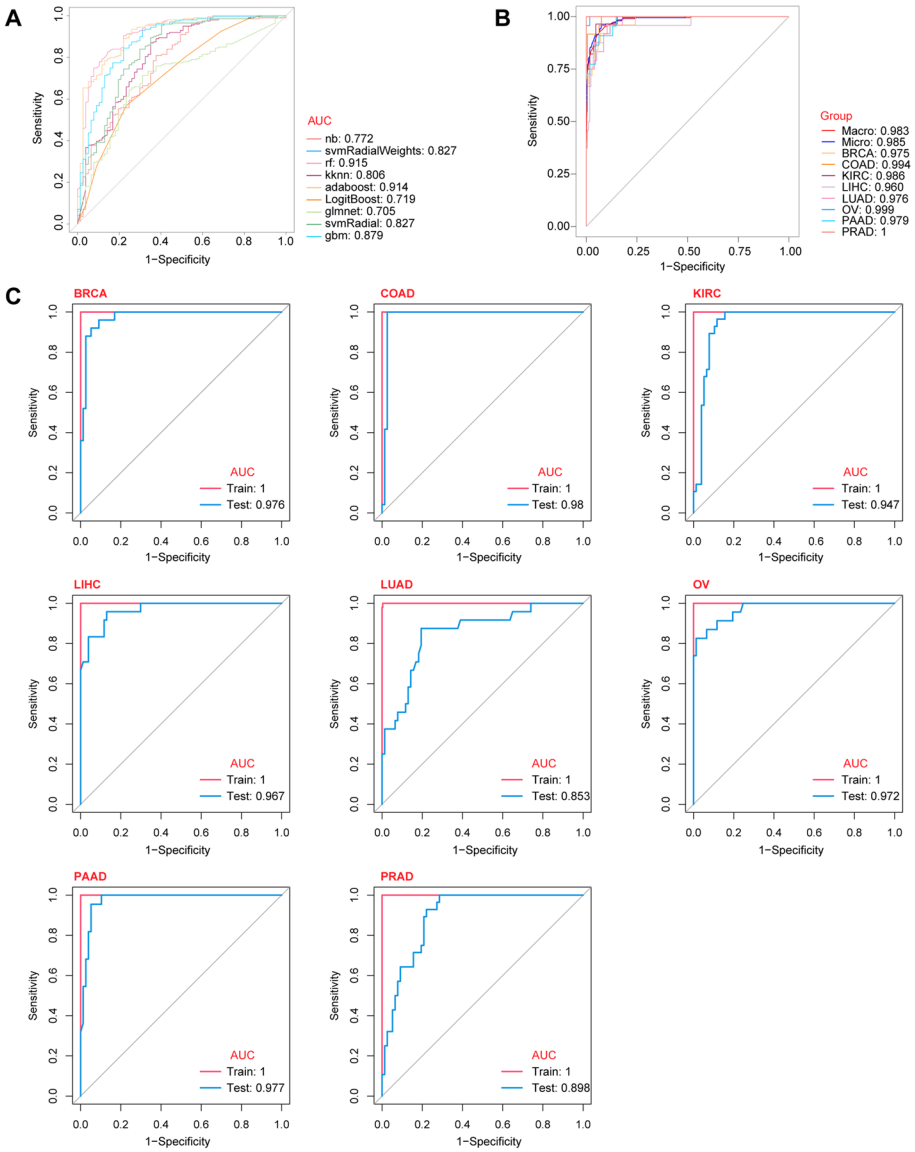

Fig. 4 Constructing a TaqMan qPCR-based blood-derived exoRNA diagnostic model. **A** ROC curves illustrating the performance comparison of nine different machine learning algorithms in the validation set. **B** ROC curves depicting the classification efficacy of the multivariate classification model constructed by the RF algorithm in the validation set. **C** ROC curves illustrating the performance of single-cancer diagnostic models in the validation cohort

database, blood-derived exoRNA-seq data, and blood-derived exosomal TaqMan qPCR data across eight cancer types.

The heatmap illustrates notable heterogeneity in the gene expression patterns of ETR.sig between blood-derived exoRNA-seq and TaqMan qPCR data (Fig. 5A), reflecting the differences between the three platforms. Additionally, the majority of key genes in the ETR.sig cohort exhibited upregulated expression according to tissue RNA-seq data, with the exception of ALB, which emerged as a potential tumor driver based on searches of cancer-promoting and -suppressing gene databases (TSGene and NCG Database; Supplementary Table S1). In addition, ETR.sig correlated significantly with tumor stage and survival outcome, with most key genes being primarily associated with advanced tumor stage and worse prognosis (Fig. 5B, Fig. S4-5). Next, enrichment of ETR.sig in a single sample was calculated to obtain the ETR.sig score for each patient, and survival analysis indicated poorer prognoses in the TCGA-COAD, TCGA-KIRC, TCGA-LIHC, PAAD, and TCGA-BRCA cohorts (Fig. 5C).

To further explore the potential secretion sources of key genes in ETR.sig at the cellular level, we downloaded scRNA datasets of KIRC tissues, adjacent normal tissues, PBMCs from KIRC cases, normal PBMCs from healthy controls, and brain metastasis samples from KIRC patients from the GEO database. For LUAD, the scRNA-seq datasets included LUAD tissue, adjacent normal lung tissue, lymph node (LN) metastasis sample from LUAD patients, normal LN tissue, brain metastasis sample from LUAD patients, adrenal metastasis sample from LUAD patients, bone metastasis sample from LUAD patients, and pleural effusion sample from LUAD patients (the sample and cell number counts are shown in Supplementary Table S18).

Following stringent quality control and identification procedures, we identified 16 and 9 distinct cell types in all KIRC and LUAD scRNA-seq datasets, respectively (Fig. 6A, C, Fig. S6-7). Based on differential expression of key genes within ETR.sig in blood-derived exosomes determined via TaqMan qPCR in KIRC and LUAD samples, we performed AUCell analysis of scRNA–seq datasets. The results indicated increased ETR.sig scores for macrophages, epithelial cells, and fibroblasts in these two cancer types (Fig. 6B, D). To further understand the expression patterns of key genes in ETR.sig across different cell types, we analyzed scRNA-seq data and visualized the expression disparities between tumor and control groups (Fig. S8-10, Supplementary Table S19). Bubble plots were generated to visualize the differential expression profiles of key genes in the ETR.sig gene set in the different datasets between the groups. Intriguingly, we observed increased expression levels of key genes in the ETR.sig population in macrophages in the KIRC TME. According to the LUAD scRNA-seq data, predominant differences in expression between the tumor and control groups were observed for epithelial cells, ciliated cells, and fibroblasts (Fig. S10).

Discussion

Through comprehensive analysis of large-scale exoRNA-seq data from 818 patients encompassing BRCA, COAD, KIRC, LIHC, LUAD, OV, PAAD, and PRAD patients and from a healthy control group, we sought to unveil the biological features of these patients and identify potential exosome-based biomarkers. Differential gene expression, as observed through between-group comparisons of blood-derived exoRNA data, revealed significant enrichment of various pathways encompassing cytokine-cytokine receptor interaction, chemokine signaling, IL-17 signaling, PI3K-Akt signaling, TNF signaling, and ECM receptor interaction, undeniably affirming the highly interactive nature of diverse cellular identities and emphasizing the pivotal role of exosomes in orchestrating the TME [40, 41].

Consistent with our findings, recent literature has consistently highlighted the crucial role of exosomes in driving tumor growth. These reports suggest that tumors create a favorable growth environment by releasing exosomes, fostering processes such as angiogenesis, inflammation, and immune suppression, which have been identified as facilitators of tumor progression [42, 43]. Notably, Wu et al. indicated that a stiff ECM could stimulate release of exosomes from cancer cells, consequently promoting tumor growth through activation of the Notch signaling pathway [38]. In summary, the substantial release of exosomes into the TME underscores the potential value of blood-derived exoRNAs in pan-cancer diagnosis.

Despite the easily accessible nature of blood samples, the mixed content compromises sensitivity and specificity as cancer biomarkers. In our study, PCA revealed that the overall gene expression profile of blood-derived exoRNAs performed suboptimally in distinguishing cancer samples from controls. Individual exoRNAs also failed to achieve ideal results in distinguishing multi-cancer patients from control individuals, suggesting that exoRNAs in the vast biological pool of blood are heterogeneous and prone to be affected by confounding factors in blood. Subsequent research revealed that the machine learning-based approach for screening potential biomarkers and constructing multi-cancer diagnostic model is more advantageous than use of a single exoRNA. Similarly, Hoshino

(See figure on next page.)

Fig. 5 Clinical correlation of key genes in ETR.sig. **A** Heatmap illustrating differences in expression of key genes within the ETR.sig across eight cancer types, utilizing blood-derived exosome RT–qPCR data, blood-derived exoRNA-seq, and TCGA RNA-seq data for comparisons between cancer and control samples. In the heatmap, red denotes significantly upregulated key genes, blue indicates significantly downregulated key genes, and white signifies no significant difference. The middle numbers represent |Log2FC| values. Gene types are categorized within ETR.sig as cancer candidate driver genes (CCG) in orange and non-TSG or POG genes in purple (genes that are neither tumor-promoting nor tumor-suppressing). The bar chart above depicts the proportion of gene types with differential expression in each column. The gene number ratio indicates the proportion of upregulated (in red), downregulated (in blue), or unchanged (in gray) genes in each column or row. **B** Bubble plot illustrating the correlation of key genes within the ETR.sig cohort with clinical tumor stage and survival based on the TCGA cohort. The X-axis represents cancer types in the TCGA cohort, while the Y-axis represents key genes in the ETR.sig cohort. Pink indicates a positive correlation between increased gene expression and advanced tumor stage or worse clinical prognosis, while blue indicates the opposite correlation. All presented results were significant at $P < 0.05$. **C** K–M curve plot: the prognostic value of the overall survival rate across cancers according to the GSVA estimation of the ETR.sig score

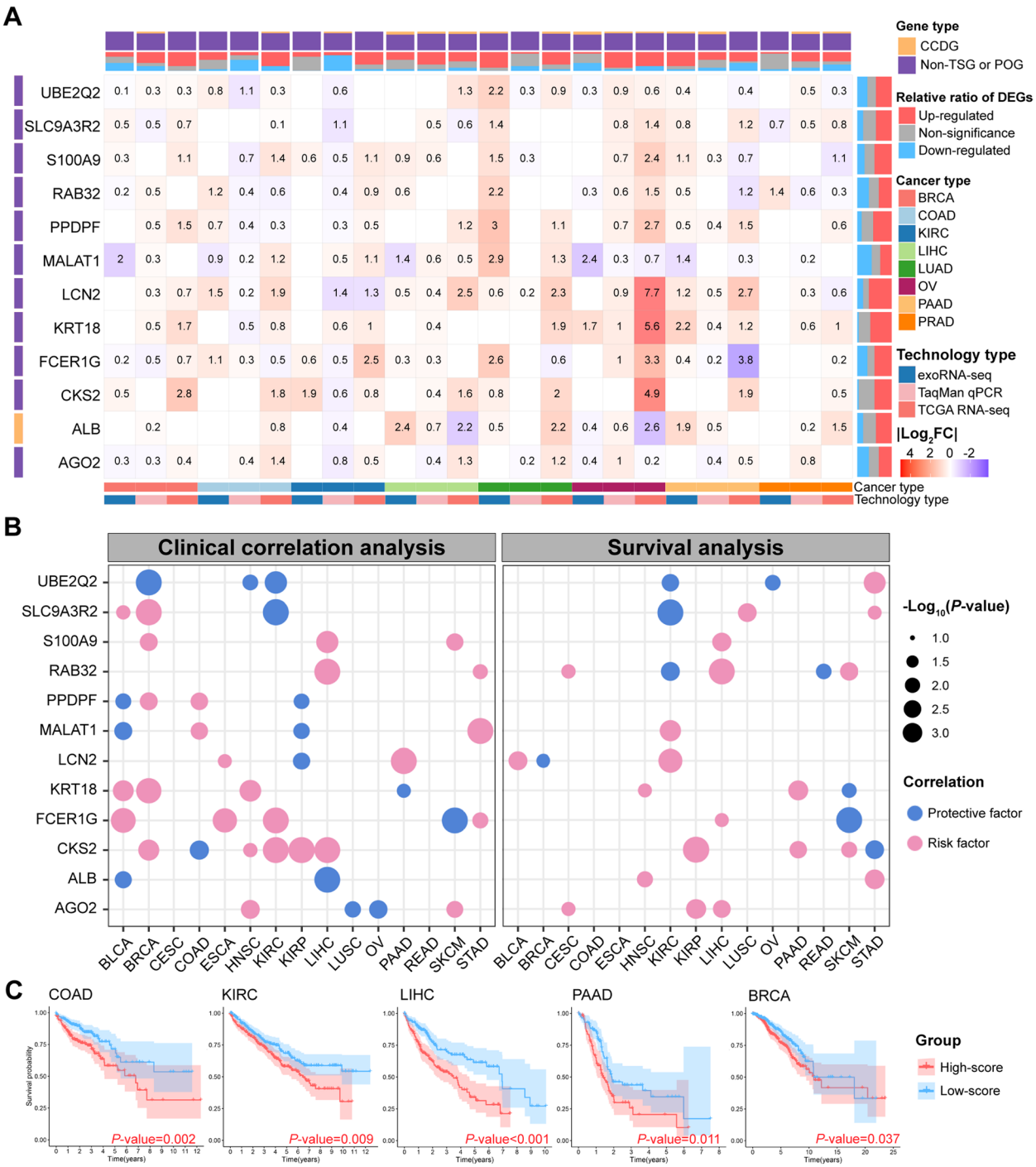

Fig. 5 (See legend on previous page.)

et al. constructed pan-cancer diagnostic model based on extracellular vesicles and particles using machine learning, demonstrating its reliability as a biomarker for pan-cancer detection and classification [44], it also outperformed in diagnosis of single cancer types [45, 46]. After filtering of candidate key genes via RNA-seq and subsequent TaqMan qPCR, we constructed ETR.sig for multi-cancer/single cancer type diagnosis; this gene set included 12 key genes: ALB, FCER1G, KRT18, LCN2, PPDPF, SLC9A3R2, AGO2, CKS2, MALAT1, RAB32, S100A9, and UBE2Q2. The multi-cancer vs. control group classification model (ROC: 0.915), multi-cancer classification model (Macro ROC: 0.983, Micro

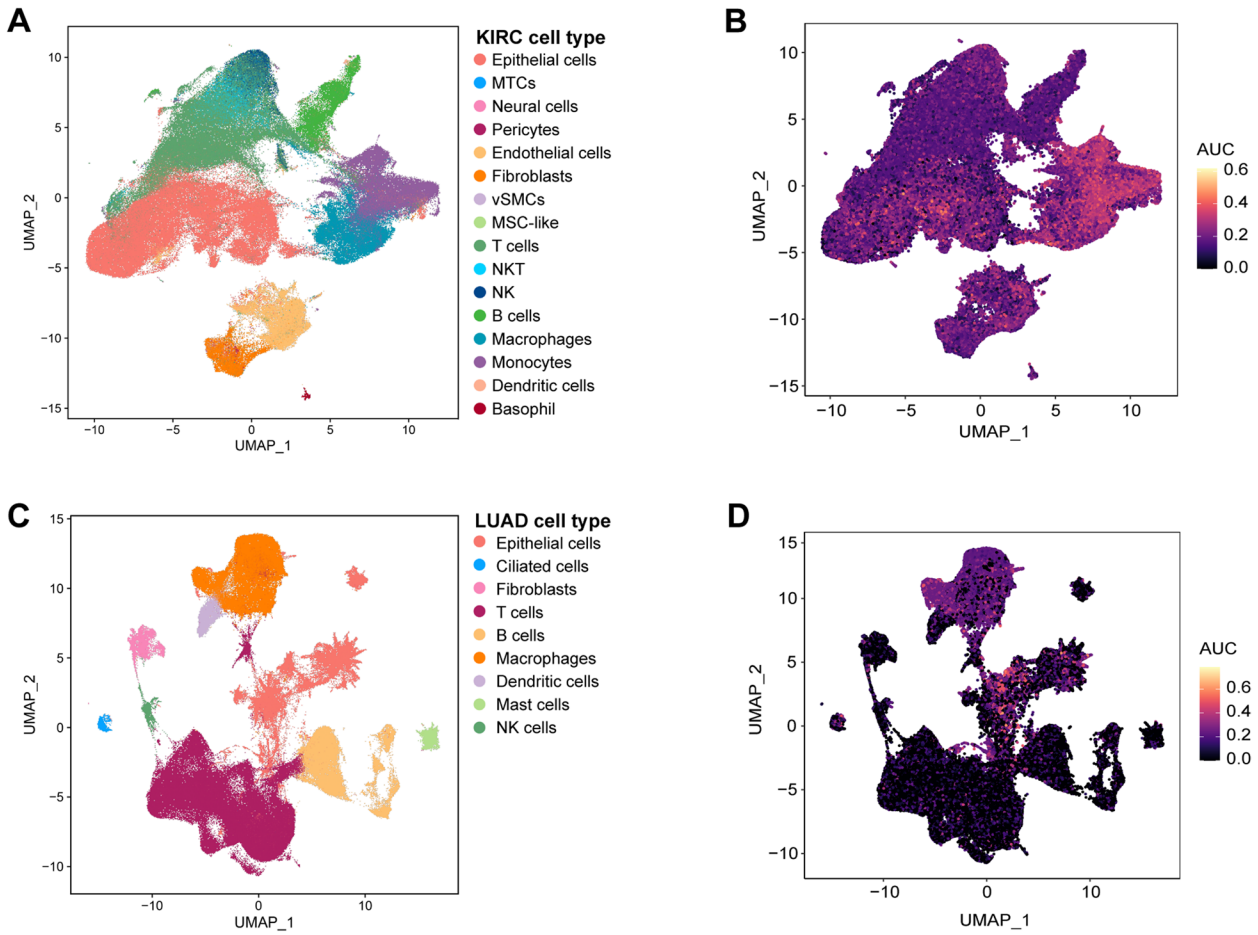

Fig. 6 Tissue origins of key genes in ETR.sig. **A** UMAP visualization depicting the major cell clusters in the KIRC scRNA-seq dataset. **B** UMAP plot illustrating the ETR.sig-related AUCell score in KIRC scRNA-seq datasets. **C** UMAP plot displaying the major cell clusters in the LUAD scRNA-seq dataset. **D** UMAP plot showing the ETR.sig-related AUCell score in LUAD scRNA-seq datasets

ROC: 0.985), and single cancer vs. all (ROC: > 0.853) built on ETR.sig-related genes demonstrated robust performance.

Mechanistically, compared with normal cells, cancer cells secrete more exosomes, a phenomenon intricately connected to the influence of the TME. *Sun* proposed that exosomal ADAM-17, a catabolic integrin and metalloproteinase derived from colon cancer cells, plays a central role in promoting cancer metastasis. This occurs through cleavage of E-cadherin junctions and active participation in formation of the premetastatic niche [47]. Additionally, owing to their compatibility with biological systems [48], exosomes are harnessed as natural drug delivery vehicles that can effectively transport various therapeutics, including genetic material, leveraging their inherent ability to deliver therapeutic cargo into cells [49, 50]. Intriguingly, key genes in the ETR.sig cohort are invariably associated with the occurrence, development, metastasis and prognosis of tumors. In our study, ETR.sig demonstrated excellent performance in multi-cancer diagnosis, indicating significant differences in expression between the multi-cancer and normal groups. According to analysis of TCGA datasets and corresponding clinical information, high expression of ETR.sig-related genes is associated with advanced tumor staging and poorer prognosis in cancer patients. The ETR.sig score exhibited high value in predicting the prognosis of cancer patients (Fig. 5), further underscoring the significance of key genes in ETR.sig in the occurrence and development of tumors, prompting us to further explore their roles in the TME.

By analyzing scRNA-seq data from tumors and their corresponding control samples, our study demonstrated that blood-derived exoRNAs may reflect the global effects of cancer, occurring not only in developing primary tumors but also in reprogramming of the microenvironment, metastatic foci, and immune system (Fig. 6, Fig. S8-10). Thus far, we have shown that cancer-related blood-derived exoRNAs primarily originate from tumor cells, macrophages, and T cells. Taking samples from KIRC and

LUAD as examples, KRT18 and MALAT1 were found to be expressed in epithelial cells and to be similarly upregulated in samples from LUAD patients with LN metastasis and distant metastasis. Additionally, MALAT1 was upregulated overall in tumor tissues (Fig. S8-10). Notably, there were widespread differences in MALAT1 and PPDPF expression between cancer and normal sample-derived single cells in both the KIRC and LUAD samples. Particularly in KIRC tissue, the genes exhibited consistent upregulation; in KIRC PBMC samples, a prevalent downregulation was observed relative to that in the control group. Violin plots of the gene expression data further indicated high MALAT1 and PPDPF expression in both cancer and normal tissues, suggesting potential alterations in the TME (Fig. S8-9). Furthermore, compared with those in control samples, S100A9 and RAB32 in macrophages exhibited specific upregulation tendencies and consistent upregulation trends in cancer and metastasis samples. Moreover, FCER1G exhibited universally high expression in myeloid immune cells, particularly in macrophages. KRT18 exhibited high expression in tumors and, notably, in metastatic tumor cells (MTCs). SLC9A3R2 was found to be highly expressed in vascular endothelial cells in both KIRC tissue and brain metastasis samples. However, changes in expression of FCER1G, KRT18, and SLC9A3R2 were not observed between tumor and control cell types.

Through comprehensive transcriptomic profiling of blood-derived exosomes, we identified and validated a novel set of multi-cancer biomarkers, ETR.sig, that exhibited promising diagnostic accuracy for both cancer detection and origin determination. In this multi-phase, multi-center study involving eight distinct cancer types, ETR.sig demonstrated potential as a minimally invasive, rapid, and reproducible diagnostic method, showing advantages over many conventional approaches. While our findings suggest that exosome-derived RNA signatures may reflect underlying tumor biology, further research is necessary to elucidate the role of exosome-mediated cell-to-cell communication.

Limitation

However, this study has several limitations. First, the inclusion of a broader range of benign diseases would improve the model's ability to differentiate between cancer and complex non-cancer conditions. Second, the incorporation of additional exosomal components, such as miRNA and methylated DNA, could enhance the richness of the model and provide deeper insights into tumor biology. Third, expanding the sample size and conducting prospective studies are crucial for refining and validating the model in larger, more diverse cohorts. Finally, further research is needed to trace and investigate the key biomarkers identified in this study. This includes leveraging additional single-cell datasets, in vitro functional assays, and fluorescence-labeled biomolecular experiments to elucidate their biological mechanisms and improve the clinical applicability of the model.

Conclusion

In conclusion, while the results of this study are promising, ETR.sig requires additional validation and refinement. Future work should focus on addressing these limitations to further optimize its performance and broaden its potential for early multi-cancer detection, ultimately contributing to the advancement of precision oncology and liquid biopsy strategies.

Abbreviations

AUC	Area under the curve
AdaBoost	Adaptive boosting classification trees
AUCell	A package for scoring gene sets in single-cell RNA-seq data
BRCA	Breast cancer
cDNA	Complementary DNA
COAD	Colon adenocarcinoma
CAN	Chromosomal copy number alteration
ChIP	Chromatin immunoprecipitation
cRNA	Circular RNA
CSF	Cerebrospinal fluid
Cox	Cox proportional hazards regression
CUT&Tag	Cleavage under targets and Tagmentation
DEGs	Differentially expressed genes
ETR.sig	Exosomal tumor RNA signatures
exoRNA-seq	Exosomal RNA sequencing
exoRbase	Exosomal RNA database
FC	Fold change
FISH	Fluorescence in situ hybridization
GLM	Generalized linear model
GEO	Gene Expression Omnibus
GSVA	Gene Set Variation Analysis
KIRC	Kidney renal clear cell carcinoma
K–M	Kaplan-Meier
LIHC	Liver hepatocellular carcinoma
LogiBoost	Boosted logistic regressions
LUAD	Lung adenocarcinoma
ML	Machine learning
Mann–Whitney U Test	A nonparametric test for independent samples
NTA	Nanoparticle tracking analysis
NB	Naïve Bayes
OS	Overall survival
PCA	Principal component analysis
PRAD	Prostate adenocarcinoma
qPCR	Quantitative polymerase chain reaction
ROC	Receiver operating characteristic
RNA-seq	RNA sequencing
Rbase ExoRNA-seq	Database for exosomal RNA data
RF	Random forest
Rpart	Classification and regression tree
scRNA-seq	Single-cell RNA sequencing
SVM	Support vector machine
t-test	Student's t-test
TEM	Transmission electron microscopy
TME	Tumor microenvironment
TaqMan	QPCR Quantitative polymerase chain reaction with TaqMan probes
UMI	Unique molecular identifier
Wilcoxon	Signed-Rank Test A nonparametric test for paired samples
WB	Western blotting
ZM	Zero mode waveguides

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12943-025-02271-4.

Supplemental Material 1: Supplementary Figure S1. Study Design. BRCA, breast cancer; COAD, colon adenocarcinoma; KIRC, kidney renal clear cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; OV, ovarian cancer; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma

Supplemental Material 2: Supplementary Figure S2. Quality control of exosome isolation and verification. (A) NTA image showing the peaks of circulating exosomes from localized RCC, RM, PCA, BPH, NSCLC, ColC, BC, HCC, PAAD, OV, and HCs and the particle size distribution of exosomes isolated from the plasma samples of localized RCC, RM, PCA, BPH, NSCLC, ColC, BC, HCC, PAAD, OV, and HCs. (B) TEM image showing the presence of exosomes isolated from the plasma samples of patients with RCC or NSCLC and healthy controls. (C) WB showing the exosomal markers CD9, CD63, TSG101, and β-actin in exosomes isolated from plasma samples of patients with RCC, RM, PCA, BPH, NSCLC, ColC, BC, HCC, PAAD, OV, or HC. TEM, transmission electron microscopy; NTA, nanoparticle tracking analysis; WB, Western blotting; RCC, renal cell carcinoma; RM, renal mass; PCA, prostate cancer; BPH, benign prostatic hyperplasia; NSCLC, non-small cell lung cancer; ColC, colorectal cancer; BC, breast cancer; HCC, hepatocellular carcinoma; PAAD, pancreatic adenocarcinoma; OV, ovarian cancer; HC, healthy control

Supplemental Material 3: Supplementary Figure S3. ROC curve for the diagnostic model constructed by the RF machine learning algorithm. (A) ROC curves for the diagnostic model constructed by RF for distinguishing multi-cancer patients from healthy controls. (B) ROC curves for the diagnostic model constructed by RF for distinguishing diverse cancer types from corresponding controls

Supplemental Material 4: Supplementary Figure S4. Violin plots illustrating associations between key genes related to ETR.sig and clinical stage

Supplemental Material 5: Supplementary Figure S5. K-M survival curves analyzing the prognostic differences between patients with high and low gene expression stratified by median gene expression

Supplemental Material 6: Supplementary Figure S6. Cellular identification in the scRNA-seq dataset of KIRC. (A) UMAP visualization displaying the major cell clusters in the scRNA-seq dataset of KIRC. (B) Bubble plots showing the expression levels of marker genes in the scRNA-seq dataset of KIRC, with a side bar indicating the proportion of major cell types of tissue origin. (C) UMAP visualization displaying the major cell clusters in the scRNA-seq data of the PBMC sample of KIRC. (D) Bubble plots showing the expression levels of marker genes in the scRNA-seq dataset of PBMC samples from KIRC patients, with a side bar indicating the proportion of major cell types of tissue origin. (E) UMAP visualization displaying the major cell clusters of the scRNA-seq dataset for brain metastasis samples of KIRC. (F) Bubble plots showing the expression levels of marker genes in the scRNA-seq dataset of brain metastasis samples from KIRC patients

Supplemental Material 7: Supplementary Figure S7. Cellular identification in the scRNA-seq dataset of LUAD. (A) UMAP visualization displaying the major cell clusters in the scRNA-seq dataset of LUAD. (B) Bubble plots showing the expression levels of marker genes in the scRNA-seq dataset of LUAD, with a side bar indicating the proportion of major cell types of tissue origin. (C) UMAP plot showing the major cell clusters in the scRNA-seq data of metastatic LUAD patient LNs. (D) Bubble plots showing the expression levels of marker genes in the scRNA-seq dataset of metastatic LUAD LN samples, with a side bar indicating the proportion of major cell types of tissue origin. (E) UMAP plot displaying the major cell clusters in the scRNA-seq data of metastatic LUAD samples. (F) Bubble plots showing the expression levels of marker genes in the scRNA-seq dataset of metastatic samples of LUAD

Supplemental Material 8: Supplementary Figure S8. Violin plots showing the expression levels of key genes in ETR.sig in different cell types (KIRC)

Supplemental Material 9: Supplementary Figure S9. Violin plots showing the expression levels of key genes in ETR.sig in different cell types (LUAD)

Supplemental Material 10: Supplementary Figure S10. Bubble plot presenting the differences in expression of ETR.sig in various cell types between the tumor and control groups, with red indicating upregulation in the tumor group and blue indicating downregulation

Supplemental Material 11: Supplementary Table S1. Table of oncogenes and tumor suppressor genes identified in the TSGene and NCG databases

Supplemental Material 12: Supplementary Table S2. Primer design for 33 target genes

Supplemental Material 13: Supplementary Table S3. Statistical table of TaqMan qPCR standard curves (Biosystems™ 7500)

Supplemental Material 14: Supplementary Table S4. Statistical table of TaqMan qPCR standard curves (Biosystems™ StepOnePlus)

Supplemental Material 15: Supplementary Table S5. Table of sample statistics for blood-derived exoRNA-seq in the discovery group

Supplemental Material 16: Supplementary Table S6. Patient demographic and clinical characteristics in the blood-derived exosome TaqMan qPCR screening and model construction cohorts

Supplemental Material 17: Supplementary Table S7. Identification of DEGs between cancer and control samples within eight distinct cancer types using corresponding exoRNA-seq data

Supplemental Material 18: Supplementary Table S8. KEGG pathway analysis of upregulated DEGs between cancer and control samples within eight distinct tumor types using corresponding exoRNA-seq data

Supplemental Material 19: Supplementary Table S9. KEGG pathway analysis of downregulated DEGs between cancer and control samples within eight distinct tumor types using corresponding exoRNA-seq data

Supplemental Material 20: Supplementary Table S10. Identification of DEGs distinguishing multi-cancer patients from healthy controls

Supplemental Material 21: Supplementary Table S11. Screening of DEGs was performed using univariate logistic regression analysis based on their association with binary variables in cancer and control samples

Supplemental Material 22: Supplementary Table S12. AUC values of training and validation sets in multi-cancer cohorts for the six candidate gene panels constructed by LASSO regression analyses

Supplemental Material 23: Supplementary Table S13. Feature genes selected by univariate logistic regression and LASSO analysis for the cancer-specific diagnostic models

Supplemental Material 24: Supplementary Table S14. Differential expression analysis of key genes in ETR.sig using TaqMan qPCR data

Supplemental Material 25: Supplementary Table S15. Performance evaluation of multi-cancer diagnostic models using nine machine learning algorithms

Supplemental Material 26: Supplementary Table S16. Performance evaluation results of multi-cancer multi-class classification models using five machine learning algorithms

Supplemental Material 27: Supplementary Table S17. List of feature genes included in single-cancer diagnostic models according to exosomal TaqMan qPCR data

Supplemental Material 28: Supplementary Table S18. Sample counts and cell number statistics for scRNA-seq data

Supplemental Material 29: Supplementary Table S19. DEGs between cancer and healthy-derived cells in KIRC and LUAD

Acknowledgements

We thank all the blood donors for their willingness to participate in this study. We gratefully acknowledge support from the National Natural Science Foundation of China (NSFC) (82372828 to F. Wang, 81960477 to Y. Liu, 82160483 to J. Cheng, 82072846 to B. Xu, and 82203134 to X. Shi), the Science Foundation for Distinguished Young Scholars of Guangxi (2023GXNSFFA026003 to F. Wang), Shanghai "Rising Stars of Medical Talent" Youth Development Program "Outstanding Youth Medical Talents" (SHWSRS(2021)_099 to B. Xu),

Guangxi Natural Science Foundation (2023GXNSFDA026041 to Y. Liu), Oriental Talents Program Youth Project (formerly Shanghai Youth Top Talents; B.Xu), the Science Foundation for Distinguished Young Scholars of Guangxi Medical University (F. Wang), the Science and Technology Major Project of Guangxi (AA22096030 to F. Wang and Z. Mo and AA22096032 to F. Wang and Z. Mo), the Yongjiang Program of Nanning (2021015 to F. Wang), Anhui Province Translational Medicine Research Fund Project (2021zhyx-C59 to D. Zhang), and Suzhou Science and Technology Project of Anhui (2021-13 to D. Zhang). The funding bodies had no role in the design of the study or the collection, analysis, or interpretation of data nor in writing the manuscript.

Authors' contributions

F. Wang: led the project, designed the experimental approach, coordinated the project, and revised the manuscript. C. B. Wang: designed the experimental approach, interpreted and analyzed the data, and wrote the manuscript. S. Chen: interpreted the data, and wrote the manuscript. C. M. Wei: designed the experimental approach, coordinated the multi-cancer sample preparation and data collection. J. Ji and Y. Liu: coordinated the multi-cancer sample preparation and data collection and performed validation experiments. L.Liang: coordinated the lung cancer sample preparation and data collection. X. Li: coordinated the ovarian cancer sample preparation and data collection. T. Li, Z. Wang, F. Liu, Z. Chen and W. Zhou. coordinated the liver cancer sample preparation and data collection. T. Li and Y. Liu: coordinated the breast cancer sample preparation and data collection. M. Li and Y. Hong: coordinated the colorectal cancer sample preparation and data collection. X. Hu, and L. Ouyang: coordinated the pancreatic cancer sample preparation and data collection. Y. Chen, L. Zhao, Y. Fang, Z. Liu, L. Mo, T. Li, Q. Zhang and B. Yang: coordinated the prostate cancer and kidney cancer sample preparation and data collection. W. Lu, D. Zhang, X. Wei, and J. Cheng: coordinated the healthy control sample preparation and data collection. X. Shi, W. Lu, X. He and J. Wen: performed sample processing. Z. Wang, X. Li, and R. Zhou: performed exosome purification and identification. W. Lu, X. Wei, L. Wang, B. Yang, S. Huang, H. Zhang and G. Pang: helped to coordinate the project, coordinated the multi-cancer data collection. L. Wang and H. Zhang: analyzed the data and critically reviewed the manuscript. L. Ouyang, Z. Wang and J. Chen: contributed to discussing the hypothesis and data analysis. B. Xu, Z. Mo and F. Wang: conceived and discussed the hypothesis. All authors read and approved the final manuscript.

Funding

We gratefully acknowledge support from the National Natural Science Foundation of China (NSFC) (82372828 to F. Wang, 81960477 to Y. Liu, 82160483 to J. Cheng, 82072846 to B. Xu, and 82203134 to X. Shi), the Science Foundation for Distinguished Young Scholars of Guangxi (2023GXNSFFA026003 to F. Wang), Shanghai "Rising Stars of Medical Talent" Youth Development Program "Outstanding Youth Medical Talents" (SHWSRS(2021)_099 to B. Xu), Guangxi Natural Science Foundation (2023GXNSFDA026041 to Y. Liu), Oriental Talents Program Youth Project (formerly Shanghai Youth Top Talents; B.Xu), the Science Foundation for Distinguished Young Scholars of Guangxi Medical University (F. Wang), the Science and Technology Major Project of Guangxi (AA22096030 to F. Wang and Z. Mo and AA22096032 to F. Wang and Z. Mo), the Yongjiang Program of Nanning (2021015 to F. Wang), Anhui Province Translational Medicine Research Fund Project (2021zhyx-C59 to D. Zhang), and Suzhou Science and Technology Project of Anhui (2021–13 to D. Zhang). The funding bodies had no role in the design of the study or the collection, analysis, or interpretation of data nor in writing the manuscript.

Data availability

Blood-derived exosomal RNA-seq data were obtained from online databases, as detailed in Supplementary Table S1. All other data supporting the findings of this study are available from the corresponding author on reasonable request. All code was implemented in R using caret as the main machine learning package. All code used to reproduce the experiments presented here are publicly available at: https://github.com/FUBO-lab/Exosome-pan-cancer, as of the date of publication.

Declarations

Ethics approval and consent to participate

This study obtained approval from the Clinical Research Ethics Committees of Guangxi Medical University (Approval Numbers: GXMU2022-0154). Written informed consent was obtained from all participants.

Competing interests

F.B. Wang, Z.N. Mo, C.B. Wang, S.H. Chen and C.M. Wei have filed a Chinese patent application related to this work (Application number: 202410931244.5).

Author details

[1]Center for Genomic and Personalized Medicine, Guangxi Key Laboratory for Genomic and Personalized Medicine, Guangxi Collaborative Innovation Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning 530021, Guangxi, China. [2]Department of Urology, The First Affiliated Hospital of Guangxi Medical University, Guangxi Medical University, Guangxi 530021, China. [3]School of Life Sciences, Guangxi Medical University, Nanning, Guangxi 530021, China. [4]Department of Urology, Guangxi Medical University Cancer Hospital, Nanning, Guangxi, China. [5]School of Public Health, Guangxi Medical University, Nanning, Guangxi 530021, China. [6]Department of Urology, Shanghai Changhai Hospital, Naval Medical University, Shanghai 200433, China. [7]Department of Urology, Naval Medical Center, Naval Medical University, Shanghai 200433, China. [8]Department of Breast, Bone and Soft Tissue Oncology, Guangxi Medical University Cancer Hospital, Nanning, Guangxi 530021, China. [9]Laboratory of Breast Cancer Diagnosis and Treatment Research of Guangxi, Department of Education, Affiliated Tumor Hospital of Guangxi Medical University, Nanningaq, Guangxi 530021, China. [10]Department of Oncology, The First People'S Hospital of Yulin, the, Sixth Affiliated Hospital of Guangxi Medical Universityaq, Guangxi 537000, China. [11]Department of Urology, The First People'S Hospital of Yulin, the, Sixth Affiliated Hospital of Guangxi Medical Universityaq, Guangxi 537000, China. [12]Department of Obstetrics and Gynecology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200080, China. [13]Department of Urology, Suzhou Hospital of Anhui Medical University, Suzhouaq, AnHui 234000, China. [14]Department of Hepatobiliary Surgery, The First People'S Hospital of Yulin, the, Sixth Affiliated Hospital of Guangxi Medical Universityaq, Guangxi 537000, China. [15]Department of Urology, Jinling Hospital, Medical School of Nanjing University, Nanjing 210002, Jiangsu, China. [16]Collaborative Innovation Centre of Regenerative Medicine and Medical Bioresource Development and Application Co-Constructed by the Province and Ministry, Guangxi Medical University, Nanning, Guangxi 530021, China. [17]Department of Hepatobiliary and Pancreatic Surgery, Changhai Hospital, Naval Medical University, Shanghai 200433, China. [18]Department of Colorectal and Anal Surgery, The First People'S Hospital of Yulin, the, Sixth Affiliated Hospital of Guangxi Medical Universityaq, Guangxi 537000, China. [19]The Third Department of Hepatic Surgery, Eastern Hepatobiliary Surgery Hospital, Naval Medical University, Shanghai 200438, China. [20]Outpatient Department, Qingdao, Special Servicemen Recuperation Center of PLA Navy, Shandong 266071, China. [21]The First Outpatient Department, General Hospital of PLA Northern Theater Command, Shenyangaq, Liaoning 110001, China. [22]Department of Laboratory Medicine, Third Affiliated Hospital of Naval Medical University, Shanghai 200438, China. [23]Department of Colorectal Surgery, Changhai Hospital, Naval Medical University, Shanghai 200433, China. [24]Suzhou Key Laboratory for Clinical Big Data and Intelligent Treatment of Urinary System Diseases, Suzhouaq, AnHui 234000, China. [25]Research Center for Intelligence Information Technology, Nantong University, Nantong, Jiangsu 226001, China. [26]Department of Urology, The First Afliated Hospital of Soochow University, Suzhou 215006, China. [27]Key Laboratory of Medical Epigenetics and Metabolism, Institutes of Biomedical Sciences, Fudan University Shanghai Cancer Center, Fudan University, Shanghai 201321, China. [28]Department of Hepatobiliary and Pancreatic Surgery, School of Medicine, Shanghai Fourth People's Hospital, Tongji University, Shanghai 200434, China. [29]Department of Urology, Shanghai Ninth People'S Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200011, China.

Received: 20 January 2025 Accepted: 13 February 2025
Published online: 01 March 2025

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209.
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024;74:12–49.

3. Crichton DJ, Altinok A, Amos CI, Anton K, Cinquini L, Colbert M, Feng Z, Goel A, Kelly S, Kincaid H, et al. Cancer biomarkers and big data: a planetary science approach. Cancer Cell. 2020;38:757–60.
4. Bartlett AH, Liang JW, Sandoval-Sierra JV, Fowke JH, Simonsick EM, Johnson KC, Mozhui K. Longitudinal study of leukocyte DNA methylation and biomarkers for cancer risk in older adults. Biomark Res. 2019;7:10.
5. Zhang J, Shi J, Zhang H, Zhu Y, Liu W, Zhang K, Zhang Z. Localized fluorescent imaging of multiple proteins on individual extracellular vesicles using rolling circle amplification for cancer diagnosis. J Extracell Vesicles. 2020;10: e12025.
6. Becker A, Thakur BK, Weiss JM, Kim HS, Peinado H, Lyden D. Extracellular vesicles in cancer: cell-to-cell mediators of metastasis. Cancer Cell. 2016;30:836–48.
7. Exosome Profiling Pinpoints Cancer Type. Cancer Discov. 2020;10:1619.
8. Yu W, Hurley J, Roberts D, Chakrabortty SK, Enderle D, Noerholm M, Breakefield XO, Skog JK. Exosome-based liquid biopsies in cancer: opportunities and challenges. Ann Oncol. 2021;32:466–77.
9. Valadi H, Ekstrom K, Bossios A, Sjostrand M, Lee JJ, Lotvall JO. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol. 2007;9:654–9.
10. Zhou W, Fong MY, Min Y, Somlo G, Liu L, Palomares MR, Yu Y, Chow A, O'Connor ST, Chin AR, et al. Cancer-secreted miR-105 destroys vascular endothelial barriers to promote metastasis. Cancer Cell. 2014;25:501–15.
11. Huang X, Yuan T, Tschannen M, Sun Z, Jacob H, Du M, Liang M, Dittmar RL, Liu Y, Liang M, et al. Characterization of human plasma-derived exosomal RNAs by deep sequencing. BMC Genomics. 2013;14:319.
12. Li Y, Zhao J, Yu S, Wang Z, He X, Su Y, Guo T, Sheng H, Chen J, Zheng Q, et al. Extracellular vesicles long RNA sequencing reveals abundant mRNA, circRNA, and lncRNA in human blood as potential biomarkers for cancer diagnosis. Clin Chem. 2019;65:798–808.
13. Saugstad JA, Lusardi TA, Van Keuren-Jensen KR, Phillips JI, Lind B, Harrington CA, McFarland TJ, Courtright AL, Reiman RA, Yeri AS, et al. Analysis of extracellular RNA in cerebrospinal fluid. J Extracell Vesicles. 2017;6:1317577.
14. Yang Y, Zhang J, Zhang W, Wang Y, Zhai Y, Li Y, Li W, Chang J, Zhao X, Huang M, et al. A liquid biopsy signature of circulating extracellular vesicles-derived RNAs predicts response to first line chemotherapy in patients with metastatic colorectal cancer. Mol Cancer. 2023;22:199.
15. Yu S, Li Y, Liao Z, Wang Z, Wang Z, Li Y, Qian L, Zhao J, Zong H, Kang B, et al. Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. Gut. 2020;69:540–50.
16. McKiernan J, Donovan MJ, O'Neill V, Bentink S, Noerholm M, Belzer S, Skog J, Kattan MW, Partin A, Andriole G, et al. A novel urine exosome gene expression assay to predict high-grade prostate cancer at initial biopsy. JAMA Oncol. 2016;2:882–9.
17. Tao W, Wang BY, Luo L, Li Q, Meng ZA, Xia TL, Deng WM, Yang M, Zhou J, Zhang X, et al. A urine extracellular vesicle lncRNA classifier for high-grade prostate cancer and increased risk of progression: A multi-center study. Cell Rep Med. 2023;4: 101240.
18. Ji J, Chen R, Zhao L, Xu Y, Cao Z, Xu H, Chen X, Shi X, Zhu Y, Lyu J, et al. Circulating exosomal mRNA profiling identifies novel signatures for the detection of prostate cancer. Mol Cancer. 2021;20:58.
19. Wei C, Chen X, Ji J, Xu Y, He X, Zhang H, Mo Z, Wang F. Urinary exosomal prostate-specific antigen is a noninvasive biomarker to detect prostate cancer: Not only old wine in new bottles. Int J Cancer. 2023;152:1719–27.
20. Wang CB, Chen SH, Zhao L, Jin X, Chen X, Ji J, Mo ZN, Wang FB. Urine-derived exosomal PSMA is a promising diagnostic biomarker for the detection of prostate cancer on initial biopsy. Clin Transl Oncol. 2023;25:758–67.
21. He X, Tian F, Guo F, Zhang F, Zhang H, Ji J, Zhao L, He J, Xiao Y, Li L, et al. Circulating exosomal mRNA signatures for the early diagnosis of clear cell renal cell carcinoma. BMC Med. 2022;20:270.
22. Chen F, Wendl MC, Wyczalkowski MA, Bailey MH, Li Y, Ding L. Moving pan-cancer studies from basic research toward the clinic. Nat Cancer. 2021;2:879–90.
23. Zhang Z, Wang ZX, Chen YX, Wu HX, Yin L, Zhao Q, Luo HY, Zeng ZL, Qiu MZ, Xu RH. Integrated analysis of single-cell and bulk RNA sequencing data reveals a pan-cancer stemness signature predicting immunotherapy response. Genome Med. 2022;14:45.
24. Liao J, Yu Z, Chen Y, Bao M, Zou C, Zhang H, Liu D, Li T, Zhang Q, Li J, et al. Single-cell RNA sequencing of human kidney. Sci Data. 2020;7:4.
25. Su C, Lv Y, Lu W, Yu Z, Ye Y, Guo B, Liu D, Yan H, Li T, Zhang Q, et al. Single-cell RNA sequencing in multiple pathologic types of renal cell carcinoma revealed novel potential tumor-specific markers. Front Oncol. 2021;11:719564.
26. Zhang M, Zhai W, Miao J, Cheng X, Luo W, Song W, Wang J, Gao WQ. Single cell analysis reveals intra-tumour heterogeneity, microenvironment and potential diagnosis markers for clear cell renal cell carcinoma. Clin Transl Med. 2022;12: e713.
27. Laughney AM, Hu J, Campbell NR, Bakhoum SF, Setty M, Lavallee VP, Xie Y, Masilionis I, Carr AJ, Kottapalli S, et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. Nat Med. 2020;26:259–69.
28. Borcherding N, Vishwakarma A, Voigt AP, Bellizzi A, Kaplan J, Nepple K, Salem AK, Jenkins RW, Zakharia Y, Zhang W. Mapping the immune environment in clear cell renal carcinoma by single-cell genomics. Commun Biol. 2021;4:122.
29. Wu TD, Madireddi S, de Almeida PE, Banchereau R, Chen YJ, Chitre AS, Chiang EY, Iftikhar H, O'Gorman WE, Au-Yeung A, et al. Peripheral T cell expansion predicts tumour infiltration and clinical response. Nature. 2020;579:274–8.
30. Cillo AR, Kurten CHL, Tabib T, Qi Z, Onkar S, Wang T, Liu A, Duvvuri U, Kim S, Soose RJ, et al. Immune landscape of viral- and carcinogen-driven head and neck cancer. Immunity. 2020;52(183–199).
31. Steele NG, Carpenter ES, Kemp SB, Sirihorachai VR, The S, Delrosario L, Lazarus J, Amir ED, Gunchick V, Espinoza C, et al. Multimodal mapping of the tumor and peripheral blood immune landscape in human pancreatic cancer. Nat Cancer. 2020;1:1097–112.
32. Zhang C, Yin K, Liu SY, Yan LX, Su J, Wu YL, Zhang XC, Zhong WZ, Yang XN. Multiomics analysis reveals a distinct response mechanism in multiple primary lung adenocarcinoma after neoadjuvant immunotherapy. J Immunother Cancer. 2021;9:e002312.
33. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, Lee JI, Suh YL, Ku BM, Eum HH, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11:2285.
34. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.
35. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, Kumar T, Hu M, Sei E, Davis A, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39:599–608.
36. Wang C, He Y, Zheng J, Wang X, Chen S. Dissecting order amidst chaos of programmed cell deaths: construction of a diagnostic model for KIRC using transcriptomic information in blood-derived exosomes and single-cell multi-omics data in tumor microenvironment. Front Immunol. 2023;14:1130513.
37. Lai H, Li Y, Zhang H, Hu J, Liao J, Su Y, Li Q, Chen B, Li C, Wang Z, et al. exoRBase 2.0: an atlas of mRNA, lncRNA and circRNA in extracellular vesicles from human biofluids. Nucleic Acids Res. 2022;50:D118–28.
38. Wu B, Liu DA, Guan L, Myint PK, Chin L, Dang H, Xu Y, Ren J, Li T, Yu Z, et al. Stiff matrix induces exosome secretion to promote tumour growth. Nat Cell Biol. 2023;25:415–24.
39. Kalluri R, McAndrews KM. The role of extracellular vesicles in cancer. Cell. 2023;186:1610–26.
40. Zhang X, Chen F, Huang P, Wang X, Zhou K, Zhou C, Yu L, Peng Y, Fan J, Zhou J, et al. Exosome-depleted MiR-148a-3p derived from hepatic stellate cells promotes tumor progression via ITGA5/PI3K/Akt axis in hepatocellular carcinoma. Int J Biol Sci. 2022;18:2249–60.
41. Lima LG, Ham S, Shin H, Chai EPZ, Lek ESH, Lobb RJ, Muller AF, Mathivanan S, Yeo B, Choi Y, et al. Tumor microenvironmental cytokines bound to cancer exosomes determine uptake by cytokine receptor-expressing cells and biodistribution. Nat Commun. 2021;12:3543.
42. Paskeh MDA, Entezari M, Mirzaei S, Zabolian A, Saleki H, Naghdi MJ, Sabet S, Khoshbakht MA, Hashemi M, Hushmandi K, et al. Emerging role of exosomes in cancer progression and tumor microenvironment remodeling. J Hematol Oncol. 2022;15:83.
43. Yang E, Wang X, Gong Z, Yu M, Wu H, Zhang D. Exosome-mediated metabolic reprogramming: the emerging role in tumor microenvironment remodeling and its influence on cancer progression. Signal Transduct Target Ther. 2020;5:242.

44. Hoshino A, Kim HS, Bojmar L, Gyan KE, Cioffi M, Hernandez J, Zambirinis CP, Rodrigues G, Molina H, Heissel S, et al. Extracellular vesicle and particle biomarkers define multiple human cancers. Cell. 2020;182(1044–1061).
45. Nakamura K, Zhu Z, Roy S, Jun E, Han H, Munoz RM, Nishiwada S, Sharma G, Cridebring D, Zenhausern F, et al. An exosome-based transcriptomic signature for noninvasive, early detection of patients with pancreatic ductal adenocarcinoma: a multicenter cohort study. Gastroenterology. 2022;163(1252–1266).
46. Lapitz A, Azkargorta M, Milkiewicz P, Olaizola P, Zhuravleva E, Grimsrud MM, Schramm C, Arbelaiz A, O'Rourke CJ, La Casta A, et al. Liquid biopsy-based protein biomarkers for risk prediction, early diagnosis, and prognostication of cholangiocarcinoma. J Hepatol. 2023;79:93–108.
47. Sun J, Lu Z, Fu W, Lu K, Gu X, Xu F, Dai J, Yang Y, Jiang J. Exosome-derived ADAM17 promotes liver metastasis in colorectal cancer. Front Pharmacol. 2021;12:734351.
48. Zhang J, Ji C, Zhang H, Shi H, Mao F, Qian H, Xu W, Wang D, Pan J, Fang X, et al. Engineered neutrophil-derived exosome-like vesicles for targeted cancer therapy. Sci Adv. 2022;8:eabj8207.
49. Qu L, Ding J, Chen C, Wu ZJ, Liu B, Gao Y, Chen W, Liu F, Sun W, Li XF, et al. Exosome-transmitted lncARSR promotes sunitinib resistance in renal cancer by acting as a competing endogenous RNA. Cancer Cell. 2016;29:653–68.
50. Rivoltini L, Chiodoni C, Squarcina P, Tortoreto M, Villa A, Vergani B, Burdek M, Botti L, Arioli I, Cova A, et al. TNF-Related Apoptosis-Inducing Ligand (TRAIL)-armed exosomes deliver proapoptotic signals to tumor site. Clin Cancer Res. 2016;22:3499–512.

Publisher's Note