Supplementary Information (SI) for Chemical Science.
This journal is © The Royal Society of Chemistry 2024

Supplementary Materials for Manuscript

A simplified and efficient extracellular vesicles-based proteomics strategy for early diagnosis of colorectal cancer

Jin Zhang,‡[a] Zhaoya Gao,‡[bc] Weidi Xiao,‡[ad] Ningxin Jin,[a] Jiaming Zeng,[d] Fengzhang Wang,[a] Xiaowei Jin,[e] Liguang Dong,[f] Jian Lin,*[gh] Jin Gu*[bcij] and Chu Wang*[adhj]

[a]Beijing National Laboratory for Molecular Sciences, Key Laboratory of Bioorganic Chemistry and Molecular Engineering of Ministry of Education, College of Chemistry and Molecular Engineering, Peking University, Beijing, China.

[b]Department of Gastrointestinal Surgery, Peking University Shougang Hospital, Beijing, China.

[c]Center for Precision Diagnosis and Treatment of Colorectal Cancer and Inflammatory Diseases， Peking University Health Science Center.

[d]Peking University Chengdu Academy for Advanced Interdisciplinary Biotechnologies, Chengdu, China.

[e]Department of Gastroenterology, Peking University Shougang Hospital, Beijing, China.

[f]Center for Health Care Management, Peking University Shougang Hospital, Beijing, China.

[g]Department of Pharmacy, NMPA Key Laboratory for Research and Evaluation of Generic Drugs, Peking University Third Hospital Cancer Center, Peking University Third Hospital, Beijing, China.

[h]Synthetic and Functional Biomolecules Center, Peking University, Beijing, China.

[i]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Gastrointestinal Surgery, Peking University Cancer Hospital & Institute, Beijing, China.

[j]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China.

Corresponding Author* E-mail: chuwang@pku.edu.cn; zlgujin@126.com; linjian@pku.edu.cn;

Table of contents

Experimental details

Collection of plasma samples. Plasma samples were obtained from 21 healthy individuals, 16 adenomatous polyp patients and 21 early-stage CRC patients at Peking University Shougang Hospital. In accordance with the diagnostic criteria of AJCC (8^{th})[1], we selected early CRC patients defined as stage I and II. All patients were pathologically diagnosed by experienced pathologists based on histologic criteria. For each patient, peripheral venous blood was drawn into Vacutainer K2-EDTA tubes (Becton Dickinson) and was kept at room temperature for 30 minutes followed by centrifugation at 1350 g for 15 minutes and centrifugation at 3000 g for 15 minutes at 4°C. The final supernatant was stored at -80 °C until further analysis. This study was performed with the approval from the medical ethics committee of Peking University Shougang Hospital. All patients were informed and signed consent forms.

Preparation of the DSPE-functionalized beads. DSPE-functionalized beads were prepared by conjugating NHS ester coated beads (Cytiva) with DSPE-PEG2000-NH2 in 0.1% triton/PBS (pH=8.0) solution at 29 ℃ for 6 h. Firstly, 50 μL NHS ester coated beads were washed twice with 1 mM HCl/PBS solution before use, respectively. Subsequently, they are resuspended in 1 mL of 0.1% triton/PBS (pH=8) buffer, and 20 μL of 50 mM DSPE-PEG2000-NH2 stock was added to react at room temperature for 6 h. After the reaction was completed, the supernatant was discarded via centrifugation, and a quenching buffer (100 mM Tris-HCl pH=8.0, 500 mM NaCl, 500 mM ethanolamine) was introduced at 4℃ overnight for blocking the remaining NHS ester. The product was washed with PBS buffer three times, and dispersed in 30% glycerol/PBS and stored as 400 μL per aliquot at -20°C until use.

EVs isolation. Ultracentrifuged EV samples were purified following by standard procedures. Firstly, 200 μL plasma were thawed on ice and then diluted by 5 folds with PBS buffer. Samples were centrifuged at 12000 g for 30 min to remove large cell debris, and the supernatant were centrifuged at 100000 g for 70 min in a Beckman OptimaTM MAX-XP ultracentrifuge. The pellet was washed in PBS and centrifuged at 100000 g for 70 min again to obtain the final EV pellet. The process of isolating and purifying EVs via size-exclusion chromatography (SEC) is as follows: to remove cellular debris, 200 μL plasma was centrifuged for 10 min at 4 °C and 3000g, the supernatant was subsequently centrifuged for 30 min at 12000g and 4 °C. Next, the supernatant was applied to size-exclusion columns (qEVoriginal, 35 nm+; Izon Science Limited) equilibrated with 10 ml of PBS to isolate EVs. We eluted 20 fractions with a volume of 500 μl. Fractions 5–8 contain the highest EV concentrations and were combined and concentrated for subsequent analysis. As a comparison, plasma samples were also isolated by polymer-based precipitation using an Exosome Isolation Q3 kit (EIQ3-02001, Wayen Biotechnologies, Shanghai, China) according to the product manual. For exosome isolation by the DSPE-functionalized beads, 200 μL plasma samples were thawed on ice and centrifuged at 3000 g for 10 min to remove large particles and other debris, and the supernatant were introduced into 10 μL of the prepared DSPE-functionalized beads to incubate for 10 min at 29℃. The EV samples captured on beads were washed with PBS buffer for four times to remove impurities and stored at -80 °C until further analysis. EV-depleted plasma was prepared by collecting the supernatant carefully without disturbing the EV pellet (after Kit preparation of total plasma EVs). EVs were isolated using DSPE-PEG2000-biotin as follows: DSPE-PEG2000-biotin was diluted to 40 mM in ethanol. 10 μL of diluted DSPE-PEG2000-biotin was added to 200

μL plasma (final concentration 2 mM) and incubated at 4°C for 30 minutes. Streptavidin-coated beads were then added and incubated at room temperature for 30 minutes. Finally, beads were washed thrice with PBS to isolate EVs.

EVs characterization. For western blotting analysis, EV samples isolated by different methods were lysed using RIPA buffer (MACGENE, MP015), and the protein concentrations were measured by BCA (Pierce, Thermo Fisher Scientific). 20 μg of proteins were added with the loading buffer and were boiled at 95 °C for 5 min. SDS-PAGE analysis and protein transfer was performed sequentially, and then the transferred membrane was blocked in 5% BSA in PBST at room temperature for 1 h. The membrane was incubated with primary antibodies overnight at 4℃ in PBST, washed 3 times with PBST, incubated with secondary antibodies for 1 h at room temperature in PBST and washed again 3 times with PBST. The results were detected with ECL (Solarbio, PE0010). The following antibodies were used: CD63 (Abcam, ab134045, 1:1000, rabbit), HSP70(Abcam, ab181606, 1:1000, rabbit), Alix (Abcam, ab186429, 1:1000, rabbit), TSG101 (Santa, sc-7964, 1:200, mouse), Syntenin (Abcam, ab133267,1:1000, rabbit), Calnexin (Sigma-Aldrich, C4731, 1:2000, rabbit). For scanning electron microscopy (SEM) characterization of EVs, the beads with EVs were diluted to an optimal concentration. A silicon wafer was placed with its smooth side facing up on carbon conductive adhesive. Using a pipette, a small amount of the sample was carefully drawn and one or two drops were dispensed onto the silicon wafer. The sample was allowed to dry before proceeding with imaging. For TEM analysis, 10 μl of the sample was applied onto a formvar/carbon coated grid and allowed to settle for 10 minutes. Excess liquid was gently blotted off, and the sample was negatively stained by applying 10 μl of 3% aqueous uranyl acetate, followed by a 1-3 minutes incubation period before excess stain was blotted away. The grid was air-dried for 10 minutes to ensure optimal sample conditions for electron microscopy analysis using a JEM-1200EX (JEOL Corporation) operating at 100 kV.

Sample preparation and LC-MS/MS analysis. Enriched EVs were lysed using the RIPA lysis buffer. In a typical experiment, 5 μg proteomes were added with 50 μL of 8M urea/PBS, and were reduced with 10 mM dithiothreitol (DTT) at 37°C for 30 minutes and alkylated with 20 mM iodoacetamide (IAA) at 35°C for 30 minutes in dark. Subsequently, 2 μL of the pre-prepared 50 μg/μL SP3 magnetic bead stocks were added with shaking and 50 μL of ethanol were added to the mixture. The sample was incubated in a ThermoMixer at 24 °C for 5 min at 1000 rpm, allowing to induce binding of the proteins to the SP3 magnetic beads. When the binding is complete, the EP tube was placed on the magnetic rack to effectively remove the unbound contaminants from the proteins, followed by washing of the SP3 magnetic beads three times with 80% ethanol. Subsequently, the SP3 magnetic beads were resuspended in 100 μL of 30 mM ammonium bicarbonate in H2O, and were subjected to brief ultrasonication for 30 s. The mixture was added with 1 μg trypsin (Meizhiyuan, Beijing) for on-beads digestion overnight at 37°C. After digestion was complete, the peptides were collected by centrifugation and magnetic separation, and were dried by vacuum centrifugation and re-dissolved in 0.1% FA/H2O for subsequent LC-MS/MS analysis.

The LC-MS/MS analysis was performed on a Q-ExactiveTM Plus mass spectrometer equipped with an Ultimate 3000 liquid chromatography (Thermo-Fisher Scientific). The loading and separation of peptide were achieved using a reversed-phase C18 column (trap column, 5 cm × 100 μm, 3 μm particle size, analytical column, 15 cm × 100 μm, 1.9 μm particle size, respectively). The mobile phase consisted of solvent A (H2O containing 0.1%

formic acid) and solvent B (80 % acetonitrile in H2O containing 0.1% formic acid). A gradient elution program was employed as follows: 0-10 min, 4% B; 10-13 min, 4-9% B; 13-47 min, 35% B; 47-50 min, 35-45% B; 50-52 min, 45-95% B; 52-57 min, 95% B; 57-58 min, 95-5% B and 58-65 min, 5% B. MS data for EVs proteomics were acquired using the data-independent acquisition (DIA) mode. Full MS scan was acquired in the range of m/z 350-2000 at a resolution of 70000, and AGC target was 3e6 with Maximum IT of 20 ms. Full MS events were followed by 28 MS/MS windows per cycle at a resolution of 17500 and AGC target value was set to 1e6 with an auto Maximum IT.

DIA quantitative data processing. The raw data were processed using DIA-NN 1.8.1[2] in an advanced library-free module. The main search settings for in silico library generation were set as following: trypsin/P with maximum 3 missed cleavage; protein N-terminal M excision on; carbamidomethyl on C as fixed modification; oxidation on M as variable modification; peptide length from 7-30; precursor charge 1-4; precursor m/z from 300 to 1800; fragment m/z from 200 to 1800. The Human UniProt isoform sequence database (3AUP000005640) was used to annotate proteins. Other search parameters were set as following: quantification strategy was set to ‘‘robust LC (high precision)’’ mode; cross-run normalization was RT-dependent; MS2 and MS1 mass accuracies were set to 0, allowing the DIA-NN to automatically determine mass tolerances; Scan window was set to 6 corresponding to the approximate average number of data points per peak; Isotopologues and MBR were turned on; neural network classifier was single-pass mode.

For identification and quantification based on the Spectronaut (version 13.12.200217.43655) software, we utilized the directDIA analysis workflow with the following specific parameter settings: Trypsin/P as specific enzyme; peptide length from 7 to 52; max missed cleavages 3; Carbamidomethyl on C as fixed modification; Oxidation on M and Acetyl at protein N-terminus as variable modifications; toggle N-terminal M turned on; precursor q value cutoff
0.01; protein q value cutoff 0.01; cross-run normalization on; data filtering set to Q value. The Human UniProt isoform sequence database (3AUP000005640) was used to annotate proteins.

Comparison of EV isolation methods. For the comparison of different EV isolation methods presented in Figure 3, we used plasma samples collected from healthy donors at Peking University Shougang Hospital. The experimental procedure, including EV isolation, proteomics sample preparation, DIA acquisition, and DIA-NN-based searching and quantification, were performed according to the protocols provided in the experimental details above.

Statistical analysis. The protein biomarkers were screened by a supervised OPLS-DA analysis, which was performed using SIMCA 14.0 (Umetrics, Sweden) software. Mann-Whitney U test were used to define statistical significance using IBM SPSS Statistics 26 (SPSS Inc., USA), $p < 0.01$ (**), $p < 0.001$ (***), $p < 0.0001$ (****), the boxplots were generated using GraphPad Prism 8 software (GraphPad Software Inc., USA). Heatmaps and ROC curves were plotted using R language. The identified proteins of interest were put into the open-source platform, DAVID Bioinformatics Resources (https://david.ncifcrf.gov/), for GO analysis. The R studio was employed for developing machine learning models to differentiate the healthy control, polyp, and early CRC cases. The support vector

machines (SVM) function in R studio was used to establish the classification model.

Supporting information tables

Table S1: Raw data on protein identification from different EV isolation methods. (see the attached excel file)

Table S2: The unique proteins identified from EVs isolated using different methods. (see the attached excel file)

Table S3: Raw data of protein identification from EV samples extracted using DSPE-Beads and DSPE-Biotin. (see the attached excel file)

Table S4: Clinical information of samples used in the discovery cohort. (see the attached excel file)

Table S5: Label free quantification of 826 EV proteins across all 30 samples in the discovery dataset using DIA-MS. (see the attached excel file)

Table S6: The quantification of 219 differential proteins across all 30 samples using DIA-MS. (see the attached excel file)

Table S7: Top 10 proteins with marked upregulation or downregulation during malignant progression from HC to CRC. The associated AUC values are reported alongside. (see the attached excel file)

Table S8: The models' performance comparison for predicting disease. (see the attached excel file)

Table S9: Clinical information of samples used in the single-blind validation cohort. (see the attached excel file)

Table S10: The combined dataset of discovery cohort and single-blind validation cohort was used for biomarker discovery. (see the attached excel file)

Supporting information figures

Fig. S1 Loading controls for Western blot analysis of EV markers. These controls correspond to the EV biomarker (CD63, TSG101, HSP70, and Alix) and exclusion marker (Calnexin) blots shown in Fig. 2C

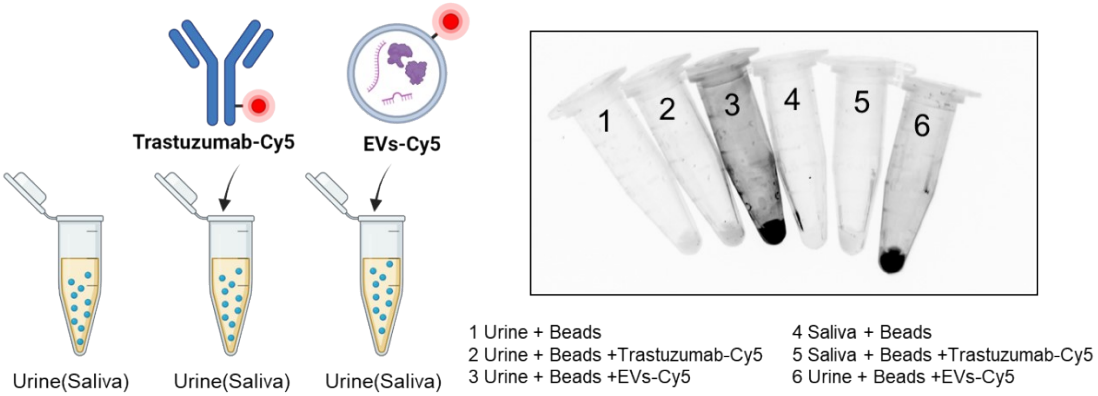

Fig. S2 Enrichment of EVs from different types of body fluids by DSPE-functionalized beads. Schematic diagram for experimentally enriching EVs from urine and saliva using DSPE-functionalized beads (left). Only tubes with the Cy5 fluorophore-conjugated EVs (tube 3 and tube 6) displayed bright fluorescence while no background fluorescence was observed in the control tubes (right).

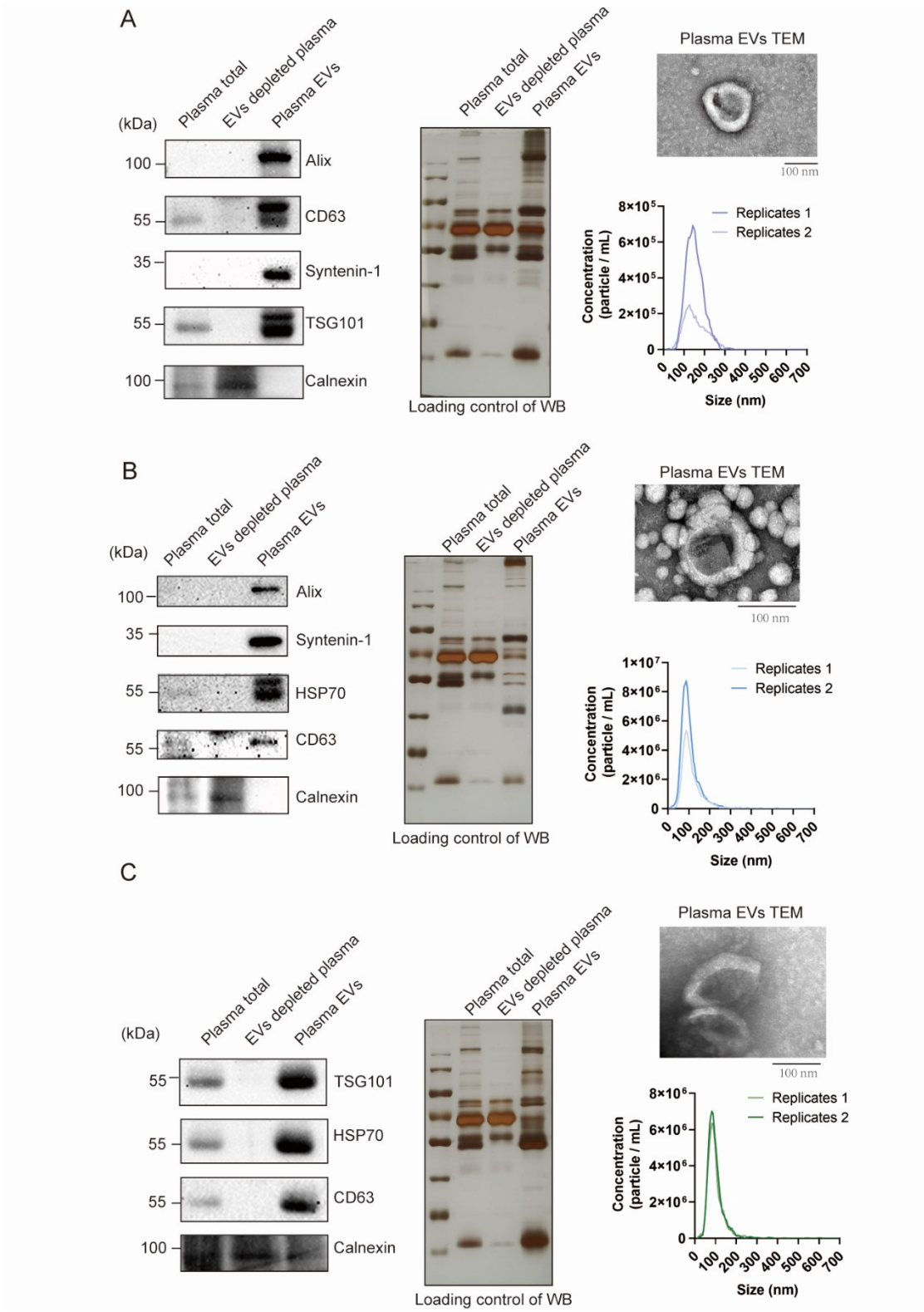

Fig. S3 The characterization of EVs isolated by using three different methods: ultracentrifugation ("UC") (A), size-exclusion chromatography ("SEC") (B), and commercial kit ("Kit") (C). Each figure involves western blot analysis of EV molecular markers, transmission electron microscopy (TEM) image of morphology of EVs and Nanoparticle tracking analysis (NTA).

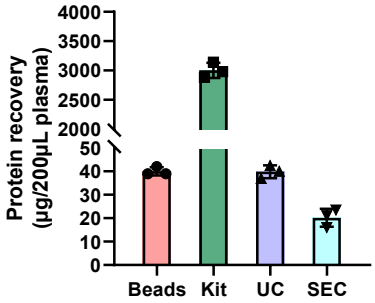

Fig. S4 Comparison of the amounts of proteins recovered from EVs isolated using different methods. DSPE-functionalized beads ("Beads"), polymer precipitation ("Kit") and ultracentrifugation methods ("UC") and size-exclusion chromatography ("SEC") were used to enrich EVs from plasma samples, and the amount proteins recovered from the resulting EV samples were measured by the BCA assay.

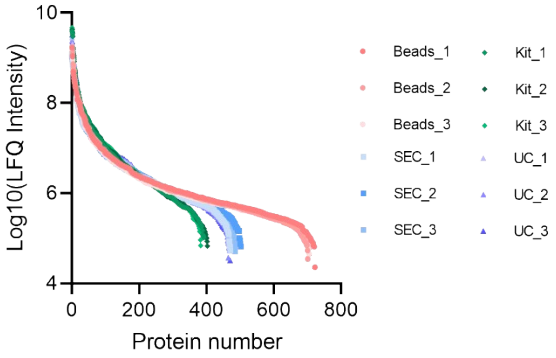

Fig. S5 Dynamic range of proteins in EV samples extracted by four different EV isolation methods based on label-free quantification results. DSPE-functionalized beads ("Beads"), polymer precipitation ("Kit") and ultracentrifugation methods ("UC") and size-exclusion chromatography ("SEC") were used to enrich EVs from plasma samples, and the intensity of proteins were measured using DIA-based mass spectrometry.

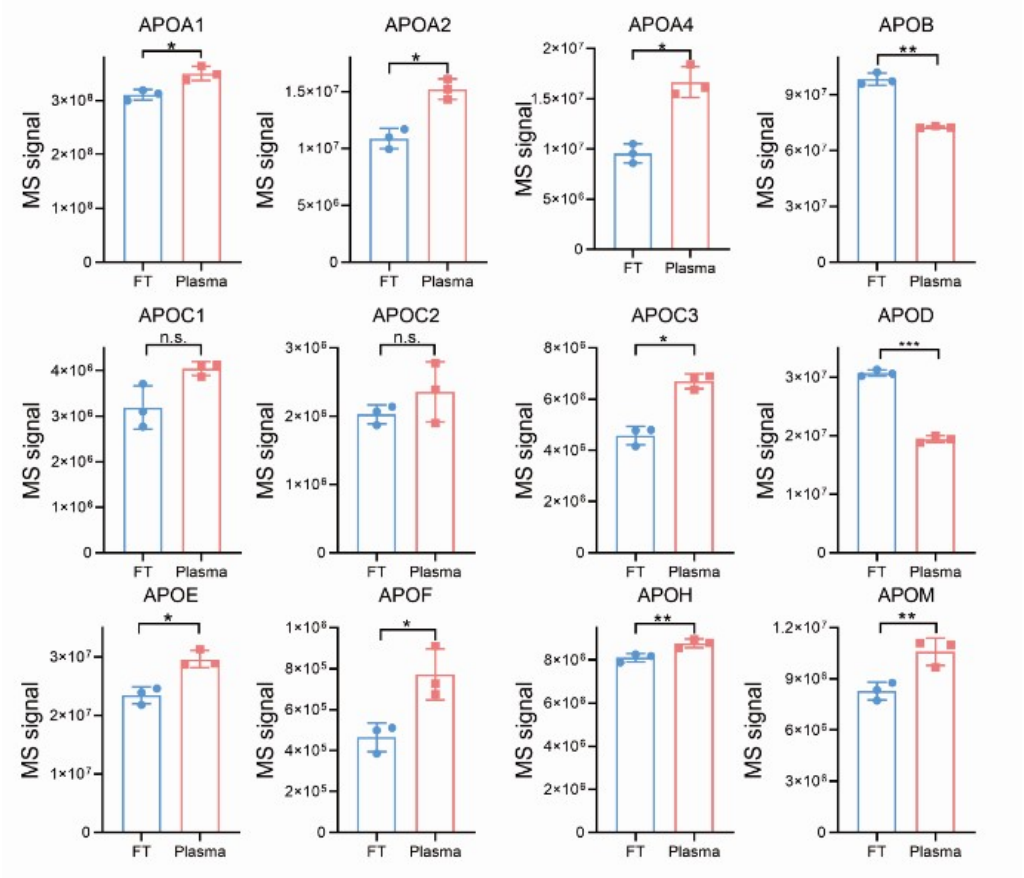

Fig. S6 Quantitative analysis of lipoprotein depletion following DSPE-functionalized bead enrichment. The relative abundance of lipoproteins in plasma (pre-enrichment) and flow-through (FT) samples was measured using DIA-based quantitative mass spectrometry. Data are presented as mean ± SD (n = 3 replicates). Statistical significance was determined using a paired t-test (*p < 0.05, **p < 0.01).

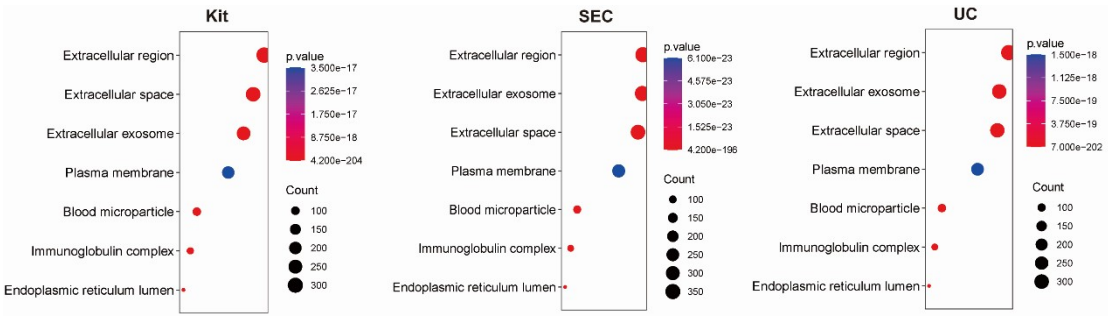

Fig. S7 Gene ontology of cellular components annotations for the identified proteins in EV samples isolated by the polymer precipitation (“Kit”) and ultracentrifugation methods (“UC”) and size-exclusion chromatography (“SEC”), respectively.

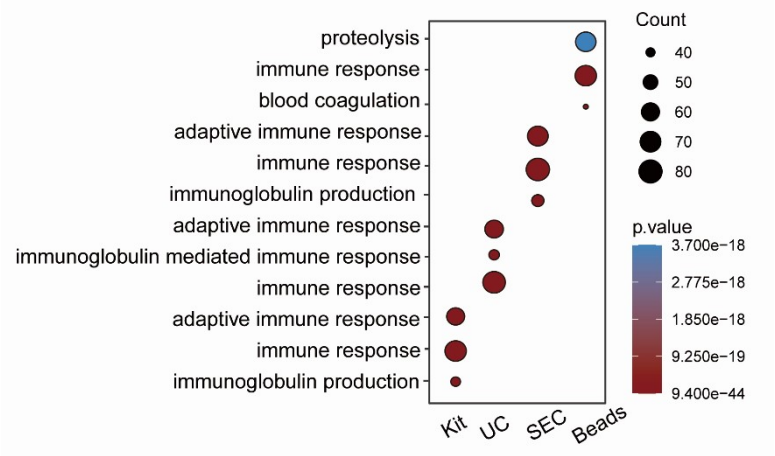

Fig. S8 Functional enrichment analysis of proteins identified by different EV isolation methods. Top 3 enriched molecular pathways from four methods, including DSPE-functionalized beads ("Beads"), polymer precipitation ("Kit"), ultracentrifugation methods ("UC") and size-exclusion chromatography ("SEC").

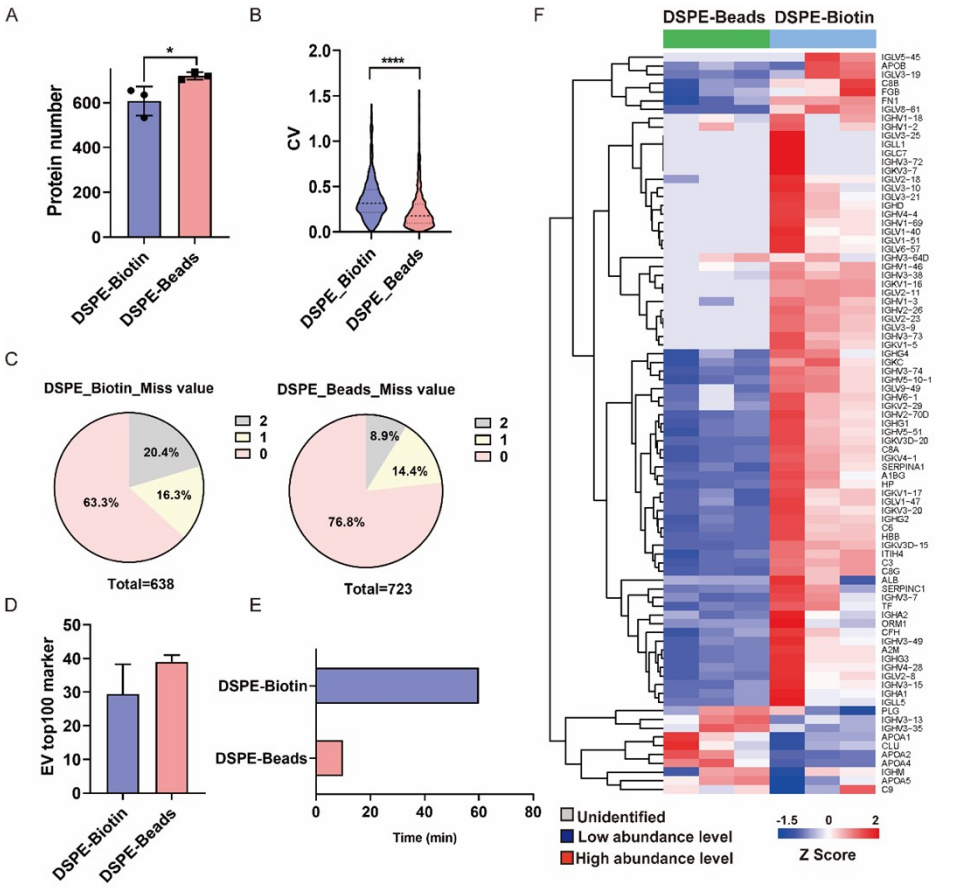

Fig. S9 Comparison of our one-step enrichment method using the DSPE-functionalized beads with the previously reported two-step method using the DSPE-biotin. (A) The number of proteins identified by the label-free MS from EVs isolated by two different methods. (B) Distribution of coefficient of variation (CV) for proteins quantified from EVs isolated by two different methods. (C) Quantitative missing value statistics from EVs isolated by two different methods. (D) Number of the top100 EV marker proteins (as defined in the Exocarta database) identified from EVs

isolated by two different methods. (E) Comparison of time costs of two different methods. (F) Heatmap of the abundance level of high-abundance plasma proteins identified from EV samples isolated by two different methods.

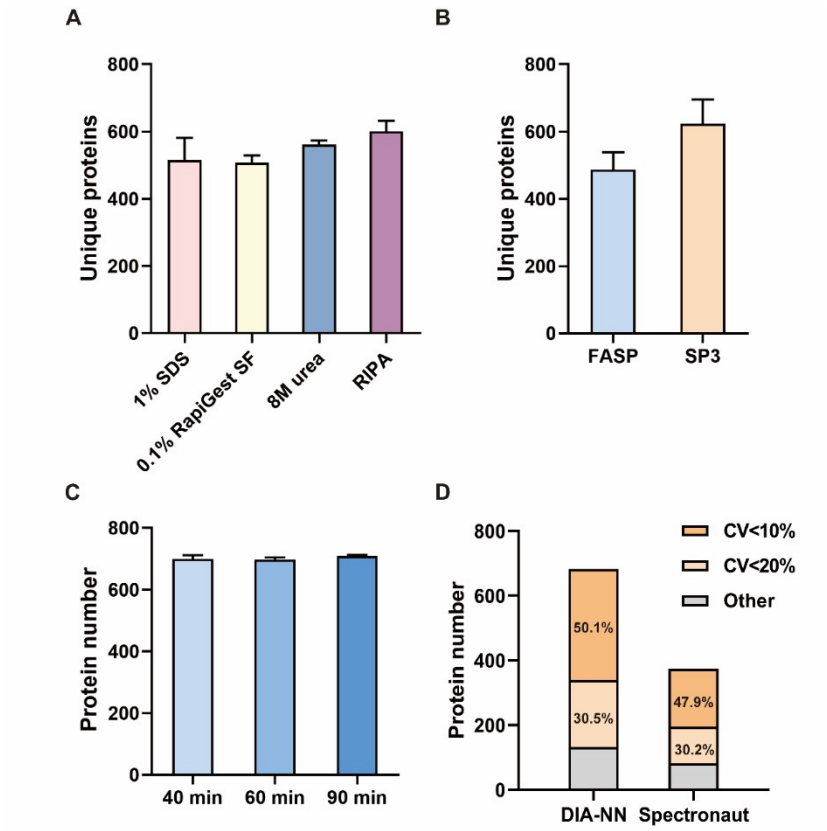

Fig. S10 Optimization of the workflow using the DSPE beads for EVs proteomics. (A) Number of protein identification from EVs samples lysed by different buffers (n=3 replicates). (B) Number of protein identification from EVs samples prepared by SP3 vs FASP (n=3 replicates). (C) Number of protein identification from EVs samples using different LC gradients (n=3 replicates). (D) Number of protein identification from EVs samples using different DIA data analysis software (n=3 replicates).

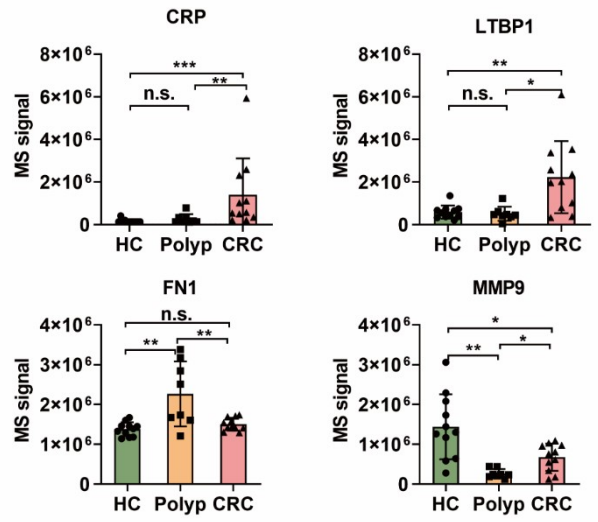

Fig. S11 Abundance distribution of four previously known protein markers as quantified in our EV proteome profiling. For each protein, the raw MS signals were plotted for the samples from each of HC, Polyp and CRC groups.

Fig. S12 Diagram showing the filtering process of selecting features for machine learning. From the sets of identified biomarkers across different groups, the intersection was obtained to derive a common biomarker subset. The random forest algorithm was utilized to prioritize the features in this subset by importance ranking.

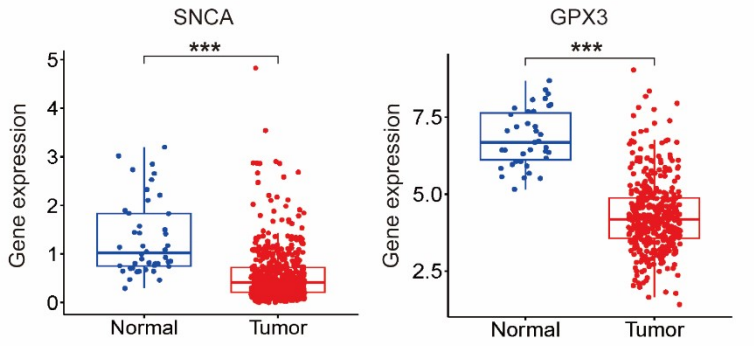

Fig. S13 Relative mRNA abundance of SNCA and GPX3. The data were extracted from cancerous and paracancerous tissues of CRC patients from TCGA database (https://portal.gdc.cancer.gov/).

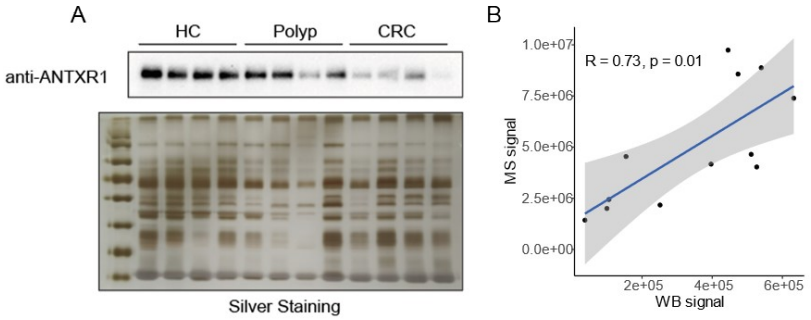

Fig. S14 Validation of ANTXR1 as a representative biomarker and correlation between mass spectrometry and Western blot analysis. (A) Western blot analysis of ANTXR1 expression in EVs isolated from plasma samples at different stages of colorectal disease progression. HC: Healthy Control; Polyp: Polyp; CRC: Colorectal Cancer. (B) Correlation analysis between ANTXR1 abundance as measured by mass spectrometry (MS) and Western blot (WB) signal intensity. Each point represents an individual sample. Pearson correlation coefficient (R) and p-value are indicated.

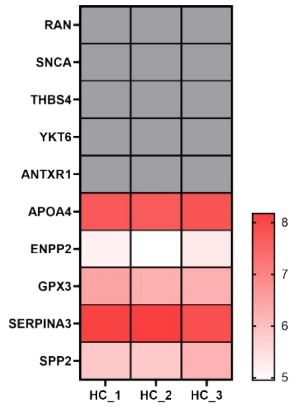

Fig. S15 Identification of the 10 protein biomarkers in plasma samples from three healthy individuals. The color scale indicates the relative abundance of each protein, with red representing high abundance, white representing low abundance, and gray indicating that the protein was not detected.

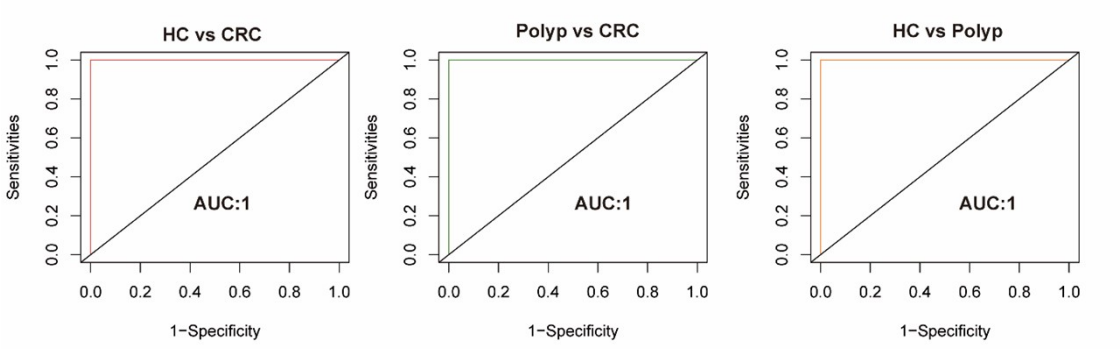

Fig. S16 Performance of the SVM classifier in distinguishing HC, Polyp and CRC samples in the original validation dataset. ROC curves and corresponding AUC values are shown between each of the two subgroups.

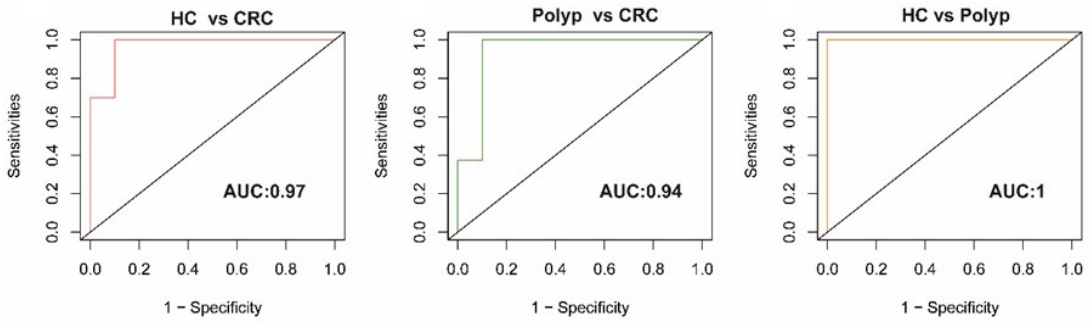

Fig. S17 Performance of SVM classifier in distinguishing HC, Polyp, and CRC samples in the single-blinded validation dataset. ROC curves and corresponding AUC values are shown between each subgroup comparison.

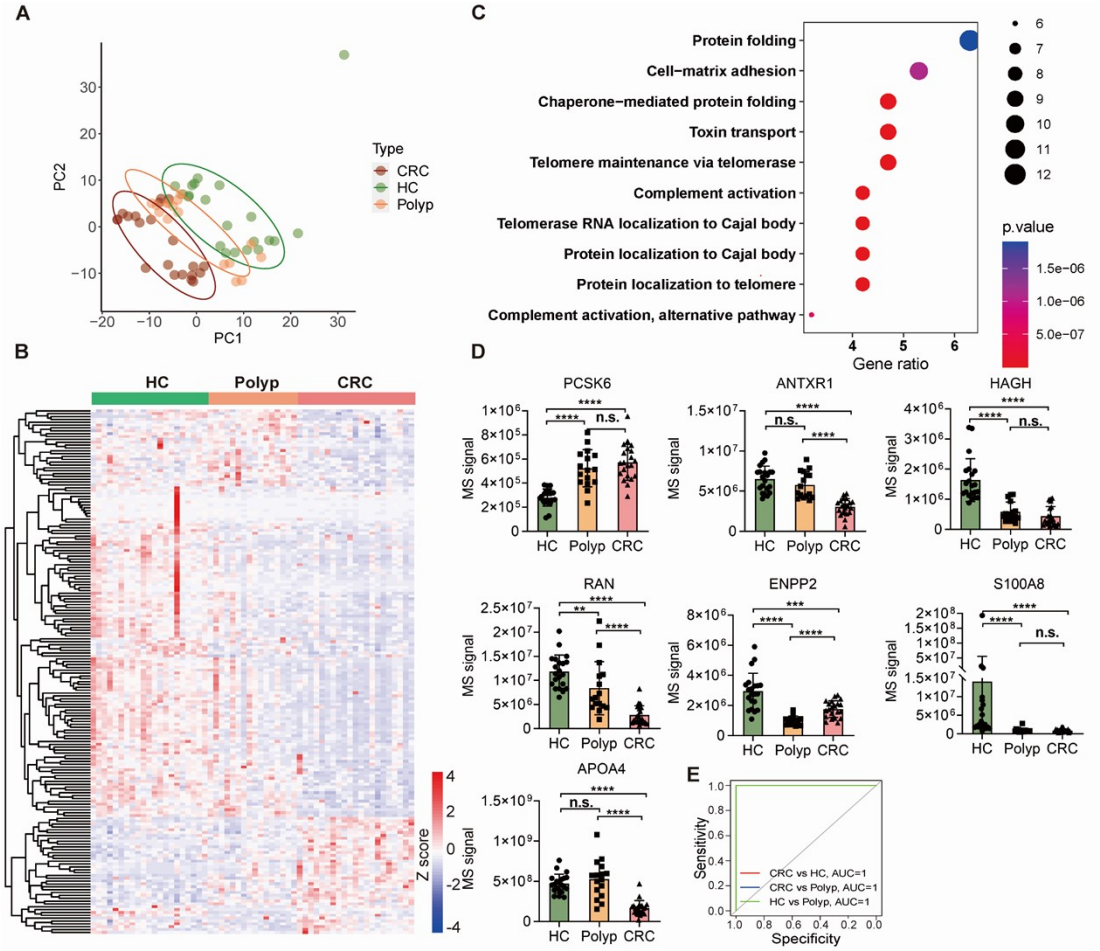

Fig. S18 Integration analysis of EV proteomic data from two separate clinical batches. (A) Principal component analysis (PCA) of EVs proteomics among the healthy controls (HC), adenomatous polyp (Polyp) and early-stage colorectal cancer (CRC). (B) Supervised cluster analysis of across the different groups using proteins dysregulated (VIP>1, p <0.05, AUC>0.85) in EV proteome during disease progression. (C) KEGG pathway enrichment for proteins significantly altered. (D) The distribution of expression levels for seven protein markers with considerable contribution to classification in different groups. Mann-Whitney U test was used to define statistical significance ($p < 0.01$ (**), $p < 0.001$ (***), $p < 0.0001$ (****)). (E) ROC curve of the combined potential biomarker panel.

Supporting information references

1. M. B. Amin, F. L. Greene, S. B. Edge, C. C. Compton, J. E. Gershenwald, R. K. Brookland, L. Meyer, D. M. Gress, D. R. Byrd and D. P. Winchester, The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging, CA Cancer J Clin, 2017, 67, 93-99.

2. V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley and M. Ralser, DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput, Nature Methods, 2019, 17, 41-44.