Supplementary Information for

Beta-sheet richness of the circulating tumor-derived extracellular vesicles for noninvasive pancreatic cancer screening.

Authors: Komila Rasuleva[1], Santhalingam Elamurugan[2], Aaron Bauer[2], Mdrakibhasan Khan[1], Qian Wen[3], Zhaofan Li[4], Preston Steen[5], Ang Guo[6], Wenjie Xia[4], Sijo Mathew[6], Rick Jansen[7,8], Dali Sun[1,2]*

Affiliations:
[1] Department of Electrical and Computer Engineering, North Dakota State University; 1411 Centennial Blvd., 101S Fargo, ND 58102.
[2] Biomedical Engineering Program, North Dakota State University; 1401 Centennial Blvd, Engineering Administration, Room 203, Fargo, ND 58102.
[3] Department of Statistics, North Dakota State University; 1230 Albrecht Blvd, Fargo, ND 58102.
[4] Department of Civil, Construction and Environmental Engineering, North Dakota State University; 1410 North 14th Avenue, CIE 201, Fargo, ND 58102.
[5] Sanford Roger Maris Cancer Center; 820 4th Street N, Fargo, ND 58122.
[6] Department of Pharmaceutical Sciences, North Dakota State University; 1401 Albrecht Blvd, Fargo, ND 58102.
[7] Department of Public Health, North Dakota State University; 1455 14th Ave N, Fargo, ND 58102.
[8] Genomics and Bioinformatics Program, North Dakota State University; 1230 16½ Street North, Fargo, ND 58102.

*Corresponding author. Email: dali.sun@ndsu.edu.

This PDF file includes:

Materials and Methods

Cell lines and cell culture

The human pancreatic cancer cell lines PANC-1, MIA PaCa-2, and the human pancreas cell line HPNE were obtained from the American Type Culture Collection (Manassas, Virginia). PANC-1, and MIA PaCa-2 cells were cultured in DMEM medium (Hyclone, GE Healthcare Life Sciences), and HPNE cells were cultured in DMEM (Hyclone, GE Healthcare Life Sciences) with 0.1ng/mL EGF (Novus Biologicals, USA). All cultures except non-starvation condition were supplemented with 10% fetal bovine serum (FBS; Life technology, Thermo Scientific Inc.), penicillin (1 U) and streptomycin (1 µg/mL). All cells were maintained in a humidified incubator with 5% CO2 at 37 °C. All cell lines were cultured in triplicate under the same conditions and then harvested to collect independent EV samples.

EV isolation from culture media

Cells were grown in culture media with serum until reached to 10^7 cells, washed two times with phosphate-buffered saline (PBS) (pH 7.0). For non-starvation collection, the cells were culture in medium with 10% EV-depleted FBS (Thermo Scientific, US) for 48 hrs. Culture supernatants were then filtered by 0.2 um filter and centrifuged at 10,000 g for 30 min to remove cell debris. The supernatant was carefully centrifuged at 200,000 g for 70 min. Resulting EV precipitates were collected, dissolved in 100 µL PBS (pH 7.0), and stored at 4 °C. TEM (JEOL JEM-2100), and a tunable resistive pulse sensing instrument (qNano system; IZON Science Ltd, Christchurch, New Zealand) were used to validate EV samples (Supporting Information Fig. S2). To remove extravescile contamination, 1 µL proteinase K (8U/mL, New England Biolabs) was added to the samples, and incubated at 37°C for 30 min. 1 µL phenylmethyl sulfonyl fluoride (1mM) was then added and incubated at room temperature for 10 min to stop the digestion. The proteinase K was removed by Ultra - 0.5 centrifugal filter device (50,000 NMWL, Amicon). The impact of extravesicular protein contamination was shown in Fig.S4.

EvIPThT assay

For the spike-in experiment, 5% tumorous EVs were spiked into 200 µL healthy serum as case, and 200µl healthy serum as control. 100 µl Invitrogen™ exosome human CD9 isolation solution or exosome human EpCAM isolation beads (Invitrogen, USA) was washed with PBS and resolved in 400 µL PBS. Serum samples and 100 µL beads were incubated for 2 hrs. at room temperature with proper shaking. After PBS wash, the beads were solved in 20 µL lysis buffer (Sigma-Aldrich, USA) for 15 min. Collected the supernatant, and the protein level was then evaluated by evaluating absorption at 280 nm using Nanodrop. 35 µL PBS was added to the supernatant, and 50 µL sample and 50 µL ThT solution (70 µM) was added to 96 well plate and incubated 30 min avoiding light before reading out with the plate reader. The final fluorescence intensity was normalized by the PBS blank with assigned ThT concentration. The quantitative characteristics of the assay were summarized in Table S3.

Circular Dichroism Measurements

The circular dichroism (CD) spectra were recorded on a Jasco J-815 spectropolarimeter (Tokyo, Japan), using a cylindrical cuvette with 0.1 cm path length. The light source system was protected by nitrogen (flow rate: 5 L/min). Proteins were obtained from cell lysate using a lysis kit (C2978, Sigma-Aldrich) following the manufacture protocol. The samples' protein contents were quantified by absorption at 280 nm in a Nanodrop ND-1000 spectrophotometer (Thermo Scientific) and diluted to 0.2 mg/ml before CD scanned at 200 nm/min in the wavelength region of 200–260 nm. Three scans were averaged for each CD spectrum. Data were analyzed and processed using the Jasco Spectra Manager 2 software package.

Flow Cytometry

Cells were stained by ThT. Briefly, the cells in the culture dish were treated with 2.0 mL of Trypsin-EDTA solution for 3 min at 37 °C, and 2 mL growth medium were added after. The cell suspensions were centrifuged to obtain the cell pellet and washed three times with PBS before incubated for 20 minutes at 37 °C in 100 µM dye. The stained cells were washed three times before diluting to 300 µL in FACS tubes. Unstained cells were used as negative controls. Cells were then subjected to flow cytometric analysis using a BD Accuri™ C6 (BD Bioscience). At least 10,000 cells were acquired from each sample. FITC channel was used to capture signal from the ThT. Flowing Software (Turku Centre from Biotechnology) was used for analysis of the cytometric data. We defined the target cell populations based on negative control. The intensity was normalized to the mean of the control.

Attenuated total reflection Fourier-transform infrared spectroscopy (ATR-FTIR)

ATR-FTIR spectra were collected using a Vertex 70 Fourier Transform Infrared Spectrometer (Bruker Optics, USA) equipped with a liquid nitrogen cooled mercury-cadmium-telluride (MCT)

detector and with a 'Golden Gate' single reflection diamond ATR accessory. All samples were diluted to the same protein level (0.5mg/mL) before testing. 5 µL of sample was mounted on the diamond ATR crystal under ambient conditions. The measurements were performed at room temperature, immediately after sample loading. 100 scans were co-added at a nominal resolution of 4 cm^{-1}. After each data acquisition ATR correction was performed. For all spectral manipulation, the OPUSTM and OPUSLabTM software package (Bruker Optics, USA) was used. The secondary derivative evaluation following a well-developed protocol. Briefly, Calculation of the second derivative with a seven-pint Savitsky-Golay smoothing; Subtraction of PBS buffer as a blank spectrum; Determination of protein content by integrated intensity of α-helix band with a peak at 1653 cm^{-1} as A_α and β-sheet with a peak at 1633 and 1644 as A_β. Calculation of 'spectroscopic β-to-α ratio': $\beta/\alpha = A_\beta / A_\alpha$.

Thioflavin T (ThT) staining assay

Thioflavin T (ThT) were purchased from Santa Cruz Biotechnology (sc-359849). 4mM Stock concentrations of ThT were made with PBS and filtered with a 2 µm syringe filter, and stored at 4ºC. Before testing the samples, the dynamic range and linearity of the ThT staining were determined (Supporting Information Fig. S3). The concentration gradients (0.01 to 1mg/mL) of Bovine Serum Albumin (BSA) was created, and tested by ThT staining following the protocol: 50µL 40 µM ThT solution in PBS is incubated with 50 µL sample for 15 min, and monitored in a fluorescence plate reader (excitation 440 nm/emission 490 nm)[1]. 50 µL of samples were stained for 30 min at room temperature with a ThT final concentration of 35 µM in a black 96-well microplate. The ThT fluorescence was measured at room temperature using a Synergy H1 Hybrid Multi-Mode Reader (BioTek, USA) through the top of the plate with excitation filter of 450 nm and emission filter of 482 nm. The concentration of ThT in solutions was determined with a spectrophotometer, using Beer's law and a molar extinction coefficient of ε_{412} = 31600 /M·cm[2,3].

Proteomics analyses

Comparative proteomics analysis was conducted by reanalyzing LC-MS/MS results from our previous study[4]**.** Briefly, HPNE and MIA PaCa-2 cell and EV lysates were diluted to 1 µg/µL with 100 mM NH_4HCO_3 supplemented with 10 mM dithiothreitol, incubated at 37 °C for 1 h, then mixed with 30 mM iodoacetamide, incubated in the dark for 30 min at room temperature before overnight digestion with 1 µg trypsin at 37 °C. Digestions were terminated by the addition of 0.1 % trifluoroacetic acid and diluted to 0.25 µg/µL protein with H_2O/acetonitrile (95:5), centrifuged at 21,000g for 20 min, and processed using ZipTip columns (Thermo Scientific) for salt removal prior to LC-MS/MS analyses that employed 5 µL of sample/injection. Samples were analyzed using an Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer (Thermo Scientific). Survey scans were collected from 500–2000 Th with an AGC target of 500 000, a resolution of 120 000 at 200 m/z, and 4 µscans averaged per spectrum. MS/MS scans were collected with a resolution of 60 000 at 200 m/z, with 4 µscans averaged per spectrum, an AGC target of 500 000, and a maximum injection time of 118 ms. All resulting MS/MS spectra were used to search Proteome Discoverer™ Software (Thermo Scientific), using a measurement tolerance of 0.5 Da.

Bioinformatics

The proteomics study provides lists of proteins with accession IDs that each point to a record in the protein databases. UniProt (https://www.uniprot.org/), a high-quality database releasing structure information of protein, was accessed for the protein structural information. R package "rvest" is used to extract secondary structure information of the proteins from UniProt database. Briefly, URL link was generated with accession ID for each protein following UniProt format. The HTML response from the target URL was then read ("read_html") and analyzed ("read_node"). Detailed table of secondary structure, primary sequence and the sequence length were unwrapped using their corresponding xpath selectors ("read_table" and "read_text"). The location and length of the secondary structure (beta-strand, helix, and turn) segments of the protein extracted from the table

were used to calculate the percentage of each secondary structure type. By analyzing the structure of each protein, the β-sheet percentage (BP) of the protein was defined as: the length of β–sheet sequence divided by the total length of the protein. In cases that structural information for the integrated protein is not available, an estimator P, defined as the percentage of the known structure partial over the protein's full length, was calculated.

Protein folding simulation.

The primary sequences of the proteins obtained from the proteomics study were retrieved from Uniprot by a getUniProt function in R package "UniprotR". The sequences were folded with PPSPred[5], a machine-learning algorithm for structural prediction. The Perl based algorithm was performed parallelly using high-performance computing resources at North Dakota State University Center for Computationally Assisted Science and Technology (CCAST). The β-sheets percentage of each sequence was extracted from the output files of the simulation.

Protein-protein interaction and enrichment analysis

Protein-protein interaction results using the STRING database (https://string-db.org). The STRING interaction map was generated using default settings (medium confidence of 0.400; criteria for linkages are neighborhood, gene fusion, co-occurrence, co-expression, experiments, databases and text mining). We constructed a PPI network using the Cytoscape software 3.8.1[6] to visualize their interactions based on our relative quantity (Q), BP result data and PPI score. The GO term enrichment was analyzed with STRING functional enrichment analysis[7] inputting IDs and corresponding BR values.

Western blotting

Cells ($2x10^6$), EVs were collected and lysed with 100 μL lysis buffer for 30 min. Total proteins of each sample were quantified by BCA. After the addition of a loading buffer containing 0.125 M Tris–HCl, pH 6.8, 4% SDS, 20% glycerol,10% 2-mercaptoethanol, 0.004% bromophenol blue, an equal amount of protein for each sample was separated by 12% SDS-PAGE and transferred to a PVDF membrane (BioRad). Incubation was conducted in diluted primary antibodies to EpCAM, CD9 (Santa Cruz) at 4 ◦C overnight. Then, membranes were probed with HRP-conjugated secondary antibody (Dako), and signals were visualized using an enhanced chemiluminescence reagent (Luminol).

Gemcitabine Treatment

Cells were seeded in 96-well plates at a density of 10^4 cells/well. After 24 hr., the medium of treatment group was replaced with medium supplemented with 20μM gemcitabine for 48 hr. Cell viability was accessed by CCK-8 assay following the manufacturer's instructions. Briefly, a mixture of 10 μL of CCK-8 and 190 μL media was added into each well, and the cells were incubated for another 1 hr. The absorbance of each well was measured at 450 nm using a microplate reader. Each experiment was performed in 6 repeats.

Statistics

Comparisons between two groups were performed using an unpaired two-tailed Mann–Whitney U-test (unpaired samples), a paired two-tailed Mann–Whitney U-test (paired samples), and a two-tailed Student's t-test (normally distributed parameters). Multiple samples were compared using a Sidak multiple comparison test, Kruskal–Wallis test (non-grouped) and ANOVA with Friedman test for multiple comparisons (grouped). All comparison groups had equivalent variances. $p < 0.05$ was considered to be statistically significant. Data analysis was performed using Oroign Pro software. Data are presented either as representative examples or means ± SEM of 3+ experiments. p values

were obtained using unpaired two-tailed Student's t test or two-way ANOVA. $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$, $****p<0.0001$. The cutoff values were estimated at various sensitivities and specificities and determined at the maximum Youden index value, that is, sensitivity + specificity – 1[8]. The diagnostic accuracy of tests were evaluated using: Accuracy = (sensitivity)(prevalence) + (specificity)(1 - prevalence)[9].

Code availability

R and Perl code as described in the manuscript is available at GitHub at https://github.com/biosunlab/EvIPThT. Additional reasonable requests for code will be promptly reviewed by the senior authors to verify whether the request is subject to any intellectual property or confidentiality obligations, and shared to the extent permissible by these obligations.

Patient samples

The goal of this study was to validate the hypothesis that tumorous cells and EVs are beta-sheet rich compared with their normal counterparts and demonstrate the clinical potential for cancer screening. All serum samples were obtained from ProteoGenex (ProteoGenex Inc., Culver City, CA) following a study protocol approved by the Russian Oncological Research Center ethic committee (No.: PG-ONC 2003/1), after obtaining informed written consent. Tumors from PDAC patients were subjected to pathological re-review and histological confirmation by two expert PDAC pathologists. A total of 15 PDAC patients (10 males and 5 females with an age range of 48–69 years old) and 6 healthy donors and 9 disease control (pancreatitis) were included in the present study. The patients' information, clinical diagnosis, TNM, and staging system of the world health organization and tumor grade established by histopathological evaluation are included in Table S1. The serum samples were stored in Nalgene storage cryogenic. The CA19-9 concentration was remeasured by a commercially available enzyme immunoassay (#EHCA199, Thermo Fisher Scientific) according to the manufacturer's protocols.

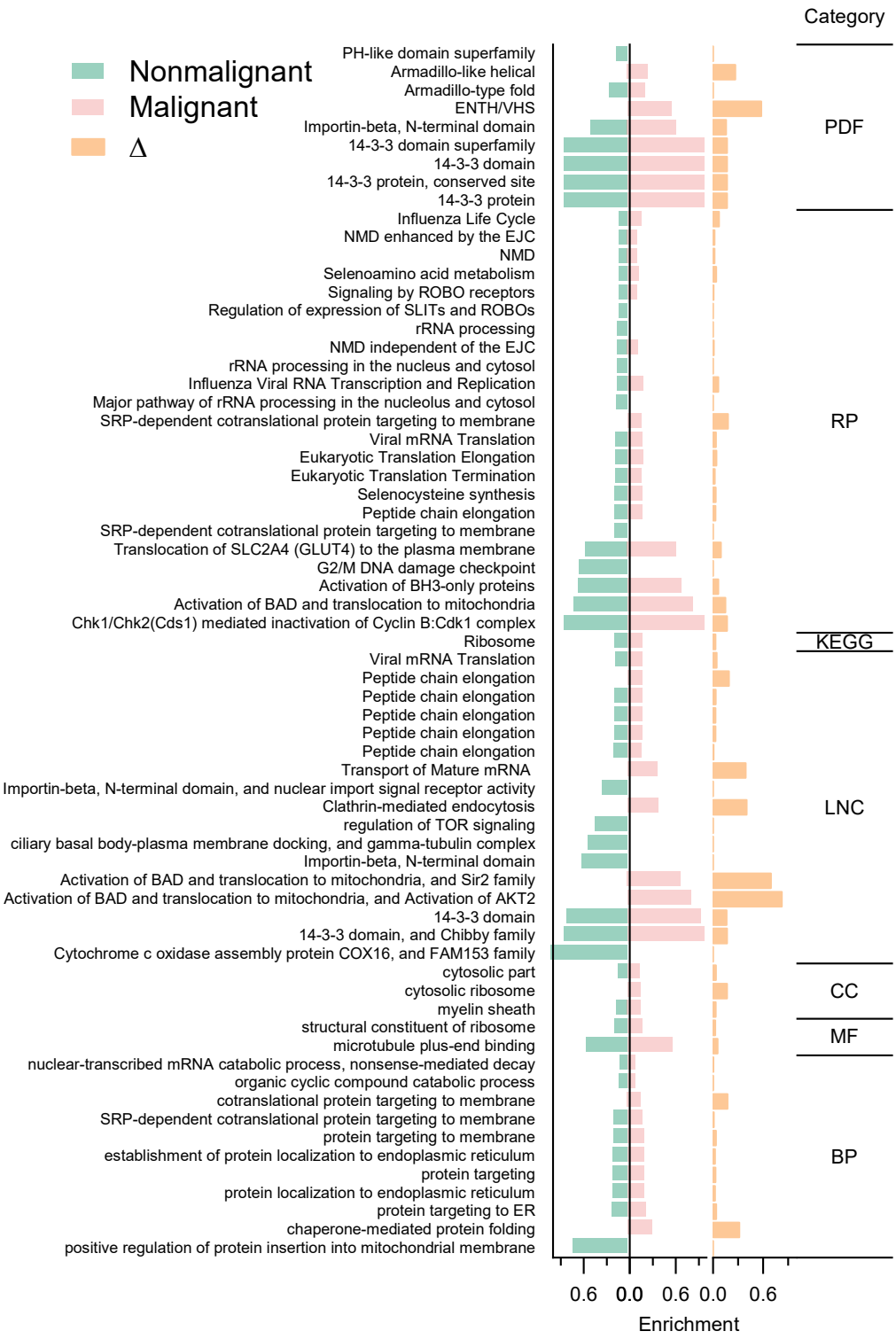

Fig. S1. Functional enrichment analysis result. ΔE denotes the enrichment score difference between nonmalignant and malignant cells. BP: Biological Process; MF: Molecular Function; CC: Cellular Component; LNC: Local Network Cluster; KEGG: KEGG Pathways; RP: Reactome Pathways; PDF: Protein Domains and Features.

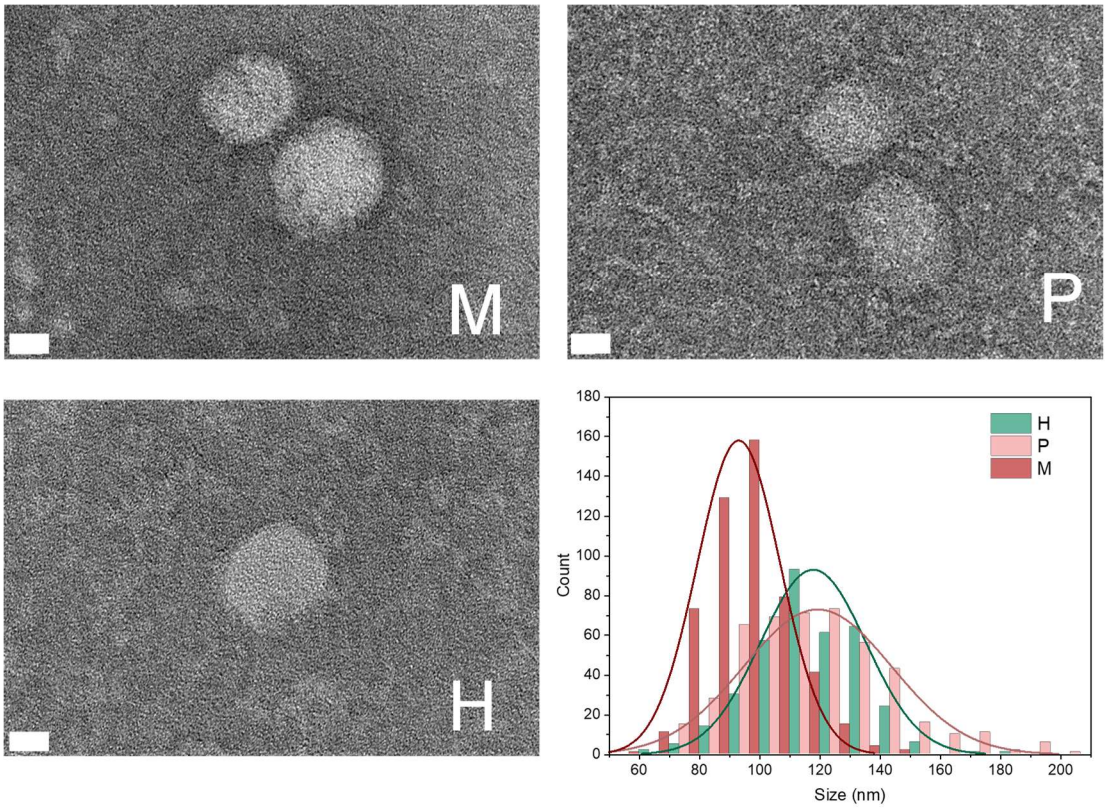

Fig. S2. TEM images and unimodal size distributions of purified EVs from cell culture. Scale bar: 10nm, H: HPNE; M: MIA PaCa-2; P: PANC-1.

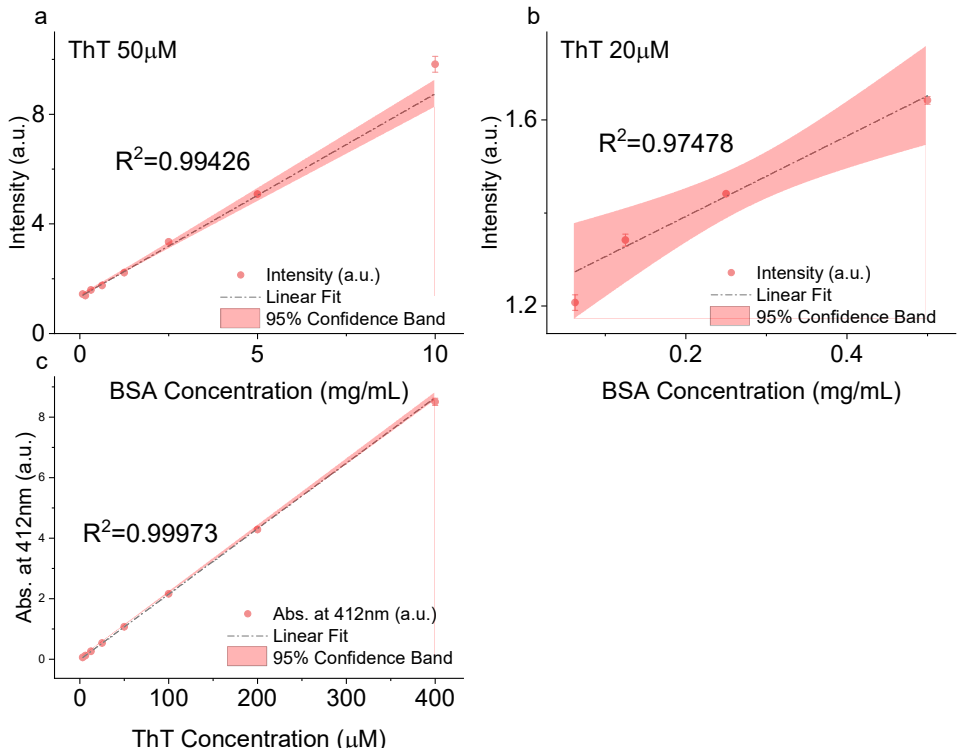

Fig. S3. The dynamic range and linearity of ThT staining. a) and b) Fluorescence response to BSA concentrations with 50 μM and 20 μM, respectively. c) Absorption at 412nm response to ThT titration.

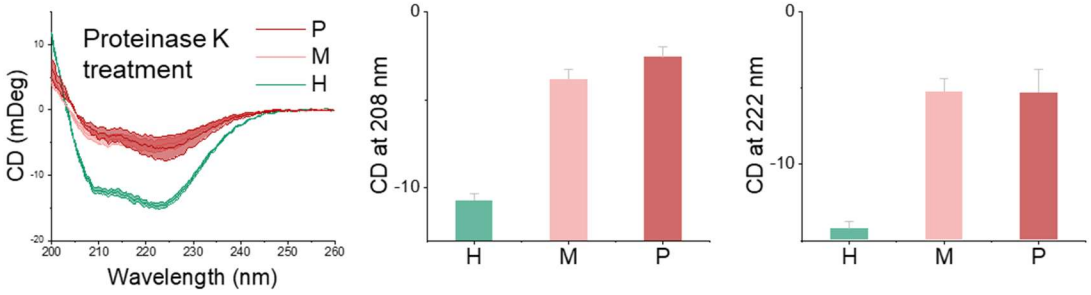

Fig. S4. Impact of extravesicular protein contamination. After proteinase K treatment, the spectrum difference between malignant and nonmalignant EV persists. Circular dichroism spectra, peak value at 222 nm and 208 nm (due to π→π* and n→π* transition respectively) of the EV proteins' CD spectra at a protein concentration of 0.5 mg/ml.

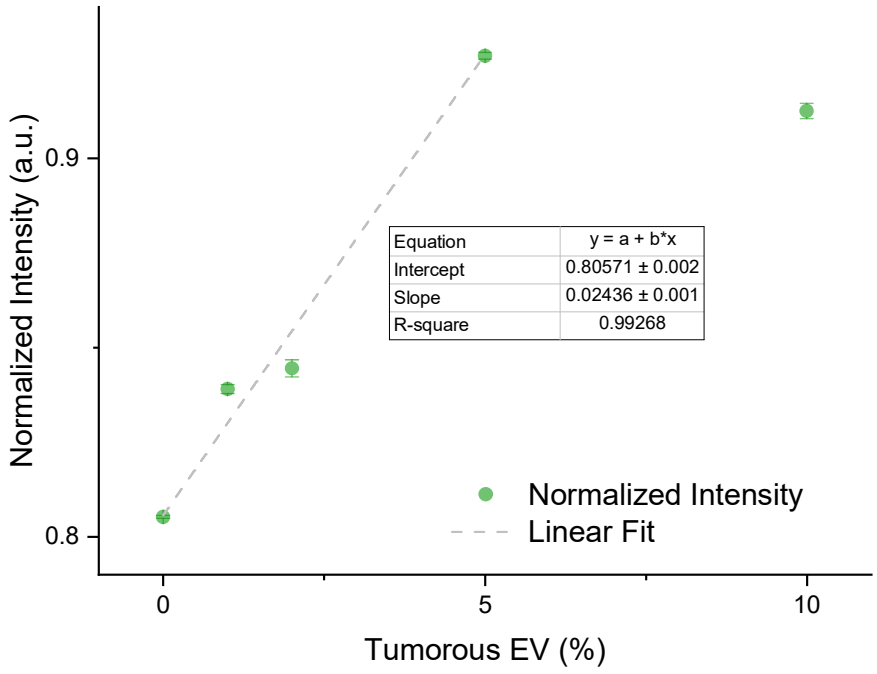

Fig. S5. Linear response of EvIPThT. Tumorous EV was spiked into the healthy human serum to designated protein percentage. The assay readout was linear <5%, and saturated beyond. Error bars, mean ± s.e.m; n=5.

Table S1. Demographic characteristics of the clinical cohort.

Characteristics	Number (n=30)	%
Age, years		
Median(range)	53 (35-69)	
Gender		
Male	16	53
Female	14	47
Ethnicity		
Caucasian	30	100
BMI		
Median(range)	25.65(19.4-29.7)	
Diagnosis		
PDAC	15	50
Pancreatitis	9	30
Healthy	6	20
CEA		
Median(range)	3.815 (1.25-48)	
CA19-9		
Median(range)	0.5845(0.41-149)	
Stage		PDAC
IA	1	6
IB	1	6
IIA	2	13
IIB	5	33
IV	3	10

Table S2. Accuracy of EvIPThT assay for screening PDAC from Healthy.

	Youden Index	Cut-off	Specificity	Sensitivity	**Accuracy**
EvIPThT	0.70	0.25	0.83	0.87	0.86
CA19-9	0.67	0.00	1.00	0.67	0.76

Table S3. Characteristics of EvIPThT assay.

Characteristics	***Value***
Coefficient of variation (CV)	Positive: 16% Negative 14%
Limits of detection (LOD)	0.5%
Limits of quantification (LOQ)	1.83%
Linear range	0~5% (Fig.S5)

Diagnostic Specificity	0.83
Diagnostic Sensitivity	0.87
Diagnostic Accuracy	0.86
Turnaround time	0.75 hr.
Cost per sample	$6

Note: Limits of detection (LOD) and quantification (LOQ) were defined as 3× and 10× the standard deviation of the assay blank, respectively. The percentage of tumorous EV proteins in the sample was used as the unit of LOD, LOQ, and linear range. CV was calculated by three repeats(between-run).

SI References

(1) Shimanovich, U.; Michaels, T. C. T.; De Genst, E.; Matak-Vinkovic, D.; Dobson, C. M.; Knowles, T. P. J. Sequential Release of Proteins from Structured Multishell Microcapsules. *Biomacromolecules* **2017**, *18* (10), 3052–3059. https://doi.org/10.1021/acs.biomac.7b00351.

(2) Sulatskaya, A. I.; Lavysh, A. V.; Maskevich, A. A.; Kuznetsova, I. M.; Turoverov, K. K. Thioflavin T Fluoresces as Excimer in Highly Concentrated Aqueous Solutions and as Monomer Being Incorporated in Amyloid Fibrils. *Sci. Rep.* **2017**, *7* (1), 1–11. https://doi.org/10.1038/s41598-017-02237-7.

(3) Chan, C. X. J.; Joseph, I. G.; Huang, A.; Jackson, D. N.; Lipke, P. N. Quantitative Analyses of Force-Induced Amyloid Formation in Candida Albicans Als5p: Activation by Standard Laboratory Procedures. *PLoS One* **2015**, *10* (6), 1–13. https://doi.org/10.1371/journal.pone.0129152.

(4) Sun, D.; Zhao, Z.; Spiegel, S.; Liu, Y.; Fan, J.; Amrollahi, P.; Hu, J.; Lyon, C. J.; Wan, M.; Hu, T. Y. Dye-Free Spectrophotometric Measurement of Nucleic Acid-to-Protein Ratio for Cell-Selective Extracellular Vesicle Discrimination. *Biosens. Bioelectron.* **2021**, 113058. https://doi.org/10.1016/j.bios.2021.113058.

(5) Yan, R.; Xu, D.; Yang, J.; Walker, S.; Zhang, Y. A Comparative Assessment and Analysis of 20 Representative Sequence Alignment Methods for Protein Structure Prediction. *Sci. Rep.* **2013**, *3*. https://doi.org/10.1038/srep02619.

(6) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**. https://doi.org/10.1101/gr.1239303.

(7) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; Jensen, L. J.; Von Mering, C. STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47* (D1), D607–D613. https://doi.org/10.1093/nar/gky1131.

(8) Youden, W. J. Index for Rating Diagnostic Tests. *Cancer* **1950**. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

(9) Zhu, W.; Zeng, N.; Wang, N. Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations. *Northeast SAS Users Gr. 2010 Heal. Care Life Sci.* **2010**.