Supporting Information

Proteomic and Phosphoproteomic Profiling of Urinary Small Extracellular Vesicles in Hepatocellular Carcinoma

Dejun Li,[a, b] Yujun Gao,[a] Chong Wang*[c] and Lianghai Hu*[a]

[a] Center for Supramolecular Chemical Biology, School of Life Sciences, Jilin University, Changchun 130012. E-mail: lianghaihu@jlu.edu.cn

[b] Prenatal Diagnosis Center, Reproductive Medicine Center, The First Hospital of Jilin University, Changchun 130021, China.

[c] Department of Hepatology, The First Hospital of Jilin University, Changchun 130021, China. E-mail: wang_chong@jlu.edu.cn

Supporting Experimental Section

Nanoparticle Tracking Analysis (NTA)

The size and concentration of sEVs were analyzed with the NanoSight NS300 instrument (NanoSight technology, Malvern, UK), which is equipped with a blue laser (488 nm) and a high-sensitivity video camera. According to the instructions from the manufacturer, sEVs were diluted with 1 × PBS to achieve the best particle per frame values for machine measurement. For each sample, five videos of 60 s were recorded. The camera level was set at 14, and the detection threshold was at 9. The particles per frame values were adjusted to 20–50 for optima counting. After capture, the data were analyzed with NanoSight NTA Software NTA 3.4 Build 3.4.003.

Transmission Electron Microscopy (TEM) Imaging

The isolated sEV samples were first placed on the 300-mesh formvar-carbon film coated copper grid for 20 min at room temperature, and the extra liquids were removed with filter paper. Then, the grid was negatively stained for one minute with 2% uranyl acetate and extra staining was removed with filter paper. Finally, the morphology of the sEV samples were viewed using a JEM-1400 transmission electron microscope operated at 120 kV (Jeol Ltd., Tokyo, Japan).

LC-MS/MS sample preparation

The isolated sEV samples were lysed to extract proteins using the phase-transfer surfactant (PTS) aided procedure as previously reported.[1] First, sEVs were resuspended in the lysis solution containing 12 mM sodium deoxycholate, 12 mM sodium lauroyl sarcosinate, 10 mM Tris (2-carboxyethyl) phosphine, 40 mM 2-chloroacetamide, and phosphatase inhibitor cocktail in 50 mM Tris· HCl (pH 8.5) by incubating 10 min at 95 °C. Then, the samples were diluted five-fold with 50 mM triethylammonium bicarbonate and digested with Lys-C (Wako) at 1:100 (w/w) enzyme-to-protein ratio for 3 h at 37 °C. For further protein digestion, trypsin was added to a final 1:50 (w/w) enzyme-to-protein ratio for overnight digestion at 37 °C. Afterwards, 10% trifluoroacetic acid (TFA) was added to acidify the sample to a final concentration of 1% TFA, and ethyl acetate was added to the abovementioned mixture at a 1:1 ratio to the samples. The resulting mixture was vortexed for 2 min and centrifuged at 15,000 × g for 3 min to obtain aqueous and organic phases. The top layer of organic phases was removed, and the bottom aqueous phase was collected and lyophilized in a refrigerated vacuum centrifuge (Laconco CentriVap). The desalting process was performed using Top-Tip C18 tips (Glygen) according to manufacturer's instructions. Each sample

was split into 98% and 2% aliquots for phosphoproteomic and proteomic experiments, respectively. For phosphoproteomic analysis, the 98% portion of each sample was subjected to phosphopeptide enrichment using PolyMAC Phosphopeptides Enrichment Kit (Tymora Analytical, West Lafayette, IN, USA) according to manufacturer's instructions.[2] All samples were freeze-dried in a refrigerated vacuum centrifuge and stored at − 80 °C.

LC−MS/MS Analysis

The mobile phase buffer consisted of 0.1% formic acid in ultrapure water (buffer A) with an eluting buffer containing 0.1% formic acid in 80% (vol/vol) acetonitrile (buffer B). Dried peptides and phosphopeptides were redissolved in 10 μL of 0.1% formic acid (FA) and were separated on a 25 cm in-house packed column (360 μm OD × 150 μm ID) containing C18 resin (2.2 μm, 100 Å; Michrom Bioresources) at a flow rate of 300 nL/min with a linear 52 min gradient from 10% to 32% B, followed by an 8 min washing gradient. The EASY-nLC 1200 was coupled online with a Q Exactive HF-X mass spectrometry (Thermo Fisher Scientific, Bremen, Germany). The mass spectrometer was operated in the data-dependent mode with the following parameters: Spray voltage was set to 2.1 kV, RF level at 40, and heated capillary at 320 °C. Mass range was set to 350–1500. Full MS resolutions were set to 60,000, and the full MS AGC target was 3×10^6 with a maximum inject time (IT) of 30 ms. The top 20 most intense precursor ions from a full MS spectrum were automatically selected for subsequent MS/MS fragment analysis with the following parameters: The AGC target for fragment spectra was set at 1×10^5, and the resolution threshold was kept at 30,000 with an IT of 50 ms. Isolation width was set at 1.6 m/z. The normalized collision energy was set at 27%. Only precursors charged between +2 and +7 that achieved a minimum AGC of 4×10^3 were acquired. Dynamic exclusion was set to 40 s to exclude all isotopes in a cluster.

LC-MS/MS Data Processing

For the global peptides and phosphopeptides, the acquired MS/ MS spectra were searched using the MaxQuant software (version 2.4.0.0) with the UniProt Human database downloaded in May 2023, which contains 20,422 proteins. The parameters were set as default if not otherwise stated. Trypsin/P was specified as cleavage enzyme allowing up to 2 missing cleavages. Carbamidomethylation on cysteines (+57.0214 Da) was set as the fixed modification, oxidation (+15.9949 Da) on methionine residues and acetylation (+42.011 Da) at N terminus of proteins were selected as variable modifications. The variable modifications of phosphorylation (+79.996 Da) on serine, threonine, or

tyrosine residues were also included for the phosphopeptides sample search. The false discovery rates (FDRs) of proteins, peptides, and phosphopeptides were all set to 1%. The minimum peptide length was set to 7 amino acids, and a minimum Andromeda score was set at 40 for modified peptides. For label free quantification method, match between run was enabled.

MaxQuant derived data were submitted to Perseus software (version 2.0.9.0) for subsequent processing. First, proteins matched to potential contaminant and reverse database were excluded. Then, for the phosphoproteomic data, only phosphosites with localization probabilities > 0.75 were kept. Proteins or phosphopepsites were further filtered with the condition of there being at least 70% valid values in at least one group. Imputation for the missing abundances was replaced by normal distribution with a downshift of 1.8 SDs and a width of 0.3 SDs. The processed dataset was exported as a data matrix for further statistical analysis. Student's t-test was used to identify differentially expressed proteins (DEPs), differentially expressed phosphosites (DEPSs), and differentially expressed phosphoproteins (DEPPs) between the HCC and NC groups. Fold change (FC) was calculated by dividing the mean values of HCC group by that of NC group. Statistically significant DEPs/DEPSs/DEPPs were defined as $|\log_2 FC| > 1$ and p value < 0.05. The volcano plots and heatmaps were generated with "ggplot2" and "complexheatmap" packages in R software (version 4.3.2).

Bioinformatics Analysis

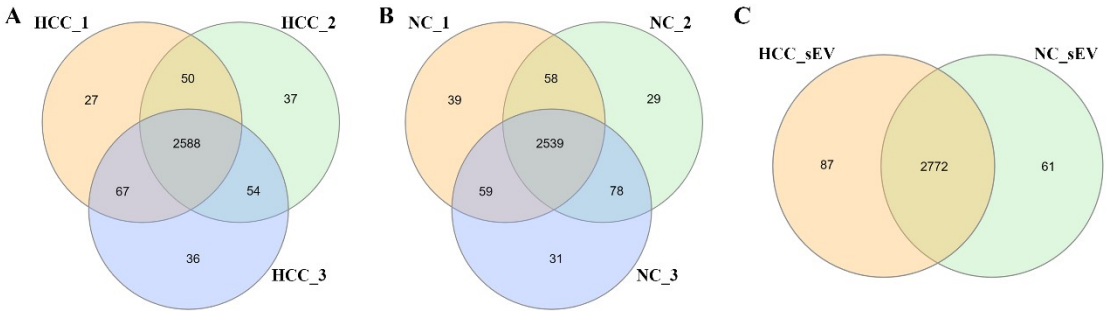
Gene Ontology (GO) enrichment analysis was conducted using the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resources 6.8 (https://david.ncifcrf.gov/).[3] GO annotation was classified into three categories, namely cellular component, biological process, and molecular functions. GO terms with a p value < 0.05 was considered significant. Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway enrichment analysis was performed by KOBAS (http://kobas.cbi.pku.edu.cn/).[4] Pathways with p value < 0.05 were considered to be significantly enriched. The proteins/phosphoproteins in the enriched KEGG pathways were submitted to STRING database (https://string-db.org/) for protein-protein interaction (PPI) analysis, with the interaction score threshold set at the highest confidence level (≥ 0.9). Cytoscape software 3.9.1(http://cytoscape.org/) was then utilized to construct and visualize the PPI network. Molecular Complex Detection (MCODE) was employed to select hub modules, and cytoHubba was utilized to calculate the top 15 hub genes of the constructed PPI

network by using six different ranking algorithms (Betweenness, Closeness, Degree, EPC, Radiality, and Stress).[5] WoLF PSORT (https://wolfpsort.hgc.jp/) was used to predict and characterize the subcellular localization of the identified proteins/phosphoproteins. Kinase-substrate enrichment analysis (KSEA) was performed to reveal the activation/inactivation of kinase with the KSEA algorithm in the KSEA app (https://casecpb.shinyapps.io/ksea/)[6] using the PhosphoSitePlus and NetworKIN database. The threshold of kinase activity was predicted according to the following parameters: NetworKIN score cutoff ≥ 2, substrate count cutoff ≥ 5, and p value cutoff < 0.05. The kinome tree was visualized using the Coral app (http://phanstiel-lab.med.unc.edu/CORAL/).[7]

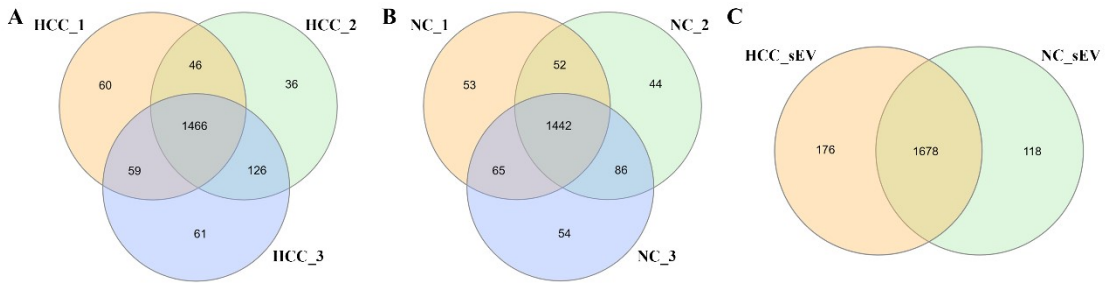
References

1. T. Masuda, M. Tomita and Y. Ishihama, *Journal of Proteome Research*, 2008, **7**, 731-740.
2. A. B. Iliuk, V. A. Martin, B. M. Alicie, R. L. Geahlen and W. A. Tao, *Molecular & Cellular Proteomics*, 2010, **9**, 2162-2172.
3. D. W. Huang, B. T. Sherman and R. A. Lempicki, *Nature Protocols*, 2008, **4**, 44-57.
4. C. Xie, X. Mao, J. Huang, Y. Ding, J. Wu, S. Dong, L. Kong, G. Gao, C.-Y. Li and L. Wei, *Nucleic Acids Research*, 2011, **39**, W316-W322.
5. C.-H. Chin, S.-H. Chen, H.-H. Wu, C.-W. Ho, M.-T. Ko and C.-Y. Lin, *BMC Systems Biology*, 2014, **8**.
6. D. D. Wiredja, M. Koyuturk and M. R. Chance, *Bioinformatics*, 2017, **33**, 3489-3491.
7. K. S. Metz, E. M. Deoudes, M. E. Berginski, I. Jimenez-Ruiz, B. A. Aksoy, J. Hammerbacher, S. M. Gomez and D. H. Phanstiel, *Cell Systems*, 2018, **7**, 347-350.e341.

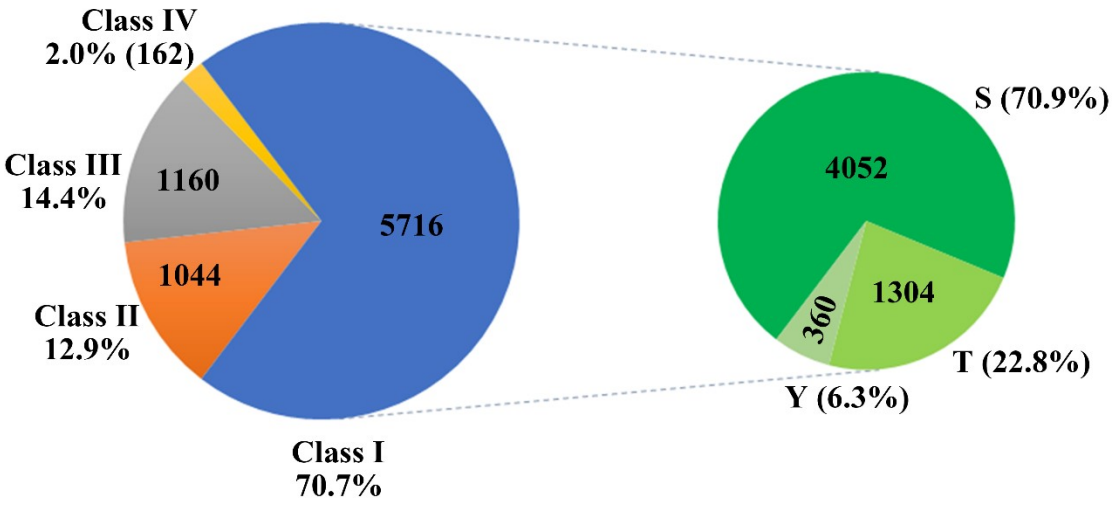
Supplementary Figures



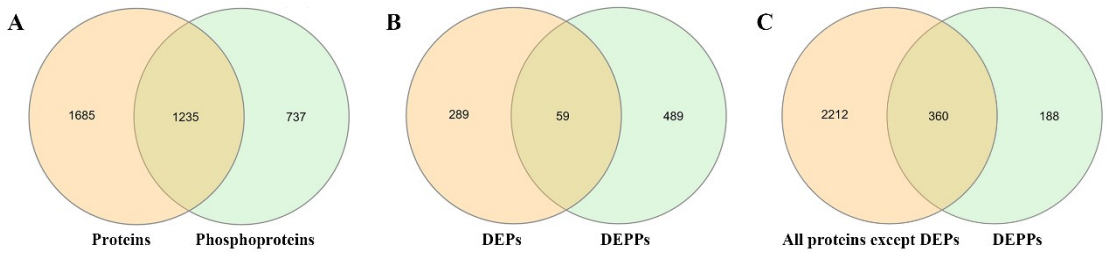
Supplementary Figure S1. Proteomics of urinary sEVs from HCC patients and normal controls (NC). (A) Venn diagram of the identified proteins among HCC groups; (B) Venn diagram of the identified proteins among NC groups; (C) Venn diagram depicted the overlap of identified proteins between the HCC and NC groups.



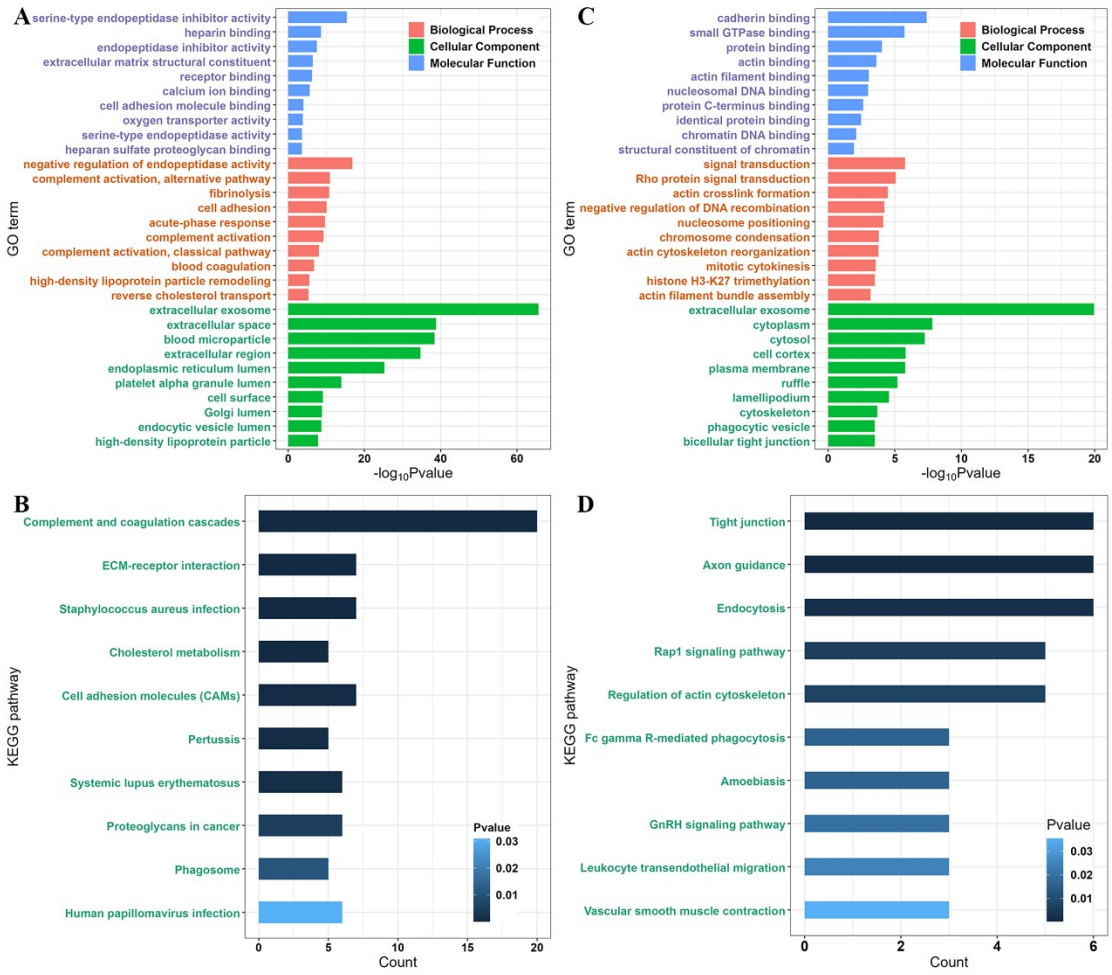
Supplementary Figure S2. Phosphoproteomics of urinary sEVs from HCC patients and normal controls (NC). (A) Venn diagram of the identified phosphoproteins among HCC groups; (B) Venn diagram of the identified phosphoproteins among NC groups; (C) Venn diagram depicted the overlap of identified phosphoproteins between the HCC and NC groups.



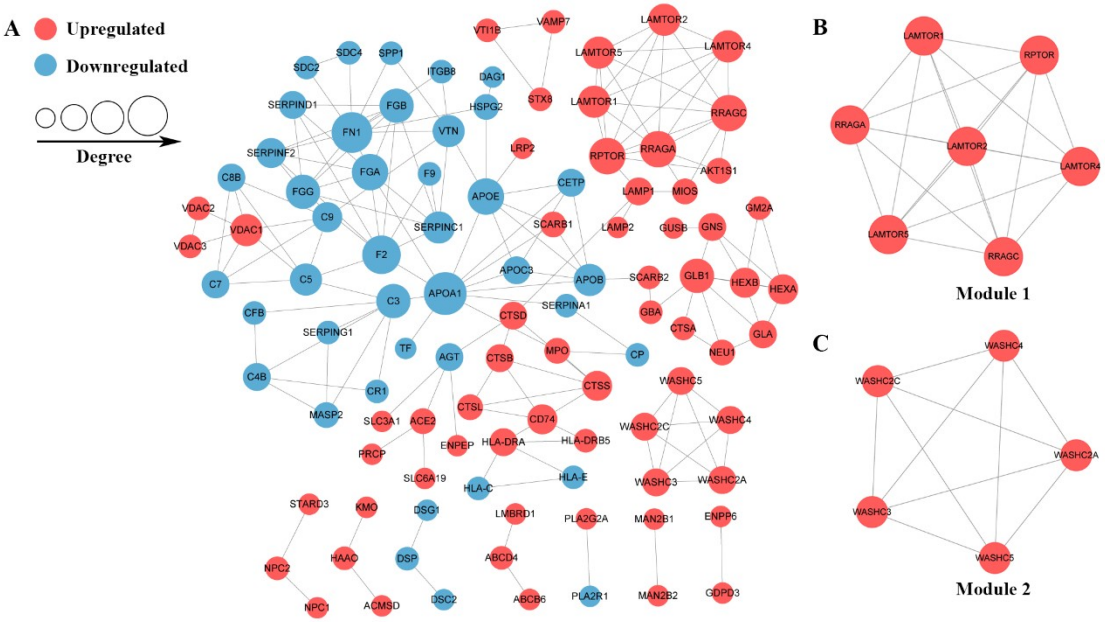
Supplementary Figure S3. The proportion of four classes in the identified phosphosites and the distribution of phosphorylated serine (S), threonine (T), and tyrosine (Y) in Class I phosphosites.



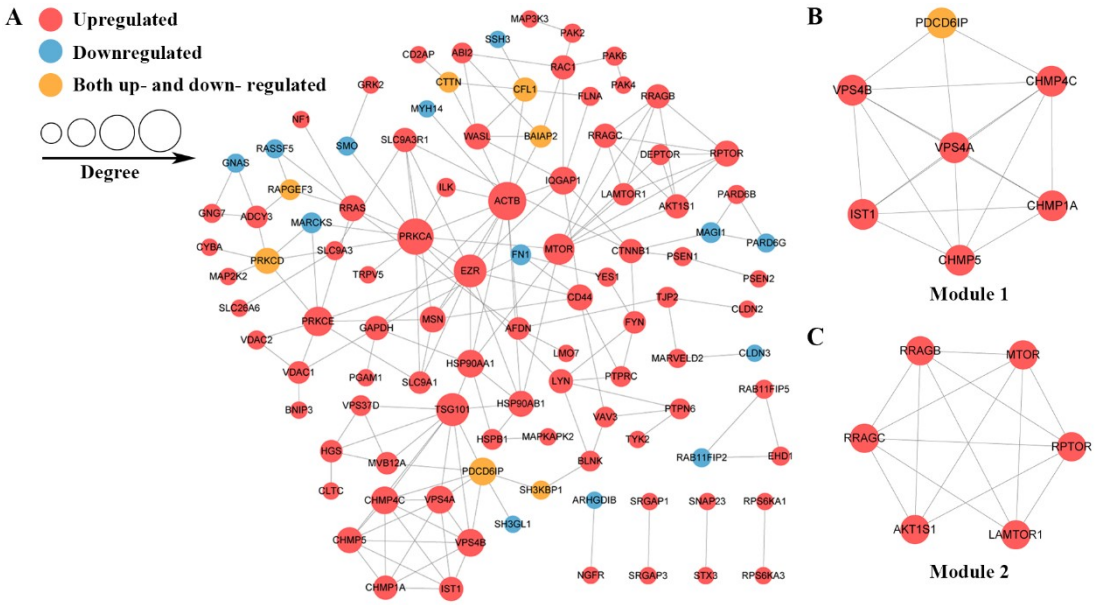
Supplementary Figure S4. Integration analysis results of the proteomics and phosphoproteomics data. (A) The overlap of identified proteins and phosphoproteins. (B) The overlap of dysregulated proteins and phosphoproteins. (C) The overlap of identified proteins (excluding DEPs) with DEPPs.



Supplementary Figure S5. GO analysis and KEGG pathway enrichment analysis of downregulated DEPs and DEPPs. (A) GO analysis of DEPs with downregulated proteins. (B) KEGG pathway analysis of DEPs with downregulated proteins. (C) GO analysis of DEPPs with downregulated phosphosites. (D) KEGG pathway analysis of DEPPs with downregulated phosphosites.



Supplementary Figure S6. Protein-protein interaction (PPI) network analysis of proteins in the enriched KEGG pathways. The two most significant modules were identified by the molecular complex detection (MCODE) algorithm. The size of nodes represents the degree value. (A) PPI network of proteins in the enriched KEGG pathways. (B) Module 1 (MCODE score = 7) was constructed with 7 nodes and 21 edges. (C) Module2 (MCODE score = 5) was composed of 5 nodes and 10 edges.



Supplementary Figure S7. Protein-protein interaction (PPI) network analysis of phosphoproteins in the enriched KEGG pathways. The two most significant modules were identified by the molecular complex detection (MCODE) algorithm. The size of nodes represents the degree value. (A) PPI network of enriched phosphoproteins in the KEGG pathways. (B) Module 1 (MCODE score = 6) was constructed with 7 nodes and 18 edges (C) Module 2 (MCODE score = 5.6) was composed of 6 nodes and 14 edges.

Supplementary Table S7. The top 15 hub genes for PPI network of DEPs were calculated by six algorithms

Category	Rank methods in cytoHubba					
	Betweenness	**Closeness**	**Degree**	**EPC**	**Radiality**	**Stress**
Gene symbol top 15	APOA1	APOA1	APOA1	F2	APOA1	APOA1
	CTSS	F2	FN1	APOA1	MPO	CTSS
	MPO	FGA	F2	FGA	APOE	MPO
	CTSD	APOE	APOE	FGG	F2	CTSD
	APOB	C3	FGA	FN1	APOB	APOB
	LAMP1	FN1	RRAGA	FGB	FGA	LAMP1
	SCARB2	VTN	RRAGC	SERPINC1	CTSS	SCARB2
	GBA	APOB	RPTOR	SERPINF2	C3	GBA
	GLB1	MPO	C3	APOE	APOC3	GLB1
	C3	C5	FGG	VTN	CETP	F2
	RPTOR	FGG	FGB	APOB	SCARB1	C3
	APOE	FGB	GLB1	SERPIND1	SERPINA1	C5
	F2	CTSS	VDAC1	APOC3	VTN	APOE
	AGT	APOC3	C5	CETP	TF	RPTOR
	C5	CETP	VTN	C5	CTSD	FGA

Supplementary Table S8. The top 15 hub genes for PPI network of DEPPs were calculated by six algorithms

Category	Rank methods in cytoHubba					
	Betweenness	**Closeness**	**Degree**	**EPC**	**Radiality**	**Stress**
Gene	ACTB	ACTB	ACTB	EZR	ACTB	TSG101
symbol	PRKCA	PRKCA	PRKCA	ACTB	EZR	ACTB
top 15	TSG101	EZR	TSG101	PRKCA	PRKCA	PRKCA
	HSP90AA1	HSP90AA1	EZR	HSP90AA1	HSP90AA1	HSP90AA1
	IQGAP1	HSP90AB1	MTOR	MSN	HSP90AB1	EZR
	EZR	IQGAP1	PRKCE	PRKCE	IQGAP1	HSP90AB1
	MTOR	MTOR	VPS4B	HSP90AB1	MSN	IQGAP1
	HSP90AB1	MSN	VPS4A	IQGAP1	GAPDH	MTOR
	CTNNB1	PRKCE	CHMP4C	CD44	SLC9A3R1	PRKCD
	AFDN	GAPDH	PDCD6IP	SLC9A1	MTOR	RAC1
	PRKCD	SLC9A3R1	IQGAP1	SLC9A3R1	ILK	RRAS
	SLC9A3R1	TSG101	HSP90AA1	MTOR	CD44	CTNNB1
	RAC1	CD44	CHMP5	WASL	PRKCE	AFDN
	RRAS	RRAS	WASL	GAPDH	SLC9A1	PRKCE
	LYN	WASL	MSN	TSG101	WASL	SLC9A3R1