Supporting Information

Wang et al. 10.1073/pnas.1712731114

SI Materials and Methods

Materials and Reagents. Human plasma depletion Seppro IgY14 LC10 column systems were purchased from Sigma-Aldrich. Tris-(2-carboxyethyl)phosphine (TCEP) and methyl methanethiosulfonate (MMTS) were purchased from Thermo Fisher Scientific. LysC and Trypsin proteases were purchased from Promega. PNGase F was purchased from New England Biolabs. PolySULFOETHYL A column (100 × 2.1 mm, 5 μm, 200 Å) for strong cation exchange (SCX) chromatography was purchased from PolyLC. C18 Cartridges for sample preparation and chromatography columns for bRPLC and online HPLC of triple-quadrupole mass spectrometer were purchased from Waters. All iTRAQ reagents and buffers were purchased from AB Sciex. Synthetic peptides were purchased from Genscript. All other reagents were purchased from Sigma-Aldrich, unless otherwise indicated.

Preparation of Solutions. SCX solvent A contained 10 mM KH_2PO_4, 25% (vol/vol) acetonitrile; SCX solvent B contained 10 mM KH_2PO_4, 350 mM KCl, 25% (vol/vol) acetonitrile; and for both SCX solvents, pH 2.75 was achieved by adding 50% H_3PO_4. bRPLC solvent A contained 10 mM TEABC; bRPLC solvent B contained 10 mM TEABC, 90% (vol/vol) acetonitrile. SAFE-SRM MS solvent A was water with 0.1% (vol/vol) formic acid; SAFE-SRM solvent B was acetonitrile with 0.1% (vol/vol) formic acid.

Pooled Plasma Samples for iTRAQ-Based Discovery Studies. Fifty normal individuals, 13 patients with pancreatic cancer, 18 with colorectal cancer, and 18 with ovarian cancer were chosen for initial analysis. One hundred microliters of plasma from each individual in one of these four groups of patients was pooled before processing through phase 1 of the study. Phase 1 of this study used these pools rather than peptides from individual patients and are referred to as “pooled peptides.”

Plasma Depletion. Abundant proteins [albumin, IgG, α_1-antitrypsin, IgA, IgM, transferrin, haptoglobin, α_2-macroglobulin, fibrinogen, complement C3, α_1-acid glycoprotein (orosomucoid), HDL (apolipoproteins A-I and A-II), and LDL (mainly apolipoprotein B)] in the plasma were depleted using a Seppro IgY14 LC10 column system. Plasma samples were diluted 5× in IgY dilution buffer, filtered (0.22 μm), and then injected into IgY LC10 columns attached to an Agilent 1200 HPLC system consisting of a binary pump, external sample injector, UV detector, and a fraction collector. The nonretained fraction was collected.

Plasma Proteome Sample Preparation. The depleted plasma proteins were denatured in 9 M urea, reduced using 5 mM TCEP at 60 °C for 15 min, and cysteine residues were alkylated with 5 mM MMTS for 15 min at room temperature in dark. The alkylated protein solution was filtered to desalt using the Amicon Ultra-15 Centrifugal Filter Unit with Ultracel-10 membrane (Millipore) and washed with 9 M urea for two times, and the desalted plasma protein was reconstituted with 4 mL of 40 mM TEABC. The samples were then digested for 3 h with LysC protease followed by an overnight digestion using sequencing-grade trypsin at 37 °C. Additional sequencing-grade trypsin was added 3 h before digestion ended, and the digestion system was incubated at 50 °C for the last 30 min before adding 1% TFA to stop the reaction. C18-mediated cleaning of the digest was performed as described (1). For samples not used in iTRAQ experiments, that is, those from individual donors rather than pooled plasma samples, 50 mM iodoacetamide (Sigma-Aldrich) rather than MMTS was used for alkylation.

N-Glycosylated Protein Enrichment and Isolation from Human Plasma Samples. One hundred microliters of pooled human plasma samples was denatured in 9 M urea and processed through reduction, alkylation, and filtration to remove salt, and then subjected to lyophilization. Lyophilized proteins were reconstituted with 5% acetonitrile with 0.1% TFA. The 10 mM sodium periodate was applied to the protein solution followed by incubation at 4 °C for 1 h in the dark. Another C8 cartridge cleaning was performed to purify the oxidized proteins. Lyophilized proteins were reconstituted with 1 mL of hydrazide resin coupling buffer (0.1 M sodium phosphate buffer, pH 7.0), and 250 μL of hydrazide resin, purchased from Bio-Rad, was added to the solution to conjugate the glycoproteome by incubation at room temperature for 5 h. The resin was then washed twice with 4 mL of 1.5 M NaCl followed by 4 mL of water, twice with 4 mL of 100 mM TEABC buffer, and finally with 4 mL of 50 mM sodium phosphate (pH 7.5). Twenty-five microliters of PNGase F was added to the resin followed by incubation at 37 °C for 4 h with agitation. The resin was then centrifuged at 8,000 × *g* for 5 min, and the supernatant was collected. The resin pellet was washed twice with 500 μL of 40 mM ammonium bicarbonate and subjected to centrifugation as above. The supernatants from these centrifugations were combined, lyophilized, and reconstituted with 40 mM ammonium bicarbonate, and subject to trypsin digestion and C18 cleaning, after which they were used for iTRAQ labeling. A total of 657 glycosylated proteins was identified and quantified (Dataset S3). There were 29 proteins identified from the N-glycosylated protein enrichment experiments that were carried forward to the validation phases of this study.

iTRAQ Labeling, SCX Cleaning, and bRPLC Fractionation. Peptides from the four pools were reconstituted in 15 μL of H_2O and 20 μL of dissolution buffer (provided with the iTRAQ labeling kit) and incubated with one of the four iTRAQ reagents diluted in 70 μL of ethanol at room temperature. The peptides from each of the four pools were labeled with iTRAQ reagents containing 114, 115, 116, or 117 reporter ions, respectively. After incubation at room temperature for 2 h, 50 μL of water was added. After another incubation for 10 min at room temperature, 100 μL of water was added. After incubation at room temperature for another 10 min, 40 μL of 40 mM ammonium bicarbonate was then added, and the reactions were incubated at 4 °C overnight. The samples were vacuum dried to 50 μL, combined, and diluted to 4 mL in 10 mM potassium phosphate buffer (pH 2.7) containing 25% acetonitrile (SCX solvent A). The pH of the sample was adjusted to 2.7 using 100 mM phosphoric acid. iTRAQ-labeled peptides were then purified using SCX chromatography with a polysulfoethyl A column (PolyLC) (300 Å, 5 μm, 100 × 2.1 mm) (2) on an Agilent 1200 HPLC system. Fractionation was carried out for a period of 45 min using a linear gradient of increasing salt concentration from 0 to 350 mM KCl in SCX solvent B. Peptide fractionations were then vacuum dried and reconstituted with 4 mL of bRPLC solvent A and subject to bRPLC fractionation with an XBridge C18 column (Waters). A total of 96 fractions from the bRPLC was deposited in a 96-well plate.

Liquid Chromatography–MS/MS and Plasma Quantitative Proteomics Data Analysis. Nanoflow electrospray ionization liquid chromatography (LC)–MS/MS analysis of the iTRAQ-labeled bRPLC-separated samples was performed with an LTQ Orbitrap Velos (Thermo Fisher Scientific) mass spectrometer interfaced with reversed-phase system controlled by Eksigent nano-LC and Agilent 1100 microwell plate autosampler. The bRPLC fractions were sequentially processed through a 75 μm × 2 cm, Magic C18AQ column (5 μm, 100 Å; Michrom Bioresources) and then separated on an analytical column (75 μm × 10 cm, Magic C18AQ, 5 μm, 100 Å; Michrom Bioresources) with a nanoflow solvent delivery. The mobile phase flow rate was 200 nL/min, composed of 3% acetonitrile/0.1% formic acid (solvent A) and 90% acetonitrile/0.1% formic acid (solvent B), and the 110-min LC-MS/MS method consisted of a 10-min column equilibration procedure, 10-min sample-loading procedure, and the following gradient profile: (min:B%) 0:0; 2:6; 72:40%; 78:90%; 84:90%; 87:50%; 90:50% (last three steps at 500 nL/min flow rate). The MS and MS/MS data were acquired in positive-ion mode at a spray voltage of 2.5 kV and at a resolution of 60,000 at *m/z* 400. For every duty cycle, the 10 most abundant peptide precursors were selected for MS/MS analysis in the LTQ Orbitrap Velos (normalized collision energy, 40%). A detailed flowchart of iTRAQ-based quantitative proteomics is shown (Fig. S1*A*).

Quantitative Proteomics Analysis. The MS data from the iTRAQ experiments were analyzed with Proteome Discoverer (version 2.1; Thermo-Fisher). MS/MS spectral data were processed using the extract feature under the MASCOT and Sequest HT search components of the program. For both components, the same search parameters were selected, and these included iTRAQ labels at tyrosine, oxidations of methionine, and deamidation at N/Q as variable modifications. iTRAQ labels at N terminus, and lysine, methylthio label at cysteine were used as fixed modifications. The MS data were searched against NCBI RefSeq 72 human protein database containing 55,692 sequences. Proteome Discoverer calculates the percentage of false identifications using a separate decoy database (reverse database) that contains the reversed sequences of the protein entries. The Proteome Discoverer counts the number of matches from both searches and calculates the false-discovery rate (FDR) by counting only the top match per spectrum, assuming that only one peptide can be the correct match. The score thresholds were adjusted to obtain 1% and 5% reverse hits compared with forward hits, resulting in an overall FDR of 5%. Precursor and reporter ion window tolerance were fixed at 20 ppm and 0.05 Da, respectively. The criteria specified for generation of peak lists included signal-to-noise ratios of 1.5 and inclusions of precursor mass ranges of 600–8,000 Da. The two validated SAFE-SRM target peptides from PPIA protein were initially identified unambiguously using a 1% FDR cutoff, as shown in Fig. S5.

Statistical Analysis of Peptide Quantification Using the *limma* Package in R/Bioconductor. Peptide expression ratios of the pooled samples were calculated based on the median value of peptide ion intensities of iTRAQ labelings 117 (pancreatic cancer pool), 116 (colorectal cancer pool), or 115 (ovarian cancer pool) relative to that of 114 (normal individual pool). Sample preparation was performed in duplicate (two biological replicates). MS analysis was performed once on the first replicate (generating dataset 1) and twice on the second replicate, generating datasets 2 and 3, which were therefore technical replicates. A matrix was generated to store the raw peptide abundance data, where row names contained all unique sequences of the peptides. Columns 1 through 4 stored the intensities of 114, 115, 116, and 117 labeling intensities from dataset 1. Columns 5 through 8 and columns 9 through 12 stored the analogous labeling intensities from datasets 2 and 3, respectively. “NA” was used to indicate that a peptide was not detected in a particular dataset with a particular label (Dataset S2).

MA plots were generated to compare the potential bias between different datasets. Because no significant bias was observed in these MA plots (Fig. S6), median normalization was chosen for subsequent analysis (Fig. S7). For this analysis, we borrowed the concepts developed for the analysis of microarray data and used R packages from the Bioconductor project to analyze peptide fold changes (3). In particular, we used the modified *t* test from limma (linear models for microarray data) to judge the statistical significance of the changes observed (3).

Let y_i and x_i denote the abundances of the *i*th protein in cancer plasma proteome and normal plasma proteome, respectively, so that

$$y_i \sim Norm\left(\mu_{y_i}, \sigma_{y_i}\right),$$

and

$$x_i \sim Norm\left(\mu_{x_i}, \sigma_{x_i}\right),$$

where μ and σ denote the mean and variance of a peptide abundance in the three datasets. To avoid identifying peptide biomarkers (highly up-regulated in cancer plasma proteome compared with normal) that have significant variance between replicates, we adopted a *t* test where

$$t\ \text{statistic} = \frac{\bar{y}-\bar{x}}{\sqrt{\frac{\widehat{\delta_x}+\widehat{\delta_y}}{n}}}.$$

The *t* test was modified by an empirical Bayes method. Instead of testing each peptide in isolation from all others, the empirical Bayes modified *t* test borrows strength from all other peptides, thus improving the error estimate of each individual peptide. The eBayes modified *t* test from *limma* R package was used to perform statistical analysis for the difference of peptide abundances between samples. In total, 208 peptides from 87 different proteins were identified as candidate cancer biomarkers and were carried on to the validation phase of this study.

Candidate Biomarkers Identified by Quantitative Plasma Proteomics. Proteomics database searches (using PRIDE, https://www.ebi.ac.uk/pride/archive/, and Peptide Atlas, www.peptideatlas.org/) were conducted for the 87 proteins, and their 253 most readily detectable peptides (other than the 208 noted above) were added to the candidate peptide list. Another 180 peptides observed repeatedly from the three discovery datasets but that did not pass the eBayes-modified *t* test were also added. In total, 641 candidate peptides were subject to further validation (Dataset S4).

Development of SAFE-SRM Assays Using Synthetic Peptides. The 641 candidate peptides were synthesized and used as standards to establish the SAFE-SRM method using a three-step optimization approach:

i) Optimization of collision energy was performed for each pair of precursor ion (usually positively charged proteotypic peptide) and product ion (peptide fragments generated from collision-induced dissociation). For each precursor ion, two steps above and two steps below (step size, 4 eV) the theoretical optimum value of collision energies were applied to fragment each precursor ion. For each peptide, five to eight fragmented ions showing the strongest intensities were selected as the detection targets. Mass-to-charge ratio (*m/z*) of the peptide, optimized collision energy values, and the *m/z* of the peptide fragmented ions were thus established for each peptide. A set of such values is typically referred to as SRM transitions for a target peptide. In total, 4,384 SRM transitions were optimized in this way to target the

641 peptides (on average, approximately seven transitions per peptide).

ii) Optimization of bRPLC fractionation. The 641 synthetic peptides were spiked into the peptides derived from the pooled normal plasma sample used in phase 1 of the study prepared as described above, and three independent HPLC fractionations were carried out. As noted above, the 96 fractions from the bRPLC fractionation were combined into "fraction groups," with each group containing three sequential fractions. The 4,384 transitions were assessed in each bRPLC fraction group, with fixed dwell time for each transition (5 ms). The bRPLC fraction group containing the highest amount of each peptide was determined, thereby defining a fraction group ID for each peptide. The standard intensity (SI) (the intensity measured by mass spectrometer for 10 fmol of the peptide) for each peptide was also recorded.

iii) SRM method assembly. A unique SRM method was created for each fraction group by compiling all of the transitions from the peptides with the same fraction group ID. The same SRM transitions were evaluated in the fraction groups eluting before and after the main fraction group. Thus, each fraction group was assessed with three different sets of SRM transitions. The dwell time for each transition was modified to be inversely proportional to the SI of the peptide, ranging from 3 to 20 ms.

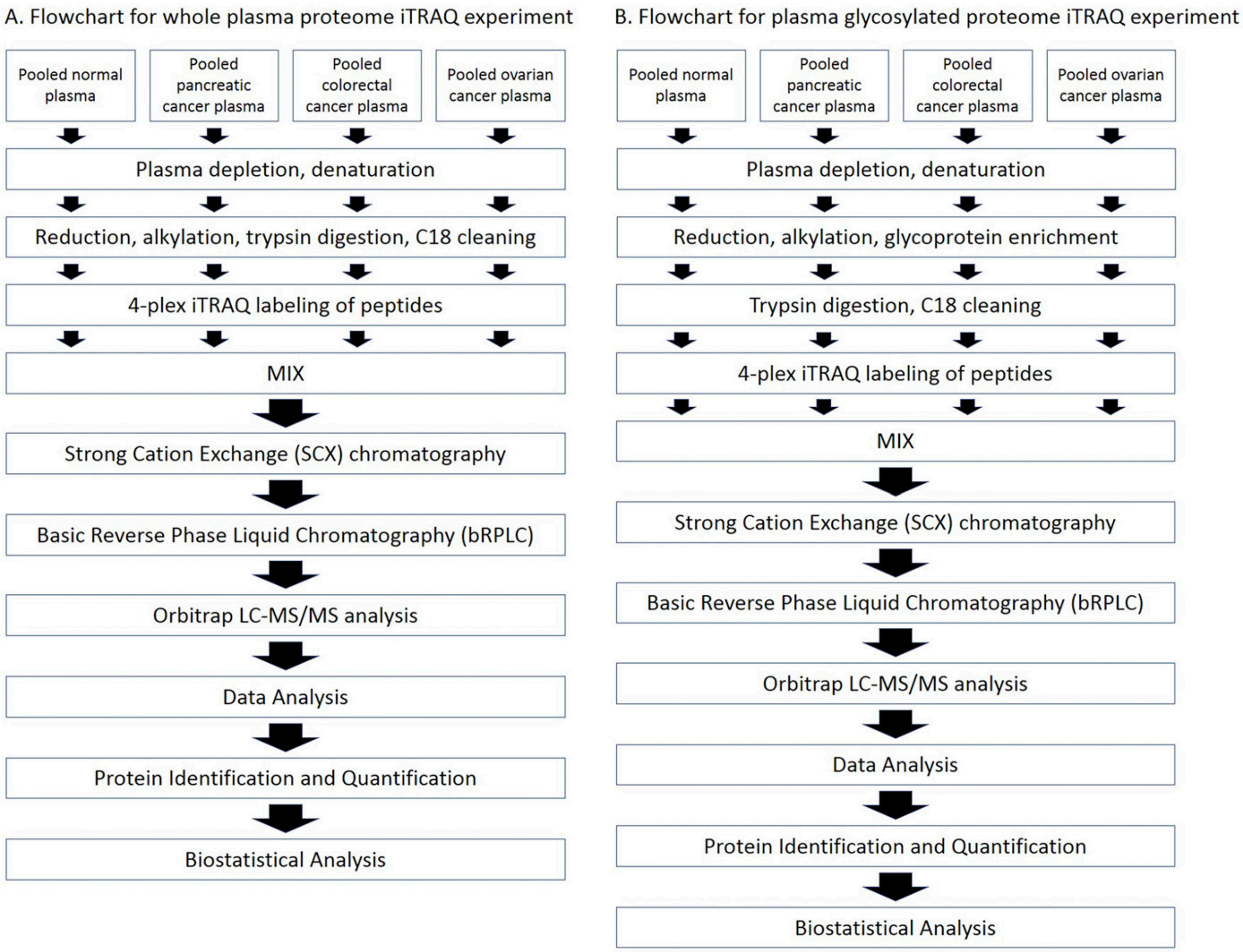
A list of the SRM transitions and fraction group IDs for all of the peptides are shown in Dataset S5. All transition parameters were manually examined and curated to exclude ions with excessive noise due to coelution with nonspecific analytes in human plasma samples. A set of 1,990 transitions was reproducibly detectable in a pool of all advanced cancer plasma samples used in phase 1, corresponding to 318 peptides (Dataset S5).

SAFE-SRM can be performed with synthetic light peptides. Shi et al. (4) reported a method called PRISM-SRM, built upon heavy-isotope–labeled peptides. The high cost of heavy peptides complicates its application to early-stage biomarker development where hundreds or thousands of biomarkers need to be validated. Heavy-isotope–labeled peptides may also lead to ion suppression, thereby compromising sensitivity.

Performance Evaluation of SAFE-SRM. Six heavy-isotope–labeled peptides (peptide 1: IQLVEEELDR*; peptide 2: VILHLK*; peptide 3: IILLFDAHK*; peptide 4: TLAESALQLLYTAK*; peptide 5: LLGHLVK*; peptide 6: GLVGEIIK*, where * indicates C13 and N15 heavy-isotope–labeled amino acids) were mixed at 1 fmol each, and the mixture was analyzed by a standard SRM method. Equal amounts (1 fmol each) of the six heavy-isotope–labeled peptides were spiked into proteolytically digested plasma peptide sample, followed by detection through a standard SRM approach (without bRPLC fractionation), a bRPLC-SRM approach, or a SAFE-SRM approach. The peptide abundance was calculated by the AUC of the peptide's SRM signal detected in each approach.

Agilent 6490 Mass Spectrometer Tuning. SAFE-SRM assays for each plasma sample were conducted only after confirmation of the instrument's performance with the manufacturer's tuning mixes (Autotune and Checktune) as well as a tuning mixture we prepared. Our tuning mixture was composed of 20 peptides representing a wide range of mass (*m*/*z* range, 200–1,400) and hydrophobicity (Table S2).

1. Wang Q, et al. (2011) Mutant proteins as cancer-specific biomarkers. *Proc Natl Acad Sci USA* 108:2444–2449.
2. Chaerkady R, et al. (2010) Comparative proteomics of human embryonic stem cells and embryonal carcinoma cells. *Proteomics* 10:1359–1373.
3. Gentleman RC, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.
4. Shi T, et al. (2012) Antibody-free, targeted mass-spectrometric approach for quantification of proteins at low picogram per milliliter levels in human plasma/serum. *Proc Natl Acad Sci USA* 109:15395–15400.

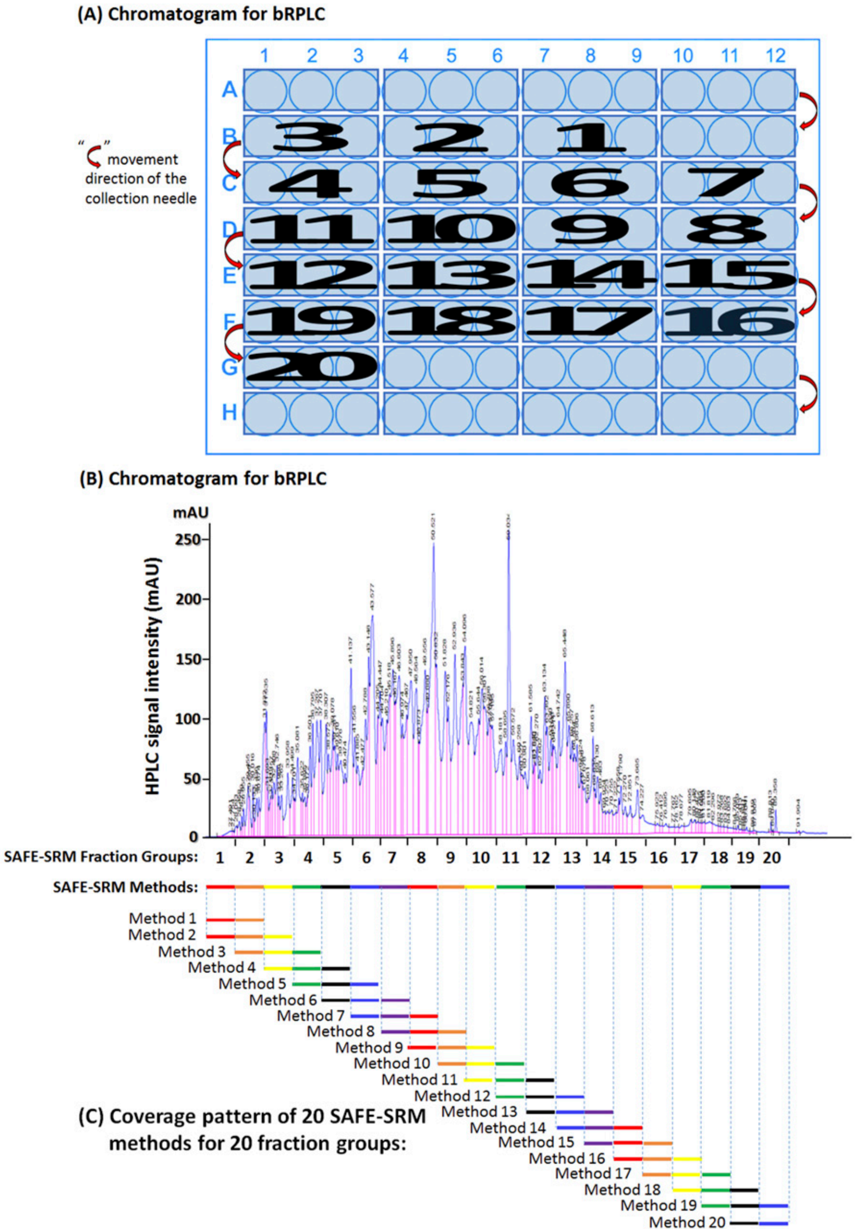


A. Flowchart for whole plasma proteome iTRAQ experiment

B. Flowchart for plasma glycosylated proteome iTRAQ experiment

Pooled normal plasma
Pooled pancreatic cancer plasma
Pooled colorectal cancer plasma
Pooled ovarian cancer plasma
Plasma depletion, denaturation
Reduction, alkylation, trypsin digestion, C18 cleaning
4-plex iTRAQ labeling of peptides
MIX
Strong Cation Exchange (SCX) chromatography
Basic Reverse Phase Liquid Chromatography (bRPLC)
Orbitrap LC-MS/MS analysis
Data Analysis
Protein Identification and Quantification
Biostatistical Analysis
Pooled normal plasma
Pooled pancreatic cancer plasma
Pooled colorectal cancer plasma
Pooled ovarian cancer plasma
Plasma depletion, denaturation
Reduction, alkylation, glycoprotein enrichment
Trypsin digestion, C18 cleaning
4-plex iTRAQ labeling of peptides
MIX
Strong Cation Exchange (SCX) chromatography
Basic Reverse Phase Liquid Chromatography (bRPLC)
Orbitrap LC-MS/MS analysis
Data Analysis
Protein Identification and Quantification
Biostatistical Analysis

Fig. S1. Detailed technical workflow for iTRAQ-labeling–based quantitative proteomics studies with total plasma proteome (*A*) and plasma glycoproteome (*B*).


(A) Chromatogram for bRPLC

(B) Chromatogram for bRPLC

(C) Coverage pattern of 20 SAFE-SRM methods for 20 fraction groups:

Fig. S2. SAFE-SRM scheme. (*A*) bRPLC fractionation was performed to separate peptides from a complicated biological sample into 96 fractions according to their hydrophobicity at high pH. The SAFE-SRM fraction groups are overlaid on the wells. (*B*) A chromatogram showing the combined signal intensities of all peptides in each of the 20 SAFE-SRM fraction groups used in the final SAFE-SRM method. (*C*) SAFE-SRM method transition coverages. For each fraction group i, the specific SAFE-SRM method i is composed of the transitions detecting peptides within that fraction group and two adjacent groups, group $i - 1$ and group $i + 1$, where $i \in [1, 20]$.

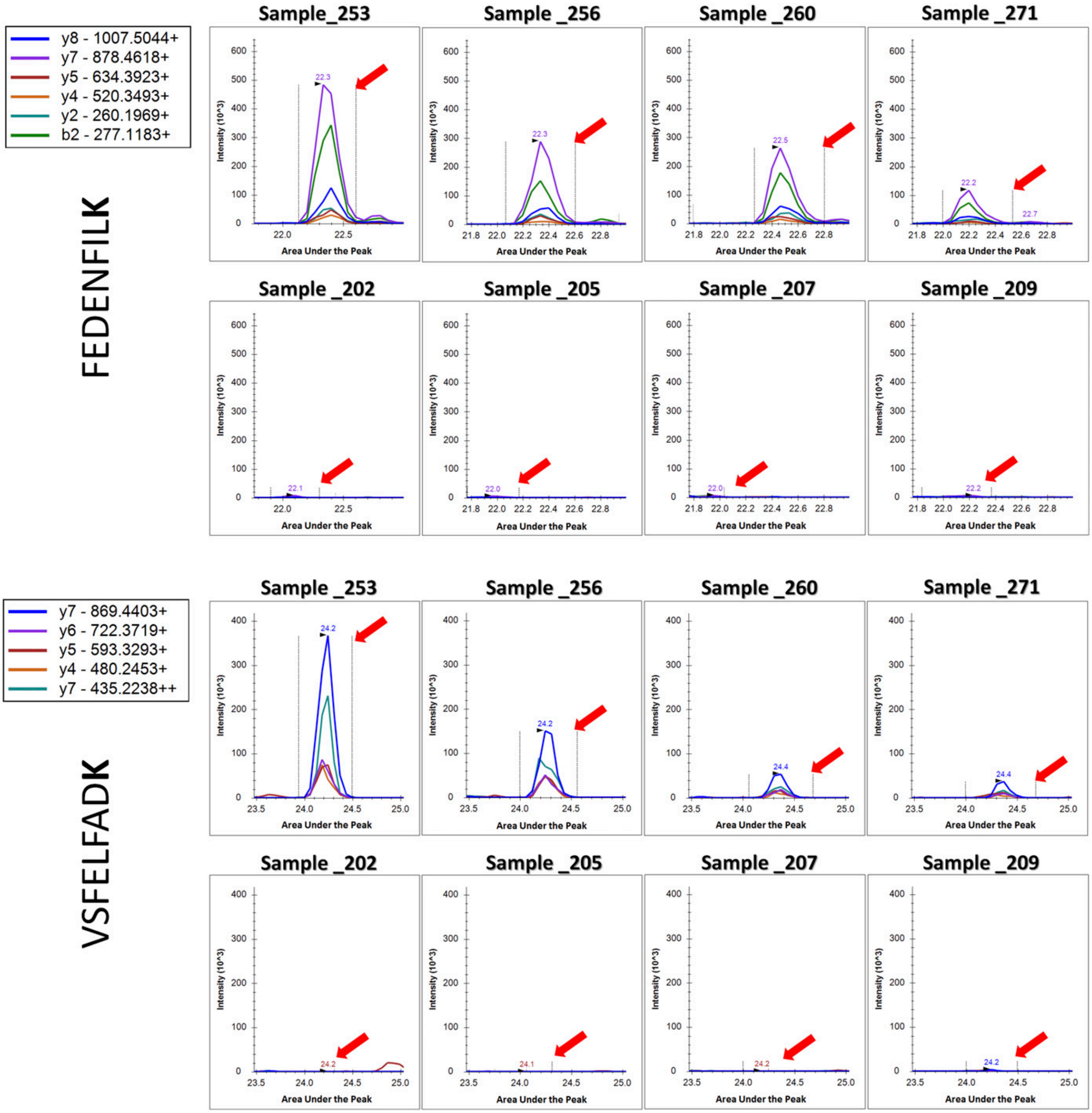


Fig. S3. SAFE-SRM profiles for three ovarian cancer biomarker peptides in eight plasma samples. Four ovarian cancer plasma samples (253, 256, 260, and 271) and four normal healthy plasma samples (202, 205, 207, and 209) were analyzed by SAFE-SRM. The areas under the peak are shown for each sample.

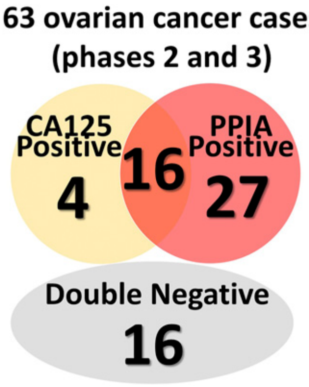


Fig. S4. Comparison of ovarian cancer diagnostic performance using SAFE-SRM–based PPIA assay and ELISA-based CA125 assay. The Venn diagram shows the number of cases identified in a cohort of 63 ovarian cancer patients.

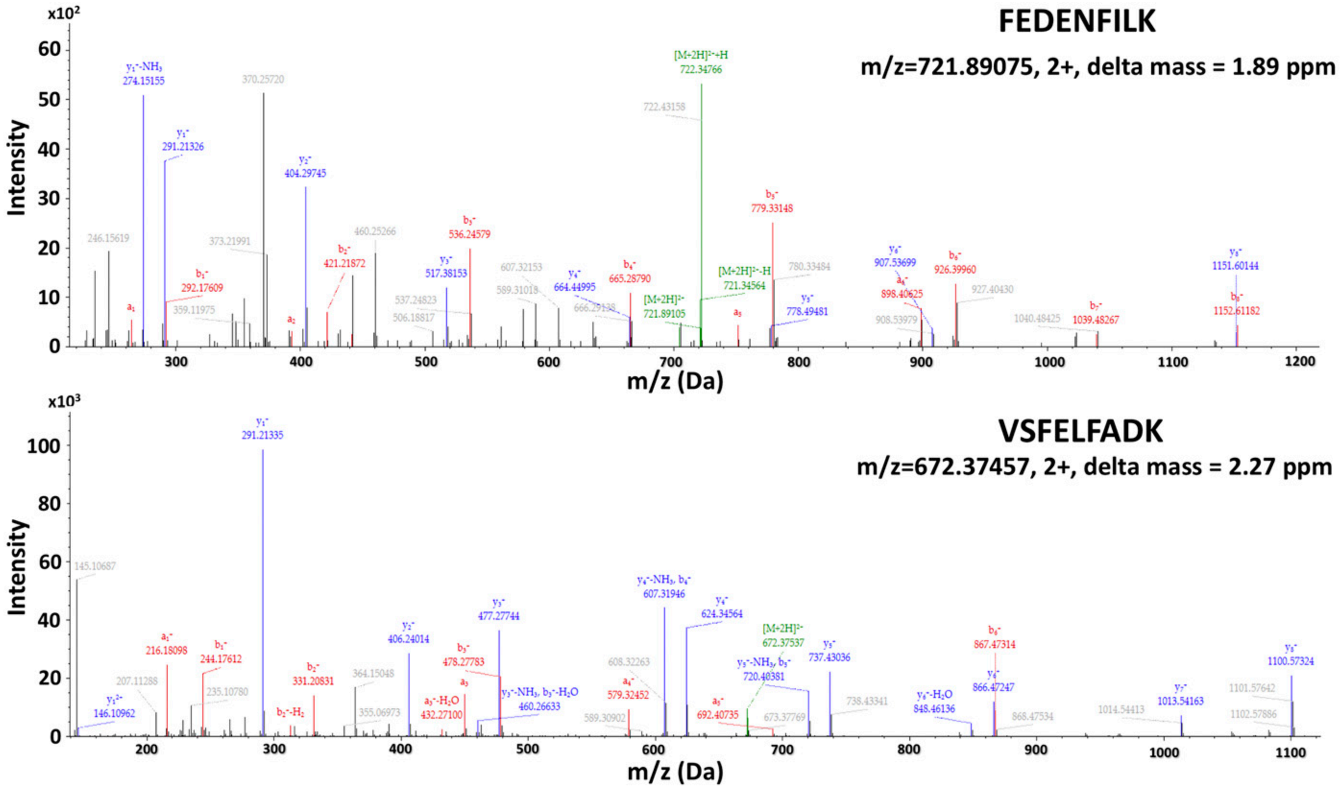


Fig. S5. MS spectra of SAFE-SRM target peptides from PPIA.

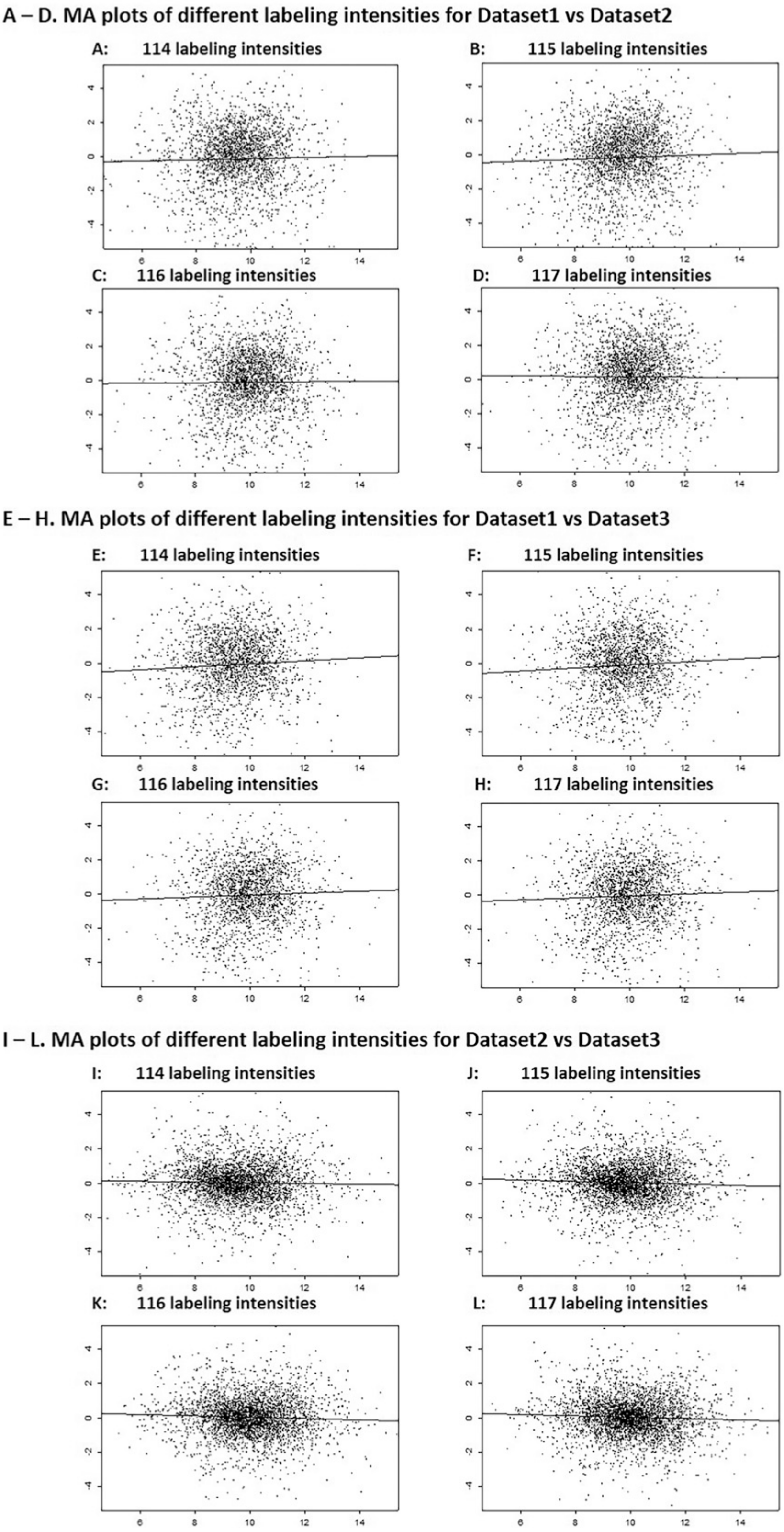


Fig. S6. MA plots for whole-plasma iTRAQ datasets. Nonnormalized peptide intensities from each of the three experiments were compared under each specific labeling (114, 115, 116, and 117) and corresponding MA plots were generated using the log-transformed raw intensities, with A ranges fixed to 6–14, and M ranges fixed to −4 to 4. There is no clear evidence of bias associated with any of the datasets. The technical variance (*I*–*L*) is significantly smaller than the biological variance (*A*–*D* or *E*–*H*).

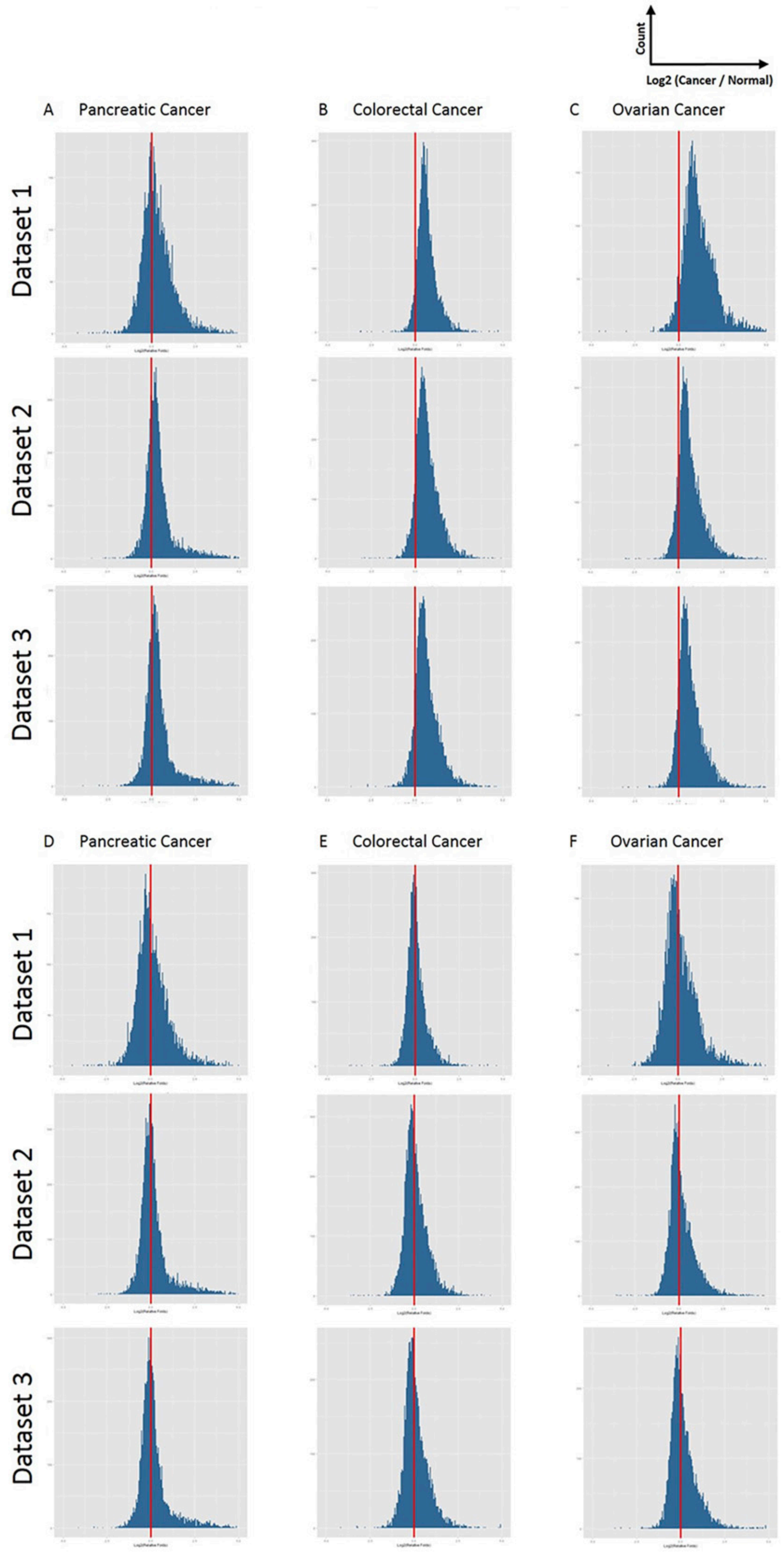


Fig. S7. Nonnormalized and median normalized histograms for cancer vs. normal in three datasets. Protein ratios of cancers/normal were plotted using $\log_2$ scale for dataset 1 (*A*–*C*, *Upper*), dataset 2 (*A*–*C*, *Middle*), and dataset 3 (*A*–*C*, *Lower*). After median normalization, the same protein ratios of cancers/normal were plotted using $\log_2$ scale for dataset 1 (*D*–*F*, *Upper*), dataset 2 (*D*–*F*, *Middle*), and dataset 3 (*D*–*F*, *Lower*). The $\log_2$ (relative ratio) = 0 lines are indicated in each plot (red line). Biased data were observed for colorectal cancer (*B*) and ovarian cancer (*C*). The bias for pancreatic cancer (*A*) is not obvious.

Table S1. Study design and cases involved in the study

Study phases	Study aims	Normal healthy individual	Pancreatic cancer	Colorectal cancer	Ovarian cancer	Total
Phase 1	Identification of candidate biomarkers from cancer patients	50	13	18	18	99
Phase 2	Development of SAFE-SRM and testing of candidate peptides by SAFE-SRM	32	14	20	28	94
Phase 3	Validation by SAFE-SRM	14	24	0	35	73
					Total	266

Shown are the study design and the cohorts evaluated.

Table S2. Standard peptides in tuning mixture (10 fmol each)

Peptide sequence	SAFE-SRM fraction group ID
DEIESVK	3
VGSAKPGLQK	4
ETIVLK	5
IQLVEEELDR	6
SIVNYKPK	7
DLQFVEVTDVK	7
TLLGDGPVVTDPK	8
GLVGEIIK	9
HFTILDAPGHK	10
LVDKFLEDVK	11
KIPVVFR	11
VILHLK	12
FPVIQHFK	12
LLGHLVK	13
FFLSHPAYR	14
LFAGLVHVK	14
TLAESALQLLYTAK	15
IILLFDAHK	15
VLDFEHFLPMLQTVAK	17
LLGNVLVCVLAHHFGK	18

Twenty peptide sequences and their SAFE-SRM fraction group IDs are shown.

Dataset S1. Summary characteristics of cases involved in this study

Dataset S1

Dataset S2. The 10,789 identified peptides and their ratios in iTRAQ experiments

Dataset S2

Dataset S3. Plasma proteomics iTRAQ datasets at protein level

Dataset S3

Dataset S4. The 641 SAFE-SRM target peptides and 318 detectable peptides in plasma

Dataset S4

Dataset S5. The 4,384 SRM optimized transitions targeting 641 peptides, and 1,990 detectable transitions in plasma

Dataset S5

Dataset S6. The 318 SAFE-SRM abundance scores for each of the 98 samples used for biomarker identification

Dataset S6

Dataset S7. Normalized SAFE-SRM abundance scores for individual cases in phases 2 and 3 of the study

Dataset S7

Values exceeding the thresholds used for scoring are in bold font.