DPHL v2: An updated and comprehensive DIA pan-human assay library for quantifying more than 14,000 proteins

Zhangzhi Xue[1,2,3], Tiansheng Zhu[1,2,3,15], Fangfei Zhang[1,2,3], Cheng Zhang[1,2,3], Nan Xiang[4], Liujia Qian[1,2,3], Xiao Yi[4], Yaoting Sun[1,2,3], Wei Liu[4], Xue Cai[1,2,3], Linyan Wang[5], Xizhe Dai[5], Liang Yue[1,2,3], Lu Li[1,2,3], Thang V. Pham[6], Sander R. Piersma[6], Qi Xiao[1,2,3], Meng Luo[7], Cong Lu[8], Jiang Zhu[8], Yongfu Zhao[9], Guangzhi Wang[9], Junhong Xiao[9], Tong Liu[10], Zhiyu Liu[11], Yi He[11], Qijun Wu[12], Tingting Gong[12], Jianqin Zhu[13,14], Zhiguo Zheng[13,14], Juan Ye[5], Yan Li[7], Connie R. Jimenez[6] *, Jun A[1,2,3] *, Tiannan Guo[1,2,3] *

[1] iMarker lab, Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, Zhejiang Province, China
[2] Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, Zhejiang Province, China
[3] Research Center for Industries of the Future, Westlake University, 600 Dunyu Road, Hangzhou, Zhejiang, 310030, China
[4] Westlake Omics (Hangzhou) Biotechnology Co., Ltd., Hangzhou, China
[5] Department of Ophthalmology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China
[6] OncoProteomics Laboratory, Department of Medical Oncology, VU University Medical Center, VU University, Amsterdam 1011, The Netherlands
[7] Songjiang research Institute and Songjiang Hospital, Department of Anatomy and Physiology, College of Basic Medical Science,
Shanghai Jiao Tong University School of Medicine, Shanghai 201600, China
[8] Center for Stem Cell Research and Application, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology Wuhan, Hubei, P. R. China
[9] Department of General Surgery, The Second Hospital of Dalian Medical University, Dalian, China
[10] Harbin Medical University Cancer Hospital, Harbin 150081, Heilongjiang Province, China

[11] Department of Urology, The Second Hospital of Dalian Medical University, No.467 Zhongshan Road, Dalian, Liaoning Province, China

[12] Department of Clinical Epidemiology, Shengjing Hospital of China Medical University, Shenyang 110000, Liaoning Province, China

[13] The Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), Hangzhou, Zhejiang, China

[14] Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, Zhejiang, China

[15] College of Mathematics and Computer Science, Zhejiang A & F University, Hangzhou, China.

Correspondence: Connie Jimenez (c.jimenez@amsterdamumc.nl), Jun A (ajun@westlake.edu.cn), Tiannan Guo (guotiannan@westlake.edu.cn)

Summary

A comprehensive pan-human spectral library is critical for biomarker discovery using mass spectrometry (MS)-based proteomics. DPHL v1, a previous pan-human library built from 1096 data-dependent acquisition (DDA) MS data of 16 human tissue types, allows quantifying 10,943 proteins. However, a major limitation of DPHL v1 is the lack of semi-tryptic peptides and protein isoforms, which are abundant in clinical specimens. Here, we generated DPHL v2 from 1608 DDA-MS data acquired using Orbitrap mass spectrometers. The data included 586 DDA-MS newly acquired from 17 tissue types, while 1022 files were derived from DPHL v1. DPHL v2 thus comprises data from 24 sample types, including several cancer types (lung, breast, kidney, and prostate cancer, among others). We generated four variants of DPHL v2 to include semi-tryptic peptides and protein isoforms. DPHL v2 was then applied to a publicly available colorectal cancer dataset with 286 DIA-MS files. The numbers of identified and significantly dysregulated proteins increased by at least 21.7% and 14.2%, respectively, compared with DPHL v1. Our findings show that the increased human proteome coverage of DPHL v2 provides larger pools of potential protein biomarkers.

Keywords

Targeted proteomics; Spectral library; Data-independent acquisition; Mass spectrometry; Cancer; Colorectal cancer

Introduction

Mass spectrometry (MS)-based quantitative proteomics is widely used for protein biomarker discovery[1-3]. The subsequent biomarker validation is often performed with targeted proteomics methods, such as selected reaction monitoring (SRM)[4] and parallel reaction monitoring (PRM)[5]. Recently, biomarker discovery and validation have been increasingly performed with targeted analysis of data-independent acquisition (DIA) MS data[6], an emerging strategy for high-throughput proteomics analyses with a high level of reproducibility[7]. A spectral library containing experimental peptide precursor information is crucial for SRM- and PRM-based protein biomarker validation, as well as DIA-based biomarker discovery[7]. In recent years, spectral libraries have been established for several organisms, such as human[8, 9], mouse[10], zebrafish[11], *Arabidopsis thaliana*[12], and *Escherichia coli*[13]. To support the identification of new protein biomarkers, the comprehensiveness of a spectral library is crucial.

The Human Proteome Project (HPP)[14] launched by Human Proteome Organization (HUPO) has reported the community-based ten-year achievement of a high-stringency proteome blueprint of 17,874 Protein Evidence 1 (PE1) proteins in 2020, covering 90.4% of the human proteome[15]. A pan-human spectral library (PHL), containing 149,130 peptide precursors and 10,322 proteins, was developed to analyze Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH-MS) data acquired on SCIEX TripleTOF Systems[8]. Another DIA pan-human library (DPHL v1) for Orbitrap data comprises 289,237 peptide precursors and 10,943 proteins[9].

However, the proteins in these two libraries are proteotypic; protein isoforms are not included. The isoforms of each protein family may result from post-translational modifications, splice variants, proteolytic products, genetic variations, or somatic recombination occurring during protein evolution[16], and participate in different biological processes[17]. Therefore, a specific protein isoform could be a valuable biomarker. A spectral library with significant coverage of the human proteome and its protein isoforms is thus needed. Additionally, previous studies demonstrated that only ~10-15% of all the tryptic peptides from a protein sample can be identified when about 50% of the protein identifications are based on a single tryptic peptide due to the intrinsic chemical properties of tryptic peptides[18-20]. Therefore, identifying more peptides (e.g., non-tryptic peptides), preferably at low computational costs, would increase the confidence in the proteins identified via tryptic peptides and increase the overall number of identifications.

Here, we present a large DIA spectral library (DPHL v2), generated from 24 different sample types and available in four variants. DPHL v2 includes more peptide precursors, peptides, and proteins than DPHL v1. It also provides higher coverage ratios, particularly for brain-, esophagus-, and ovary-specific or -enriched proteins, as well as FDA-approved drug targets. Two variants of DPHL v2 generated better identifications of the hallmark gene sets than DPHL v1. Finally, using a publicly available colorectal cancer (CRC) cohort, DPHL v2 provided larger numbers of protein and differentially expressed protein identifications than DPHL v1 and library-free method.

Results and Discussion

Data sources for generating DPHL v2

A total of 1608 raw MS data files were collected to build our spectral library. Among these, 586 files were newly generated from various samples, including tissue biopsies of prostate cancer (PCa), hepatocellular carcinoma (HCC), triple-negative breast cancer (TNBC), lung adenocarcinoma (LUAD), esophageal carcinoma, thyroid diseases, eyelid tumors, glioblastoma multiforme (GBM), healthy brain tissues, oral squamous cell carcinoma (OSCC), thymic diseases, ovarian cancer (OV), and cervix cancer. Additionally, blood plasma samples from acute myelocytic leukemia (AML), blood diseases, T-lineage acute lymphoblastic leukemia (T-ALL), and normal plasma exosome were included. Human chronic myelogenous leukemia cell line K562 was also included. Finally, the remaining 1022 files were derived from the DPHL v1 study by Zhu *et al*[9]. The sample types and number of patients contributing to DPHL v2 are summarized in Figure 1A and Table S1.

Four variants of the pan-human spectral libraries

All the 1608 raw files were centroided and converted into mzXML as previously described[9]. These files were then combined to build our new spectral library. Two different annotation files (i.e., reviewed and isoform-reviewed fasta files) were used to search the mzXML spectra against two digestion modes (i.e., full-specific and semi-specific) using MS-Fragger (version 3.0)[21]. The reviewed fasta file was obtained from the UniProt database[22] (accessed on 17 Jul. 2020); it included 20,361 reviewed human proteins and was used as the reference. The isoform-reviewed annotation file was also downloaded from UniProt (accessed on 5 Aug. 2020) and comprised 42,347 proteins, including 22,201 human isoforms. Philosopher[23] (version 3.2.9) was used for library searching based on the spectra matches with a maximum of two missed cleavages and a false discovery rate < 0.01 for spectra, peptides, and proteins. By differently combining the two annotation files and the two digestion modes, we generated four library variants: RF (reviewed fasta sequence & full-specific digestion mode), RS (reviewed fasta sequence & semi-specific digestion mode), IF (isoform fasta sequence & full-specific digestion mode), IS (isoform fasta sequence & semi-specific digestion mode).

Next, in order to ensure the consistency of the results of different time gradients of the mass spectrum, we used EasyPQP (version 0.1.9, https://github.com/grosenberger/easypqp) to anchor the CiRT[21] peptides for retention time (RT) normalization. Quality controls (QC) were then performed using an R script with the criteria next described to remove data of low quality. First, only precursors with multiple fragments (≥ 2) and a normalized RT range from -60 to 200 were retained. Second, fragments with a library intensity < 10 or a precursor charge of +1 were removed. Finally, peptides with only one precursor were retained. However, when a peptide had two precursors, we kept the one with the highest intensity if the absolute difference of the normalized RT between the two precursors was > 5; otherwise, both precursors were kept. When a peptide has more than two precursors, the averaged normalized RTs of all precursors and their differences with respect to their mean RT were calculated. Next, peptides with an absolute difference > 5 were excluded. When all the absolute values were > 5, the median normalized RT of all the precursors and their difference from the median RT were further calculated: only the peptides with a difference < 5 were then selected. The normalized RT correlations (+2/+3 states of each peptide) after these filtering steps are shown in Figure S1. Default parameters were used for all software unless otherwise indicated. The computational pipeline is schematized in Figure 1B.

Characteristics of DPHL v2

We next evaluated DPHL v2 using DIALib-QC[24] and found that all four variants of our pan-human spectral library are of high quality (Figure S2-5). We also characterized the four libraries in terms of peptide and protein identifications. As shown in Figure 1C, the RF library includes 601,982 peptide precursors, 441,141 peptides, and 13,465 proteins; the IF library includes 604,748 peptide precursors, 443,150 peptides, and 14,375 proteins. IS, another isoform-based library, comprises 808,672 peptide precursors, 624,467 peptides, and 14,555 proteins. Finally, the RS library contains 772,401 peptide precursors, 588,984 peptides, and 13,570 proteins. We then evaluated the protein identifications of the four libraries for each of the 24 sample types. As shown in Figures S6-7, the brain had the highest number of total and unique proteins among all sample types, possibly due to the larger number of brain tissues included (n = 163).

Next, we compared our four libraries with the PHL and DPHL v1 and found that our four libraries exhibited at least a 23.0% and 30.4% increase in protein coverage compared to DPHL v1 and the PHL, respectively. Among our four libraries, the isoform-based ones (IS and IF) comprise relatively high numbers of proteins (Figure 2A). Similarly, our four libraries exhibit considerably larger numbers of peptide (Figure 2C) and precursor (Figure 2E) identifications when compared to DPHL v1 and the PHL. In particular, the semi-specific digestion libraries (IS and RS) have the most significant numbers of peptide and precursor identifications. As shown in Figure 2B, 2D, and 2F, 7262 proteins, 89,328 peptides, and 103,704 precursors are shared among these six libraries, while 1,144 proteins, 165,041 peptides, and 253,673 precursors are shared only by our four libraries. These findings indicate that DPHL v2 provides higher coverage among precursors, peptides, and proteins than DPHL v1 and the PHL.

We next compared the numbers of shared proteins and peptides between our four library variants (*i.e.*, between fasta files and digestion models) (Figure 3A, 3B). We found that protein identifications were affected mainly by the fasta file, while peptide identifications were affected by the digestion model. We also compared our four libraries with DPHL v1 in terms of the enriched/specific proteins from three tissues (brain, ovary, and esophagus; Figure 3C) obtained from the Human Protein Atlas (https://www.proteinatlas.org/, data available from v21.0.proteinatlas.org). Our results indicated that the coverages of our four libraries are superior to that of DPHL v1. Similarly, our four libraries provided higher coverage of FDA-approved drug targets than DPHL v1 (Figure 3C). In addition, the hallmark gene sets from the MSigDB v7.4 database (http://www.broad.mit.edu/gsea/msigdb/, accessed on 22 Nov. 2021)[25, 26] were analyzed using these five libraries. We found that RF and RS cover more than

44% of the genes with well-defined biological states or processes, and both provide better coverages than DPHL v1 (Figure 3C). However, fewer coverages were found in the isoform-based libraries. One possible reason is that most genes from the hallmark gene sets are reviewed.

Applicability of DPHL v2 for DIA targeted data analysis

To assess the applicability of DPHL v2, we used our four libraries, DPHL v1, or a library-free method to analyze a CRC cohort, including 201 CRC cases, 40 benign samples, and 45 biological/technical replicates[27]. The missing values generated by our four libraries or DPHL v1 were comparable. On the other hand, the library-free method generated fewer missing values (Figure 4A). As shown in Figure 4B, the number of proteins identified with any variant of DPHL v2 was significantly higher than with DPHL v1 or the library-free method. A total of 978 proteins were identified by all six methods, while 166 were shared by our four libraries only (Figure 4C).

In order to demonstrate the applicability of the library, we performed differential expression analyses of the CRC data generated using the six methods described above. Differential expressions were considered significant if their adjusted p-values were < 0.01 and their $\log_2$ (fold-change) absolute values were > 1. We obtained 1997 (RF), 1984 (RS), 2024 (IF), 1992 (IS), 1783 (DPHL v1), and 1737 (library-free) up-regulated (adjusted p-value < 0.01 & $\log_2$ (fold-change) > 1) proteins, and 330 (RF), 359 (RS), 346 (IF), 370 (IS), 255 (DPHL v1), and 230 (library-free) down-regulated (adjust p-value < 0.01 & $\log_2$ (fold-change) < -1) proteins (Figure 4B). Compared with the DPHLv1, the numbers of identified and significantly dysregulated proteins increased by at least 21.7% (RF) and 14.2% (RF). Compared with the analysis using only SwissProt reviewed proteins sequences, 463 and 472 differentially expressed protein isoforms were identified using IF and IS, respectively. Similarly, 94 and 92 proteins were dysregulated in the CRC tissues compared with the benign samples by semi-specific digestion modes. These findings show that DPHL v2 allows identifying a larger number of differentially expressed proteins or protein isoforms between tumors and benign samples, providing more options for subsequent investigations.

We next used our four libraries and DPHL v1 to analyze the CRC cohort using the sub-library strategy[27], which refines a pan-human spectral library based on the tissue specificity. Compared with the conventional library search method, the sub-library strategy improved our results in all aspects (Figure S8A-C). First, the missing values were reduced by about 1% on average. The protein identifications increased by 22 (RF), 344 (RS), 103 (IF), 405 (IS), or 193 (DPHL v1). In the subsequent differential expression analysis, the total number of dysregulated proteins increased by 70 (RF), 163 (RS), 42 (IF), 203 (IS), and 20 (DPHL v1).

Finally, we built a random forest model based on the overlap dysregulated proteins generated by the four libraries to find new biomarkers. The 241 samples with 1426 proteins were randomly divided into the training set (N = 200) and the test set (N = 41). After a 5-fold cross validation, we identified 14 features that provided the highest accuracy for colorectal cancer, including S100A11, CEACAM6, GARS1, CDYL2, POTEKP, SCGN, SNCG, S100B, SCG2, NCAM1, OGN, CD81, COL28A1, CNRIP1 (Figure 5A). The area under the curve (AUC) of the training set and the test set achieved 1, 0.903 (Figure 5B), and the accuracy (ACC) achieved 0.988, 0.927, respectively (Figure 5C). Among these, S100A11[28, 29], CEACAM6[30, 31], CDYL2[32], SCGN[33], SNCG[34, 35], S100B[36], SCG2[37, 38], NCAM1[39], OGN[40], CD81[41], CNRIP1[42], have been reported to be closely related to colorectal cancer. Three features (GARS1, POTEKP, COL28A1) may be new biomarkers for colorectal cancer.

Analysis of protein isoforms and semi-tryptic peptides

We next checked whether this resource could be used to analyze specific protein isoform. Among the dysregulated proteins from IF , we identified SPTBN1 (SPTBN1-long) and one of its isoforms (SPTBN1-short)[43]. As reported in literature, SPTBN1 is significantly dysregulated and plays an essential role in liver cancer[44], colorectal cancer, and breast cancer, among others[45, 46]. To assess the accuracy of the identification, we showed the sequence of SPTBN1-long and SPTBN1-short identified in the library, in addition to the common parts of the two sequences, our library had also identified the peptide (TSSISGPLSPAYTGQVPYNYNQLEGR) specific in SPTBN1-short (Figure 6A). The Skyline software (Skyline-daily version) was used to show the peak spectrum of this peptide and a common peptide form these two proteins within the DIA raw file (Figure 6B-C).

Regarding those were only characterized through semi-specific peptides in our semi-specific libraries (IS and RS), including VWF, LMO7, ALDH2, NPEPL1, NUAK1, and TPT1, many of them have important biologic implications. ADAM22 is a new therapeutic option for treating metastatic brain disease and may be appropriate for treatment of breast cancer[47, 48]. By analyzing mRNA expression profiles, Xin et al.

found that *ASPM* is highly expressed in GBM, and patients with high *ASPM* expression have poor prognoses[49]. LRP6 inhibits cell proliferation and delays tumor growth in vivo, especially in colon, liver, breast, and pancreatic cancers[50, 51]. CHD9 was reported as a potential biomarker for clear cell renal cell carcinoma[52]. In addition, FAIM2 promotes non-small cell lung cancer growth and bone metastasis formation by regulating the epithelial-mesenchymal transformation process and the Wnt/β-catenin signaling pathway[53]. In our analysis, all these proteins showed significant differences between tumor and non-tumor samples, indicating that DPHL v2 can assist with the discovery of new potential protein biomarkers.

Conclusion

We present DPHL v2: four comprehensive spectral libraries (RF, RS, IF, and IS) derived from 1608 DDA MS raw files, including 24 sample types. By identifying over 440,000 peptides and more than 14,000 proteins, DPHL v2 can confidently detect and quantify more than 66.1% of the reviewed human proteins annotated by UniProtKB/Swiss-Prot. Our results suggest that DPHL v2 could support protein biomarker identification, especially for protein isoforms and semi-tryptic peptides. DPHL v2 outperforms previous DIA libraries in the following aspects. Firstly, five additional tissue types (oral cavity, thymus, esophagus, eyelid, and ovary) and one blood plasma sample from T-ALL were included. Secondly, protein isoforms and semi-trypsin digestion were used for library searching. In addition, these libraries are compatible with various commonly used DIA tools, with or without format transformation, such as OpenSWATH[54], DIA-NN[55], Skyline[56], and Spectronaut[57].

Materials and Methods

All chemicals used in this study were purchased from Sigma. All MS-grade reagents were acquired from Thermo Fisher Scientific (Waltham, MA).

Clinical samples

Formalin-fixed paraffin-embedded, fresh or fresh frozen tissue biopsies from GBM, healthy human brain, eyelid tumor, thyroid disease, sarcoma, OSCC, thymus, LUAD, TNBC, HCC, gastric cancer, diffuse large B-cell lymphoma, pancreatic ductal adenocarcinoma, bladder cancer, PCa, and OV were collected in this study. Human plasma samples, including acute lymphoblastic leukemia (ALL), AML, T-ALL, normal plasma exosome, and blood disease, were also analyzed, as well as K562 cells. Six of these tissues were new additions compared to the DPHL v1. Eyelid samples were obtained from the Second Affiliated Hospital of Zhejiang University School of Medicine, China. The ovary cohort was obtained from The Cancer Hospital of the University of Chinese Academy of Sciences. The OSCC, esophagus, T-ALL, and thymus cancer samples were collected at Amsterdam UMC/VU Medical Center, Amsterdam, and Erasmus University Medical Center. Sample details are provided in Table S1.

To compare our libraries with the DPHL v1 and library-free method, we used the DIA data of a CRC cohort generated by Ge et al.[27], which consists of 201 cancer samples, 40 para-cancer tissues, and 45 biological and technical replicates from 40 CRC patients and four healthy controls. The detailed sample information is given in Table S2.

MS Data acquisition

Among the newly added 586 DDA raw data files, 108 were derived from Dutch cohorts generated at the Jimenez lab and 404 from Chinese cohorts generated at the Guo lab. The pipeline for generating these DDA files coincided with that used for the DPHL v1. The DDA raw files were centroided and converted into mzXML using ProteoWizard[58] (version 3.0.11579). Carbamidomethylation was set as fixed modification at cysteine residues; oxidation was set as variable modification at methionine residues.

DIA data analysis

The DIA raw files were submitted to DIA-NN (1.7.15), a tool for DIA or SWATH proteomics data analysis[55]. Our four libraries were used as a reference, and no other fasta sequences were added. The library inference was set to "off". All other parameters were kept to their default values. The tools we used for the DIA data analysis, as described above, are publicly available[55].

Machine learning

The random forest analysis was performed with the R package "randomForest" (version 4.6.14). 1426 proteins were firstly selected as input features to build 1000 trees with 5-fold cross validation and repeated 10 times to optimize the model. The Mean Decrease Accuracy was set 4 to 6, with step size of 0.5. The final performance was evaluated by mean accuracy (ACC) and mean area under curve (AUC) in a receiver operating characteristic curve across 5-folds.

Ethical statement

Ethics approvals for this study were obtained from the Ethics Committee or Institutional Review Board of each participating institution.

Acknowledgments

This work is supported by grants from National Key R&D Program of China (2021YFA1301603; 2021YFA1301602; 2020YFE0202200).

Author contributions

T.G. conceived the project. Z.X. and T.Z. built all the libraries. Z.X. processed and analyzed data. T.G., Z.X., T.Z., J.A., and F.Z. wrote the manuscript. Y.L. collected the brain samples. J.Y. provided the eyelid tumor samples. T.L. offered the lung cancer samples. J.Z. and C.L. collected the liver cancer samples. Y.H. offered the prostate cancer samples. Q.W. provided the cervix cancer samples. J.Z. and Z.Z. collected the ovarian cancer samples. Others prepared peptides for the study. T.G. supervised the work. All authors reviewed and approved the manuscript.

Competing interests

T.G. is shareholder of Westlake Omics Inc. N.X., X.Y. and W.L. are employees of Westlake Omics Inc. The other authors declare no competing interests in this paper.

Data Availability

All newly added raw DDA MS data, spectral libraries, and protein results are publicly available at iProX[59] (PXD015314) and ProteomeXchange (PXD015314). All the R scripts were uploaded to GitHub (https://github.com/zhutiansheng/DPHLv2).

References

[1] Y. Zhu, R. Aebersold, M. Mann, T. Guo. (2021). SnapShot: Clinical proteomics, Cell, 184, 4840-4840 e4841. https://doi.org/10.1016/j.cell.2021.08.015.

[2] Q. Xiao, F. Zhang, L. Xu, L. Yue, O.L. Kon, Y. Zhu, T. Guo. (2021). High-throughput proteomics and AI for cancer biomarker discovery, Adv Drug Deliv Rev, 176, 113844. https://doi.org/10.1016/j.addr.2021.113844.

[3] R. Aebersold, M. Mann. (2016). Mass-spectrometric exploration of proteome structure and function, Nature, 537, 347-355. https://doi.org/10.1038/nature19949.

[4] V. Lange, P. Picotti, B. Domon, R. Aebersold. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial, Mol Syst Biol, 4, 222. https://doi.org/10.1038/msb.2008.61.

[5] A.C. Peterson, J.D. Russell, D.J. Bailey, M.S. Westphall, J.J. Coon. (2012). Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics, Mol Cell Proteomics, 11, 1475-1488. https://doi.org/10.1074/mcp.O112.020131.

[6] L.C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis, Mol Cell Proteomics, 11, O111.016717. https://doi.org/10.1074/mcp.O111.016717.

[7] F. Zhang, W. Ge, G. Ruan, X. Cai, T. Guo. (2020). Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020, Proteomics, 20, e1900276. https://doi.org/10.1002/pmic.201900276.

[8] G. Rosenberger, C.C. Koh, T. Guo, H.L. Rost, P. Kouvonen, B.C. Collins, M. Heusel, Y. Liu, E. Caron, A. Vichalkovski, M. Faini, O.T. Schubert, P. Faridi, H.A. Ebhardt, M. Matondo, H. Lam, S.L. Bader, D.S. Campbell, E.W. Deutsch, R.L. Moritz, S. Tate, R. Aebersold. (2014). A repository of assays to quantify 10,000 human proteins by SWATH-MS, Sci Data, 1, 140031. https://doi.org/10.1038/sdata.2014.31.

[9] T. Zhu, Y. Zhu, Y. Xuan, H. Gao, X. Cai, S.R. Piersma, T.V. Pham, T. Schelfhorst, R. Haas, I.V. Bijnsdorp, R. Sun, L. Yue, G. Ruan, Q. Zhang, M. Hu, Y. Zhou, W.J. Van Houdt, T.Y.S. Le Large, J. Cloos, A. Wojtuszkiewicz, D. Koppers-Lalic, F. Bottger, C. Scheepbouwer, R.H. Brakenhoff, G. van Leenders, J.N.M. Ijzermans, J.W.M. Martens, R.D.M. Steenbergen, N.C. Grieken, S. Selvarajan, S. Mantoo, S.S. Lee, S.J.Y. Yeow, S.M.F. Alkaff, N. Xiang, Y. Sun, X. Yi, S. Dai, W. Liu, T. Lu, Z. Wu, X. Liang, M. Wang, Y. Shao, X. Zheng, K. Xu, Q. Yang, Y. Meng, C. Lu, J. Zhu, J. Zheng, B. Wang, S. Lou, Y. Dai, C. Xu, C. Yu, H. Ying, T.K. Lim, J. Wu, X. Gao, Z. Luan, X. Teng, P. Wu, S. Huang, Z. Tao, N.G. Iyer, S. Zhou, W. Shao, H. Lam, D. Ma, J. Ji, O.L. Kon, S. Zheng, R. Aebersold, C.R. Jimenez, T. Guo. (2020). DPHL: A DIA Pan-human Protein Mass Spectrometry Library for Robust Biomarker Discovery, Genomics Proteomics Bioinformatics, 18, 104-119. https://doi.org/10.1016/j.gpb.2019.11.008.

[10] T. Lu, L. Qian, Y. Xie, Q. Zhang, W. Liu, W. Ge, Y. Zhu, L. Ma, C. Zhang, T. Guo. (2022). Tissue-Characteristic Expression of Mouse Proteome, Mol Cell Proteomics, 21, 100408. https://doi.org/10.1016/j.mcpro.2022.100408.

[11] P. Blattmann, V. Stutz, G. Lizzo, J. Richard, P. Gut, R. Aebersold. (2019).

Generation of a zebrafish SWATH-MS spectral library to quantify 10,000 proteins, Sci Data, 6, 190011. https://doi.org/10.1038/sdata.2019.11.

[12] H. Zhang, P. Liu, T. Guo, H. Zhao, D. Bensaddek, R. Aebersold, L. Xiong. (2019). Arabidopsis proteome and the mass spectral assay library, Sci Data, 6, 278. https://doi.org/10.1038/s41597-019-0294-0.

[13] M.K. Midha, U. Kusebauch, D. Shteynberg, C. Kapil, S.L. Bader, P.J. Reddy, D.S. Campbell, N.S. Baliga, R.L. Moritz. (2020). A comprehensive spectral assay library to quantify the Escherichia coli proteome by DIA/SWATH-MS, Sci Data, 7, 389. https://doi.org/10.1038/s41597-020-00724-7.

[14] G.S. Omenn, L. Lane, C.M. Overall, C. Pineau, N.H. Packer, I.M. Cristea, C. Lindskog, S.T. Weintraub, S. Orchard, M.H.A. Roehrl, E. Nice, S. Liu, N. Bandeira, Y.J. Chen, T. Guo, R. Aebersold, R.L. Moritz, E.W. Deutsch. (2022). The 2022 Report on the Human Proteome from the HUPO Human Proteome Project, J Proteome Res. https://doi.org/10.1021/acs.jproteome.2c00498.

[15] S. Adhikari, E.C. Nice, E.W. Deutsch, L. Lane, G.S. Omenn, S.R. Pennington, Y.K. Paik, C.M. Overall, F.J. Corrales, I.M. Cristea, J.E. Van Eyk, M. Uhlen, C. Lindskog, D.W. Chan, A. Bairoch, J.C. Waddington, J.L. Justice, J. LaBaer, H. Rodriguez, F. He, M. Kostrzewa, P. Ping, R.L. Gundry, P. Stewart, S. Srivastava, S. Srivastava, F.C.S. Nogueira, G.B. Domont, Y. Vandenbrouck, M.P.Y. Lam, S. Wennersten, J.A. Vizcaino, M. Wilkins, J.M. Schwenk, E. Lundberg, N. Bandeira, G. Marko-Varga, S.T. Weintraub, C. Pineau, U. Kusebauch, R.L. Moritz, S.B. Ahn, M. Palmblad, M.P. Snyder, R. Aebersold, M.S. Baker. (2020). A high-stringency blueprint of the human proteome, Nat Commun, 11, 5301. https://doi.org/10.1038/s41467-020-19045-9.

[16] M. Uhlen, L. Fagerberg, B.M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C.A. Szigyarto, J. Odeberg, D. Djureinovic, J.O. Takanen, S. Hober, T. Alm, P.H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J.M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Ponten. (2015). Proteomics. Tissue-based map of the human proteome, Science, 347, 1260419. https://doi.org/10.1126/science.1260419.

[17] M. Stastna, J.E. Van Eyk. (2012). Analysis of protein isoforms: can we do it better?, Proteomics, 12, 2937-2948. https://doi.org/10.1002/pmic.201200161.

[18] P. Mallick, M. Schirle, S.S. Chen, M.R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, R. Aebersold. (2007). Computational prediction of proteotypic peptides for quantitative proteomics, Nat Biotechnol, 25, 125-131. https://doi.org/10.1038/nbt1275.

[19] H. Tang, R.J. Arnold, P. Alves, Z. Xun, D.E. Clemmer, M.V. Novotny, J.P. Reilly, P. Radivojac. (2006). A computational approach toward label-free protein quantification using predicted peptide detectability, Bioinformatics, 22, e481-488. https://doi.org/10.1093/bioinformatics/btl237.

[20] D.J. States, G.S. Omenn, T.W. Blackwell, D. Fermin, J. Eng, D.W. Speicher, S.M. Hanash. (2006). Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study, Nature Biotechnology, 24, 333-338. https://doi.org/10.1038/nbt1183.

[21] A.T. Kong, F.V. Leprevost, D.M. Avtonomov, D. Mellacheruvu, A.I. Nesvizhskii. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics, Nat Methods, 14, 513-520. https://doi.org/10.1038/nmeth.4256.

[22] M. Magrane, C. UniProt. (2011). UniProt Knowledgebase: a hub of integrated protein data, Database (Oxford), 2011, bar009. https://doi.org/10.1093/database/bar009.

[23] F. da Veiga Leprevost, S.E. Haynes, D.M. Avtonomov, H.Y. Chang, A.K. Shanmugam, D. Mellacheruvu, A.T. Kong, A.I. Nesvizhskii. (2020). Philosopher: a versatile toolkit for shotgun proteomics data analysis, Nat Methods, 17, 869-870. https://doi.org/10.1038/s41592-020-0912-y.

[24] M.K. Midha, D.S. Campbell, C. Kapil, U. Kusebauch, M.R. Hoopmann, S.L. Bader, R.L. Moritz. (2020). DIALib-QC an assessment tool for spectral libraries in data-independent acquisition proteomics, Nat Commun, 11, 5251. https://doi.org/10.1038/s41467-020-18901-y.

[25] T.P. Subramanian A, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. . Gene set enrichment analysis a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A, 102. https://doi.org/10.1073/pnas.0506580102.

[26] A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J.P. Mesirov, P. Tamayo. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection, Cell Syst, 1, 417-425. https://doi.org/10.1016/j.cels.2015.12.004.

[27] W. Ge, X. Liang, F. Zhang, Y. Hu, L. Xu, N. Xiang, R. Sun, W. Liu, Z. Xue, X. Yi, Y. Sun, B. Wang, J. Zhu, C. Lu, X. Zhan, L. Chen, Y. Wu, Z. Zheng, W. Gong, Q. Wu, J. Yu, Z. Ye, X. Teng, S. Huang, S. Zheng, T. Liu, C. Yuan, T. Guo. (2021). Computational Optimization of Spectral Library Size Improves DIA-MS Proteome Coverage and Applications to 15 Tumors, J Proteome Res, 20, 5392-5401. https://doi.org/10.1021/acs.jproteome.1c00640.

[28] A.J. Guo, F.J. Wang, Q. Ji, H.W. Geng, X. Yan, L.Q. Wang, W.W. Tie, X.Y. Liu, R.F. Thorne, G. Liu, A.M. Xu. (2021). Proteome Analyses Reveal S100A11, S100P, and RBM25 Are Tumor Biomarkers in Colorectal Cancer, Proteomics Clin Appl, 15, e2000056. https://doi.org/10.1002/prca.202000056.

[29] Y. Niu, Z. Shao, H. Wang, J. Yang, F. Zhang, Y. Luo, L. Xu, Y. Ding, L. Zhao. (2016). LASP1-S100A11 axis promotes colorectal cancer aggressiveness by modulating TGFbeta/Smad signaling, Sci Rep, 6, 26112. https://doi.org/10.1038/srep26112.

[30] E. Ferlizza, R. Solmi, R. Miglio, E. Nardi, G. Mattei, M. Sgarzi, M. Lauriola. (2020). Colorectal cancer screening: Assessment of CEACAM6, LGALS4, TSPAN8 and COL1A2 as blood markers in faecal immunochemical test negative subjects, J Adv Res, 24, 99-107. https://doi.org/10.1016/j.jare.2020.03.001.

[31] M.T. Rodia, R. Solmi, F. Pasini, E. Nardi, G. Mattei, G. Ugolini, L. Ricciardiello, P. Strippoli, R. Miglio, M. Lauriola. (2018). LGALS4, CEACAM6, TSPAN8, and COL1A2: Blood Markers for Colorectal Cancer-Validation in a Cohort of Subjects With Positive Fecal Immunochemical Test Result, Clin Colorectal Cancer, 17, e217-e228. https://doi.org/10.1016/j.clcc.2017.12.002.

[32] S.I. Kim ST, DO IG, Jang J, Kim SH, Jung IH, Park JO, Park YS, Talasaz A, Lee J, Kim HC. . (2014). Transcriptome analysis of CD133-positive stem cells and prognostic value of survivin in colorectal cancer, Cancer Genomics Proteomics, 11, 259-266.

[33] X.Y. Yang, Q.R. Liu, L.M. Wu, X.L. Zheng, C. Ma, R.S. Na. (2018). Overexpression of secretagogin promotes cell apoptosis and inhibits migration and invasion of human SW480 human colorectal cancer cells, Biomed Pharmacother, 101, 342-347. https://doi.org/10.1016/j.biopha.2018.01.147.

[34] H. Hu, L. Sun, C. Guo, Q. Liu, Z. Zhou, L. Peng, J. Pan, L. Yu, J. Lou, Z. Yang, P. Zhao, Y. Ran. (2009). Tumor cell-microenvironment interaction models coupled with clinical validation reveal CCL2 and SNCG as two predictors of colorectal cancer hepatic metastasis, Clin Cancer Res, 15, 5485-5493. https://doi.org/10.1158/1078-0432.CCR-08-2491.

[35] D.B. Liu C, Lu A, Qu L, Xing X, Meng L, Wu J, Eric Shi Y, Shou C. . (2010). Synuclein gamma predicts poor clinical outcome in colon cancer with normal levels of carcinoembryonic antigen, BMC Cancer, 359, 1471-2407.

[36] M.Y. Huang, H.M. Wang, T.S. Tok, H.J. Chang, M.S. Chang, T.L. Cheng, J.Y. Wang, S.R. Lin. (2012). EVI2B, ATP2A2, S100B, TM4SF3, and OLFM4 as potential prognostic markers for postoperative Taiwanese colorectal cancer patients, DNA Cell Biol, 31, 625-635. https://doi.org/10.1089/dna.2011.1365.

[37] H. Wang, J. Yin, Y. Hong, A. Ren, H. Wang, M. Li, Q. Zhao, C. Jiang, L. Liu. (2021). SCG2 is a Prognostic Biomarker Associated With Immune Infiltration and Macrophage Polarization in Colorectal Cancer, Front Cell Dev Biol, 9, 795133. https://doi.org/10.3389/fcell.2021.795133.

[38] S. Weng, Z. Liu, X. Ren, H. Xu, X. Ge, Y. Ren, Y. Zhang, Q. Dang, L. Liu, C. Guo, R. Beatson, J. Deng, X. Han. (2022). SCG2: A Prognostic Marker That Pinpoints Chemotherapy and Immunotherapy in Colorectal Cancer, Front Immunol, 13, 873871. https://doi.org/10.3389/fimmu.2022.873871.

[39] T. Kok-Sin, N.M. Mokhtar, N.Z. Ali Hassan, I. Sagap, I. Mohamed Rose, R. Harun, R. Jamal. (2015). Identification of diagnostic markers in colorectal cancer via integrative epigenomics and genomics data, Oncol Rep, 34, 22-32. https://doi.org/10.3892/or.2015.3993.

[40] X. Hu, Y.Q. Li, Q.G. Li, Y.L. Ma, J.J. Peng, S.J. Cai. (2018). Osteoglycin (OGN) reverses epithelial to mesenchymal transition and invasiveness in colorectal cancer via EGFR/Akt pathway, J Exp Clin Cancer Res, 37, 41. https://doi.org/10.1186/s13046-018-0718-2.

[41] H. Yuan, J. Zhao, Y. Yang, R. Wei, L. Zhu, J. Wang, M. Ding, M. Wang, Y. Gu. (2020). SHP-2 Interacts with CD81 and Regulates the Malignant Evolution of Colorectal Cancer by Inhibiting Epithelial-Mesenchymal Transition, Cancer Manag Res, 12, 13273-13284. https://doi.org/10.2147/CMAR.S270813.

[42] T. Zhang, G. Cui, Y.L. Yao, Q.C. Wang, H.G. Gu, X.N. Li, H. Zhang, W.M. Feng, Q.L. Shi, W. Cui. (2017). Value of CNRIP1 promoter methylation in colorectal cancer screening and prognosis assessment and its influence on the activity of cancer cells, Arch Med Sci, 13, 1281-1294. https://doi.org/10.5114/aoms.2017.65829.

[43] N.V. L., Hayes, Catherine Scott, Egidius Heerkens, Vasken Ohanian, Alison M. Maggs, J. C., Pinder, Ekaterini Kordeli, A.J. Baines. (2000). Identification of a novel C-terminal variant of βII spectrin two isoforms of βII spectrin have distinct intracellular locations and activities, Journal of Cell Science, 113, 2023-2034.

[44] Shuyun Rao, Xiaochun Yang, Kazufumi Ohshiro, Sobia Zaidi, Zhanhuai Wang, Kirti Shetty, Xiyan Xiang, Md. Imtaiyaz Hassan, Taj Mohammad, Patricia S. Latham, Bao-Ngoc Nguyen, Linda Wong, Herbert Yu, Yousef Al-Abed, Bibhuti Mishra, Michele Vacca, Gareth Guenigault, Michael E. D. Allison, Antonio Vidal-Puig, Jihane N. Benhammou, Marcus Alvarez, Päivi Pajukanta, Joseph R. Pisegna, L. Mishra. (2021). β2-spectrin (SPTBN1) as a therapeutic target for diet-induced liver disease and preventing cancer development, SCIENCE TRANSLATIONAL MEDICINE, 13.

[45] P. Yang, Y. Yang, P. Sun, Y. Tian, F. Gao, C. Wang, T. Zong, M. Li, Y. Zhang, T. Yu, Z. Jiang. (2021). betaII spectrin (SPTBN1): biological function and clinical potential in cancer and other diseases, Int J Biol Sci, 17, 32-49. https://doi.org/10.7150/ijbs.52375.

[46] Z.X. Yao, W. Jogunoori, S. Choufani, A. Rashid, T. Blake, W. Yao, P. Kreishman, R. Amin, A.A. Sidawy, S.R. Evans, M. Finegold, E.P. Reddy, B. Mishra, R. Weksberg, R. Kumar, L. Mishra. (2010). Epigenetic silencing of beta-spectrin, a TGF-beta signaling/scaffolding protein in a human cancer stem cell disorder: Beckwith-Wiedemann syndrome, J Biol Chem, 285, 36112-36120. https://doi.org/10.1074/jbc.M110.162347.

[47] S. Charmsaz, B. Doherty, S. Cocchiglia, D. Vareslija, A. Marino, N. Cosgrove, R. Marques, N. Priedigkeit, S. Purcell, F. Bane, J. Bolger, C. Byrne, P.J. O'Halloran, F. Brett, K. Sheehan, K. Brennan, A.M. Hopkins, S. Keelan, P. Jagust, S. Madden, C. Martinelli, M. Battaglini, S. Oesterreich, A.V. Lee, G. Ciofani, A.D.K. Hill, L.S. Young. (2020). ADAM22/LGI1 complex as a new actionable target for breast cancer brain metastasis, BMC Med, 18, 349. https://doi.org/10.1186/s12916-020-01806-4.

[48] J. Li, M. Lu, J. Jin, X. Lu, T. Xu, S. Jin. (2018). miR-449a Suppresses Tamoxifen Resistance in Human Breast Cancer Cells by Targeting ADAM22, Cell Physiol Biochem, 50, 136-149. https://doi.org/10.1159/000493964.

[49] Xin Chen, Lijie Huang, Yang Yang, Suhua Chen, Jianjun Sun, Changcheng Ma, Jingcheng Xie, Yongmei Song, J. Yang. ASPM promotes glioblastoma growth by regulating G1 restriction point progression and Wnt-β-catenin signaling, 224-241.

[50] J. Raisch, A. Cote-Biron, N. Rivard. (2019). A Role for the WNT Co-Receptor LRP6 in Pathogenesis and Therapy of Epithelial Cancers, Cancers (Basel), 11. https://doi.org/10.3390/cancers11081162.

[51] J. Zhang, J. Chen, D. Wo, H. Yan, P. Liu, E. Ma, L. Li, L. Zheng, D. Chen, Z. Yu, C. Liang, J. Peng, D.N. Ren, W. Zhu. (2019). LRP6 Ectodomain Prevents SDF-1/CXCR4-Induced Breast Cancer Metastasis to Lung, Clin Cancer Res, 25, 4832-4845. https://doi.org/10.1158/1078-0432.CCR-18-3557.

[52] Bo Guan, Xian-Gui Ran, Yong-Qiang Du, Feng Ren4, Ye Tian, Ying Wang, M.-M. Chen. (2018). High CHD9 expression is associated with poor prognosis in clear cell renal cell carcinoma, Int J Clin Exp Pathol, 11, 3697-3702.

[53] K. She, W. Yang, M. Li, W. Xiong, M. Zhou. (2021). FAIM2 Promotes Non-Small

Cell Lung Cancer Cell Growth and Bone Metastasis by Activating the Wnt/beta-Catenin Pathway, Front Oncol, 11, 690142. https://doi.org/10.3389/fonc.2021.690142.

[54] H.L. Rost, G. Rosenberger, P. Navarro, L. Gillet, S.M. Miladinovic, O.T. Schubert, W. Wolski, B.C. Collins, J. Malmstrom, L. Malmstrom, R. Aebersold. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data, Nat Biotechnol, 32, 219-223. https://doi.org/10.1038/nbt.2841.

[55] V. Demichev, C.B. Messner, S.I. Vernardis, K.S. Lilley, M. Ralser. (2020). DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput, Nat Methods, 17, 41-44. https://doi.org/10.1038/s41592-019-0638-x.

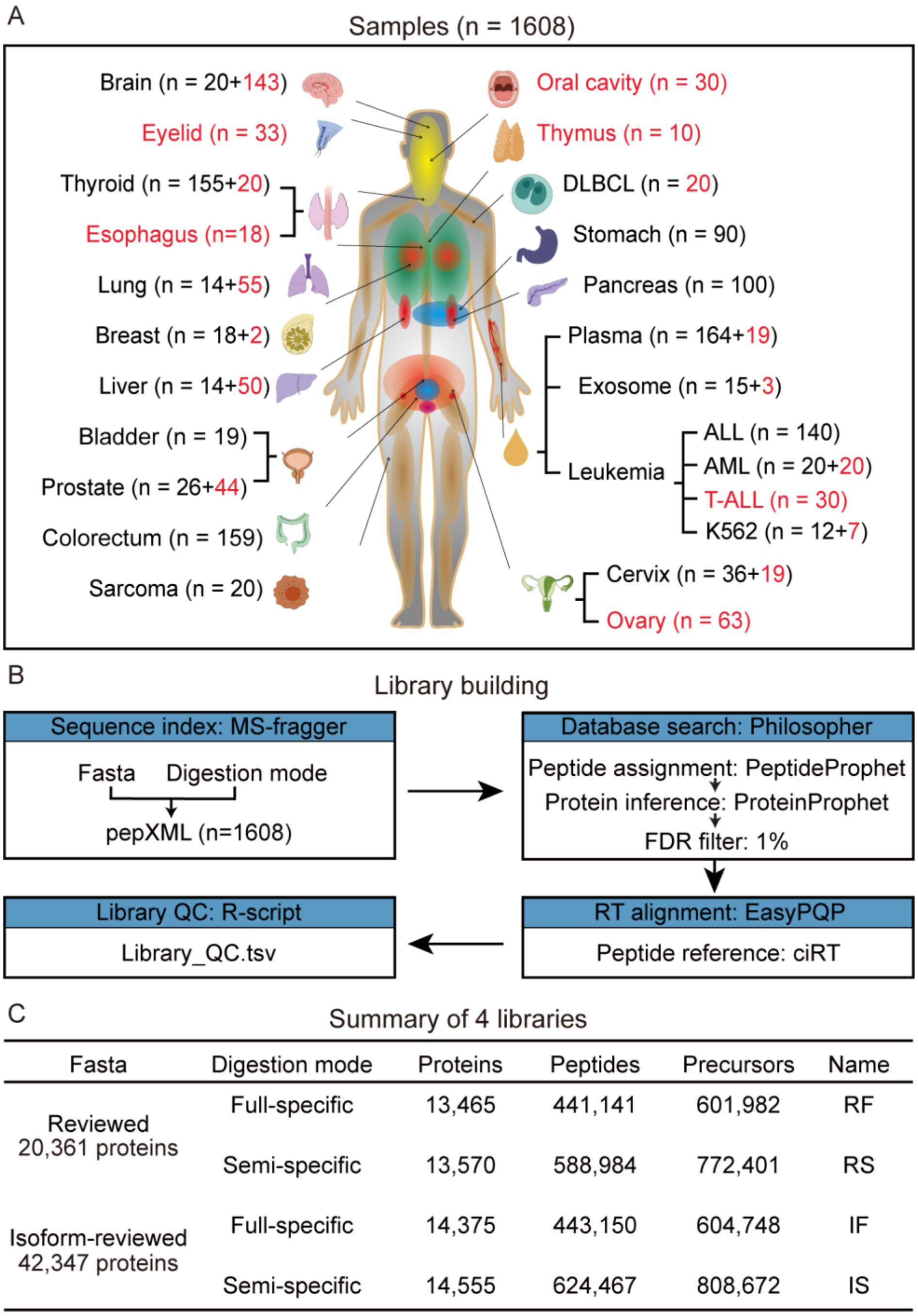
[56] B. MacLean, D.M. Tomazela, N. Shulman, M. Chambers, G.L. Finney, B. Frewen, R. Kern, D.L. Tabb, D.C. Liebler, M.J. MacCoss. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments, Bioinformatics, 26, 966-968. https://doi.org/10.1093/bioinformatics/btq054.

[57] Ana Martinez-Val, Dorte Breinholdt Bekker-Jensen, Alexander Hogrebe, J.V. Olsen. (2021). Data Processing and Analysis for DIA-Based Phosphoproteomics Using Spectronaut, Proteomics Data Analysis, Methods in Molecular Biology, 2361, 95-107. https://doi.org/https://doi.org/10.1007/978-1-0716-1641-3_6.

[58] M.C. Chambers, B. Maclean, R. Burke, D. Amodei, D.L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T.A. Baker, M.Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S.L. Seymour, L.M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E.W. Deutsch, R.L. Moritz, J.E. Katz, D.B. Agus, M. MacCoss, D.L. Tabb, P. Mallick. (2012). A cross-platform toolkit for mass spectrometry and proteomics, Nat Biotechnol, 30, 918-920. https://doi.org/10.1038/nbt.2377.

[59] J. Ma, T. Chen, S. Wu, C. Yang, M. Bai, K. Shu, K. Li, G. Zhang, Z. Jin, F. He, H. Hermjakob, Y. Zhu. (2019). iProX: an integrated proteome resource, Nucleic Acids Res, 47, D1211-D1217. https://doi.org/10.1093/nar/gky869.

Figure 1.



Fasta	Digestion mode	Proteins	Peptides	Precursors	Name
Reviewed 20,361 proteins	Full-specific	13,465	441,141	601,982	RF
	Semi-specific	13,570	588,984	772,401	RS
Isoform-reviewed 42,347 proteins	Full-specific	14,375	443,150	604,748	IF
	Semi-specific	14,555	624,467	808,672	IS

Figure 1. Sample types and workflow for building DPHL v2. (A) Number and type of samples included in this study. The ones that were missing from DPHL v1 are highlighted in red. (B) Computational pipeline for building DPHL v2. (C) Overview of the number of identified proteins, peptides, and precursors using our four library variants.

Figure 2.

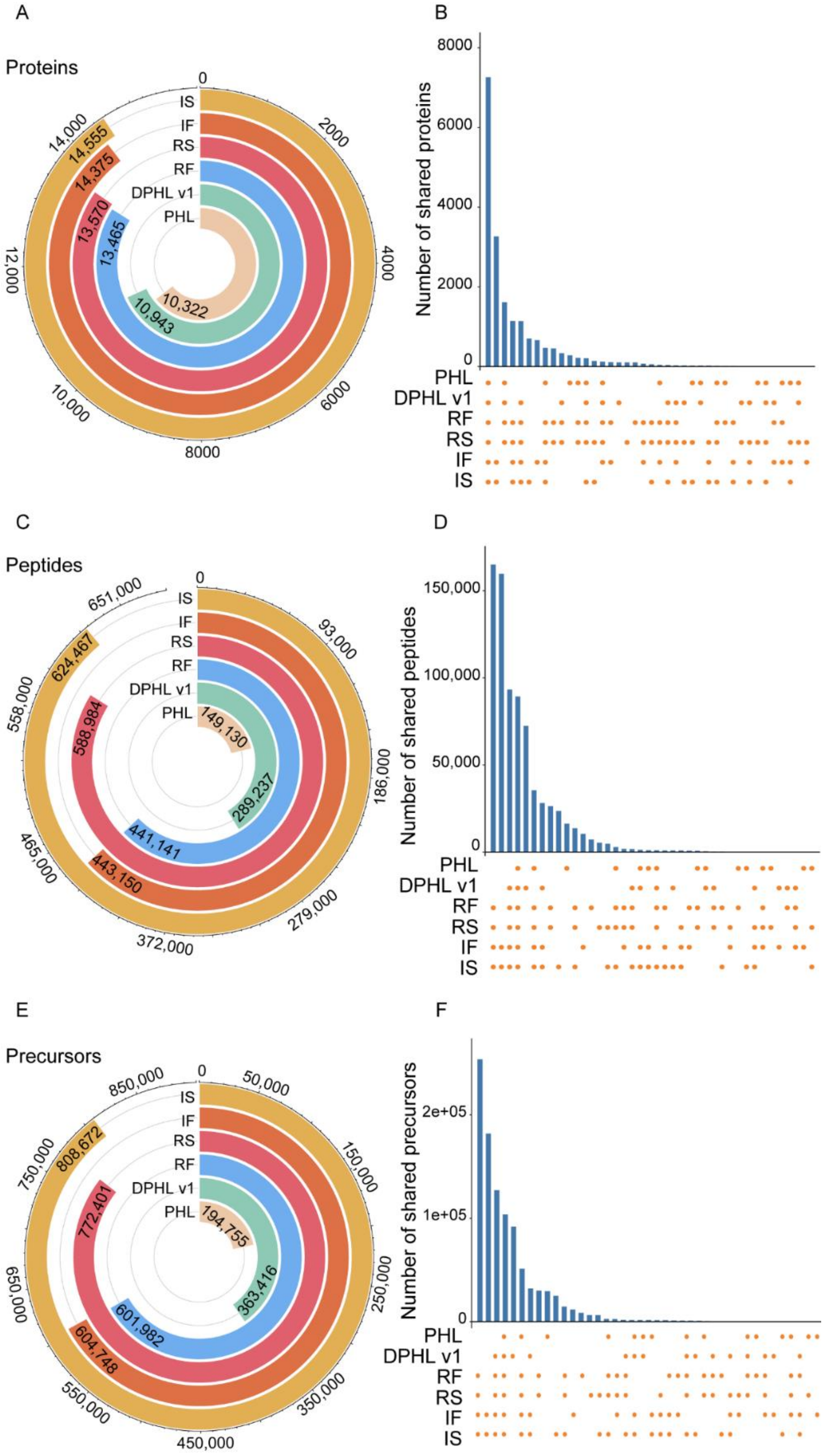


Figure 2. Comparison of the four variants of DPHL v2 (i.e., RF, RS, IF, and IS) with DPHL v1 and PHL. The circular bars show the protein (A), peptide (C), and precursor identifications (E) of the six libraries. The UpSet plots show the shared and

unique protein (B), peptide (D), and precursor identifications (F) of the six libraries. PHL, pan-human spectral library; DPHL v1, DIA pan-human library generated by Zhu *et al*; RF, reviewed fasta sequence & full-specific digestion mode; RS, reviewed fasta sequence & semi-specific digestion mode; IF, isoform fasta sequence & full-specific digestion mode; IS, isoform fasta sequence & semi-specific digestion mode.

Figure 3.

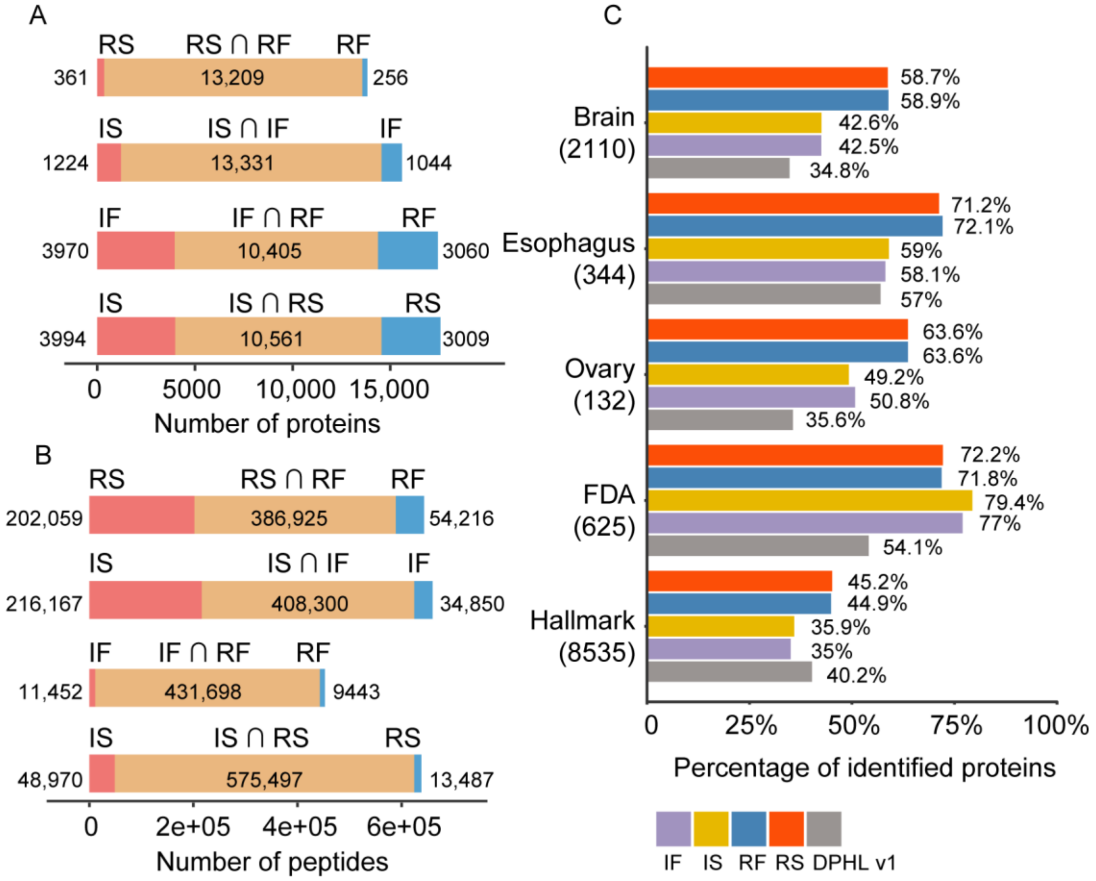


Figure 3. Comparison of the number of proteins (A) and peptides (B) identified with the same fasta sequence and the same digestion mode. (C) Percentage of proteins identified among DPHL v1 and our four libraries using hallmark gene sets, FDA-approved drug targets, and tissue-specific or tissue-enriched/enhanced proteins from brain, esophagus, and ovary samples.

Figure 4.

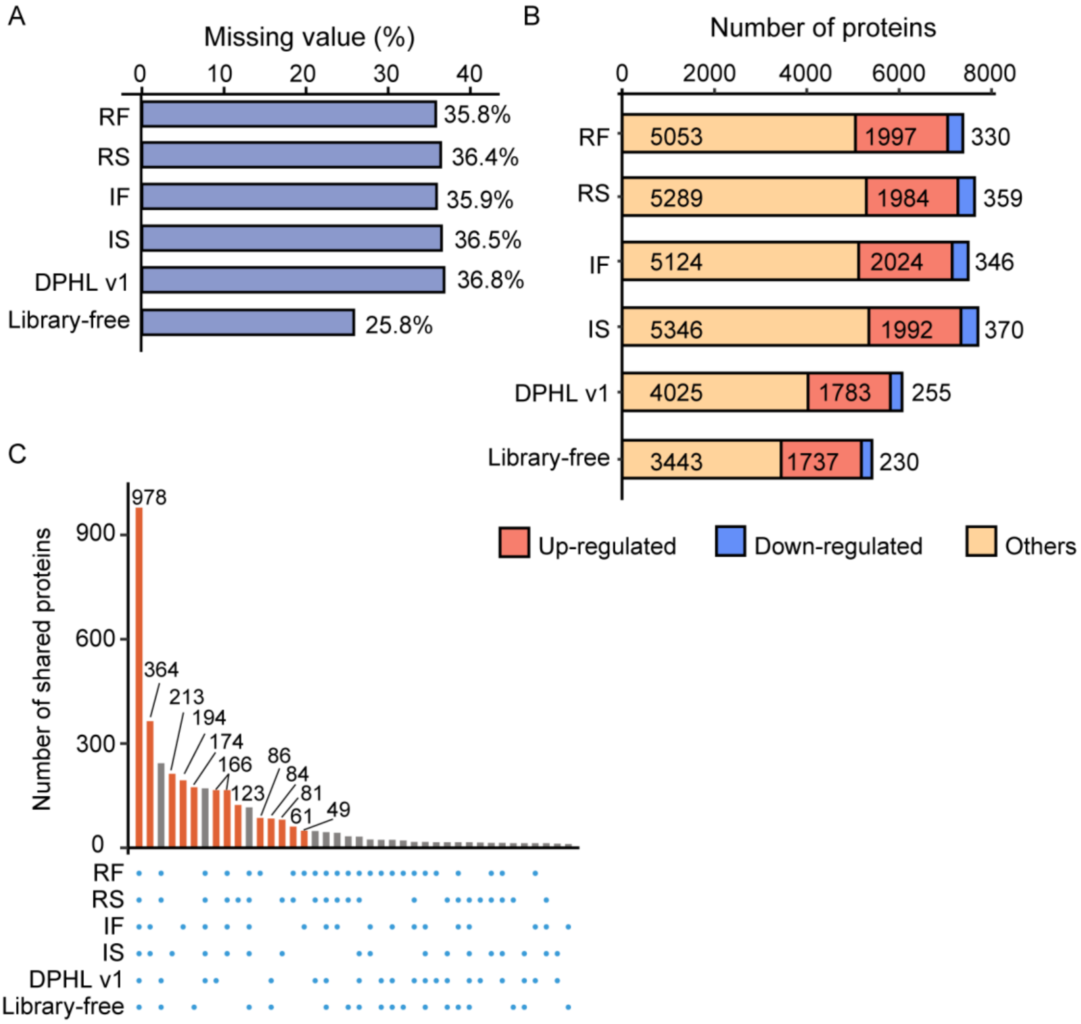


Figure 4. DIA analysis of CRC and benign samples. (A) Number of missing values obtained using the five libraries and library-free method. (B) Number of differentially expressed proteins between CRC and benign samples obtained using the five libraries and library-free method. Proteins with adjusted p-value < 0.01 and |FC| > 4 were selected as significantly differentially expressed. FC, fold change. (C) Protein identification overlaps across the six libraries.

Figure 5.

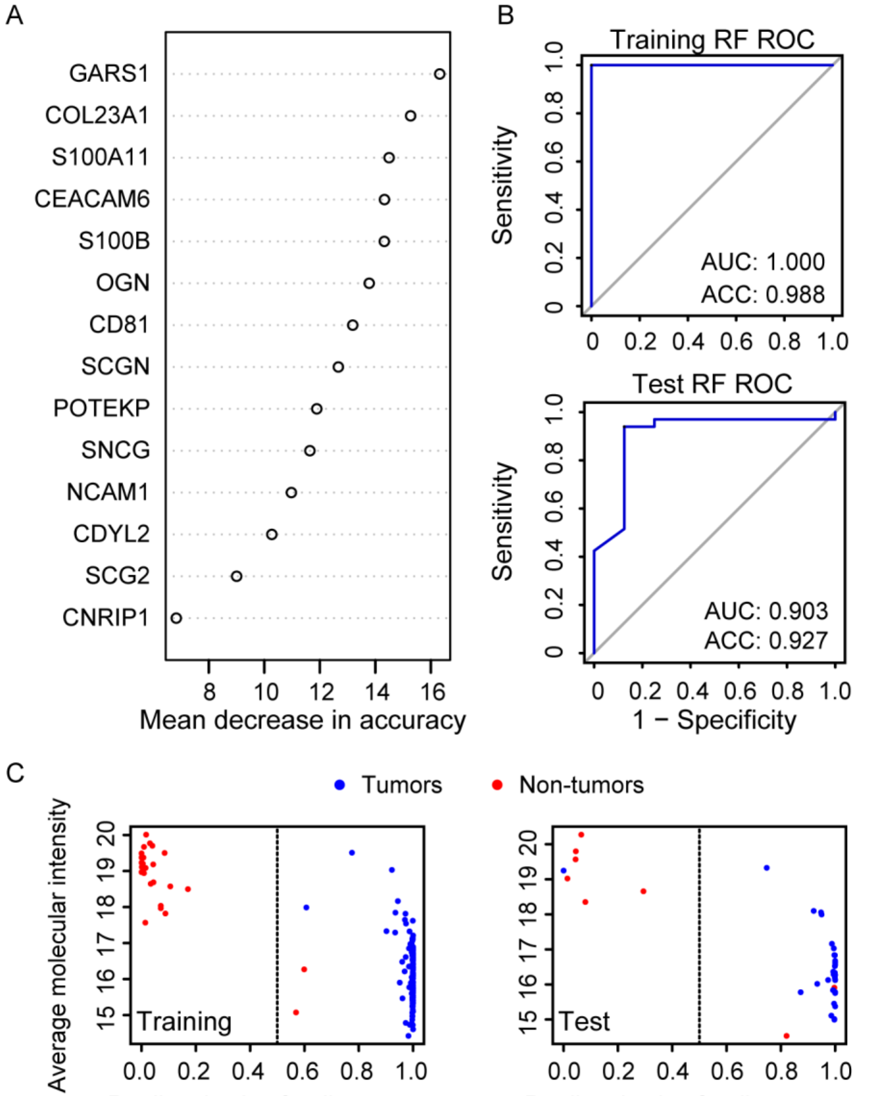


Figure 5. Machine learning to identify potential CRC biomarkers. (A) Prioritization of 14 important variables. (B) ROC plots for the training set (up) and the test set (down). (C) Performance of the model in the training set and test set.

Figure 6.

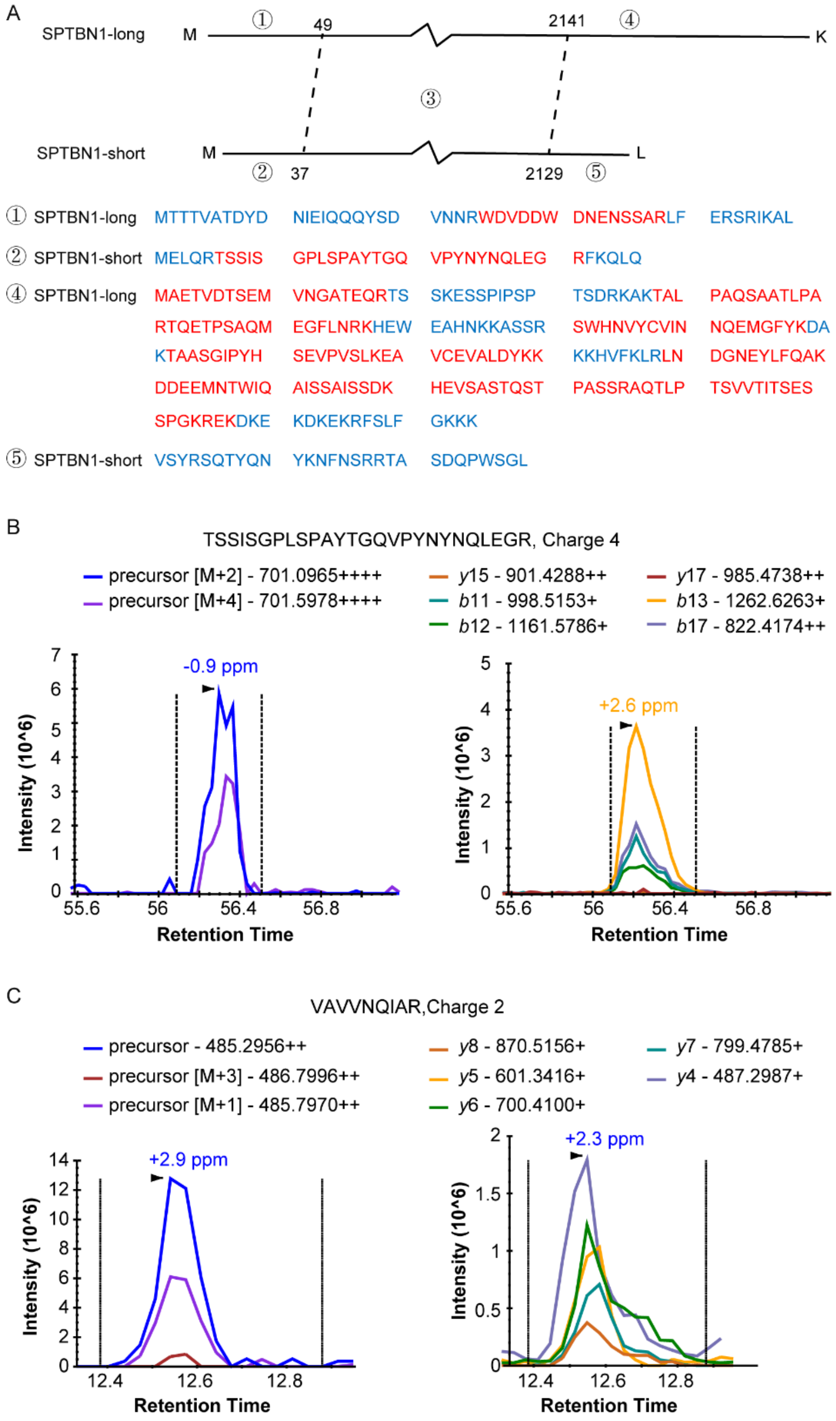


Figure 6. SPTBN1 protein identification in our DIA search results. (A) Sequences of SPTBN1 and its isoform. Blue: sequences that were not identified; red: identified sequences. (B) The peak spectrum of peptide SSISGPLSPAYTGQVPYNYNQLEGR in our DIA raw file (obtained using Skyline).