Optimization of Spectral Library Size Improves DIA-MS Proteome Coverage

Weigang Ge [1,2,3,4 #], Xiao Liang [1,2,3 #], Fangfei Zhang [1,2,3 #], Luang Xu [1,2,3], Nan Xiang [1,2,3], Rui Sun [1,2,3], Wei Liu [1,2,3] Zhangzhi Xue [1,2,3], Xiao Yi [1,2,3], Bo Wang [5], Jiang Zhu [6], Cong Lu [6], Xiaolu Zhan [7], Lirong Chen [8], Yan Wu [9,10], Zhiguo Zheng [11,12], Wangang Gong[11,12], Qijun Wu [13], Jiekai Yu [14], Zhaoming Ye [9,10], Xiaodong Teng [5], Shiang Huang [6], Shu Zheng [14], Tong Liu [7*], Chunhui Yuan [1,2,3*], Tiannan Guo [1,2,3*]

1, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China.

2, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China

3, Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, Zhejiang, China

4, Westlake Omics (Hangzhou) Biotechnology Co.,Ltd. No.1, Yunmeng Road, Cloud Town, Xihu District, Hangzhou 310024, Zhejiang Province, China

5, Department of Pathology, The First Affiliated Hospital of College of Medicine, Zhejiang University, Hangzhou, Zhejiang Province, China

6, Center for Stem Cell Research and Application, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei Province, China

7, Harbin Medical University Cancer Hospital, Harbin 150081, China

8, Department of Pathology, The Second Affiliated Hospital of College of Medicine, Zhejiang University, Hangzhou 310009, Zhejiang Province, China

9, Department of Orthopaedics, The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou 310009, Zhejiang Province, China

10, Key Laboratory of Motor System Disease Research and Precision Therapy of Zhejiang Province, Hangzhou 310020, Zhejiang Province, China

11, The Cancer Hospital of the University of Chinese Academy of Sciences, Zhejiang Cancer Hospital, Hangzhou 310022, Zhejiang Province, China

12, Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou 310022, Zhejiang Province, China

13, Department of Clinical Epidemiology, Shengjing Hospital of China Medical University, Shenyang Province, China

14, Cancer Institute, Key Laboratory of Cancer Prevention and Intervention, Ministry of Education, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China

#, co-first authors

*, corresponding author. Tong Liu: liutong@hrbmu.edu.cn; Chunhui Yuan: yuanchunhui@westlake.edu.cn; Tiannan Guo: guotiannan@westlake.edu.cn

Running title: Optimization of Spectral Library Size for DIA

Keywords: Data-independent acquisition; protein identification; Pan-human library; Spectral library optimization.

Abstract

Efficient peptide and protein identification from data-independent acquisition mass spectrometric (DIA-MS) data typically rely on an experiment-specific spectral library with a suitable size. Here, we report a computational strategy for optimizing the spectral library for a specific DIA dataset based on a comprehensive spectral library, which is accomplished by *a priori* analysis of the DIA dataset. This strategy achieved up to 44.7% increase in peptide identification and 38.1% increase in protein identification in the test dataset of six colorectal tumor samples compared with the comprehensive pan-human library strategy. We further applied this strategy to 389 carcinoma samples from 15 tumor datasets and observed up to 39.2% increase in peptide identification and 19.0% increase in protein identification. In summary, we present a computational strategy for spectral library size optimization to achieve deeper proteome coverage of DIA-MS data.

Introduction

Data-independent acquisition mass spectrometry (DIA-MS) based proteomics coupled with targeted data analysis is playing an increasing role in biomedical studies (1), owing to its high degree of reproducibility, quantitative accuracy, and high throughput (2, 3). Both spectral library-free and library-based strategies are being applied to analyze DIA-MS data (4). While the library-free strategies (5, 6) could identify peptides directly from DIA-MS itself without the requirement of an external spectral library, the depth of proteomic coverage is limited at the moment (7-9). The more widely adopted strategy is based on building a spectral library using the corresponding data-dependent acquisition mass spectrometry (DDA-MS) datasets of the samples of interest (10), or a pre-built library from public data repositories (11-14).

The size of the spectral library has a direct impact on the performance of DIA-MS data analysis (15). A larger number of DDA-MS runs, particularly from fractionated samples, leads to a more comprehensive spectral library enabling potential detection of a larger number of peptides and proteins from the DIA-MS datasets (15). However, it also generates a larger search space and reduces the statistical power to detect true positives (16, 17). Extra concerns are raised where the proteins and peptides within the library may not be specific to a particular specimen, potentially introducing more false positives (18). Other drawbacks include the prolonged computational time which is approximately linearly correlated with the size of the library (19), and distortion of retention time (RT) distribution for alignment (20).

The spectral library size could be optimized to improve DIA-MS performance. The Van Eyk group have reported that applying a comprehensive fractionated library led to higher number of protein/peptide identifications from DIA-MS datasets than un-fractionated libraries with limited sizes (15). Similar results have been reported by Uszkoreit group, where they found larger library led to higher peptide and protein identification but the increase was minimal when the library is comprehensive enough (21). The combination of an in-house built library with external libraries from public data improves DIA data analysis performance (17). Inclusion of internal library extracted from DIA files also improved peptide and protein identification (9). On the other hand, it has also been observed that libraries of very large size led to higher FDRs in the DIA-MS analyses and hence compromises the identification results (17). It was further demonstrated that, even within the same spectral library, controlling the confidence of peptide identifications to exclude redundant peptides could improve peptide and protein identification results (16). Although these studies have repetively reported the importance of the size of spectral library size, a systematic evaluation and optimization of library size is still lacking.

Here, we propose a two-step strategy called subLib to generate the experiment-specific subset libraries using *a priori* analysis of the DIA data to improve the proteomic coverage. The strategy to derive a subset library of optimal size was further applied to analyze the DIA data of 15 human tumors.

Materials and Methods

Colorectal cancer dataset

To evaluate our strategy, the DIA-MS datasets were collected from a colorectal cancer proteomic project in our group (Xiang *et al.*, manuscript in preparation). Briefly, 286 FFPE samples from 44 colorectal cancer patients were processed into peptides with a pressure cycling technology (PCT)-based protocol as described in the previous study (22). They were subjected to data acquisition on the nanoflow EASY-nLC™ 1200 System coupled with Q Exactive HF hybrid Quadrupole-Orbitrap in DIA mode over a gradient of 60 min using 24 DIA windows spanning from 400 Da to 1200 Da.

Fifteen datasets of multiple tumor types

A total of 389 tumor tissue samples from 15 tumor types were collected. The gastric carcinoma (n=30) and thyroid carcinoma (n=30) samples were collected from the First Affiliated Hospital College of Medicine, Zhejiang University. The prostate carcinoma (n=30) and bone carcinoma (n=30) samples were collected from the Second Affiliated Hospital College of

Medicine, Zhejiang University. The liver carcinoma (n=33) and leukemia (n=27) samples were collected from Wuhan Union Hospital. The ovarian carcinoma (n=30) samples were collected from Zhejiang Cancer Hospital. The cervical carcinoma (n=28) samples were collected from Shengjing Hospital of China Medical University. The lung adenocarcinoma (n=32) , gallbladder carcinoma (n=20), pancreatic adenocarcinoma (n=20), myosarcoma (n=19), clear cell renal cell carcinoma (CCRCC, n=20), diffuse large B-cell lymphoma (DLBCL, n=19), and papillary thyroid cancer (PTC, n=21) were collected from Harbin Medical University Cancer Hospital. All samples were approved by the ethics committees of their respective hospitals. The tissue samples were prepared with PCT-based tissue lysis and protein digestion protocol (22) and analyzed by DIA-MS, as listed in Table S1. Ethics approvals for this study were obtained from the Ethics Committee or Institutional Review.

Proteomic data analysis workflow

The raw DIA-MS data files were converted to mzXML format using the msConvert tool in ProteomeWizard (23). The DIA-MS datasets were analyzed using the open-source software OpenSWATH (version 2.4.0) (24) with the following criteria: common internal reference peptides (CiRTs) of each tissue were applied respectively for retention time alignment; m/z extraction window was set to 30 ppm, and RT extraction window was set between 200-800 seconds, depending on different gradients of the DIA-MS module (Table S1). PyProphet (version 2.1.3) (24) was used for statistical validation via setting the global cutoff of FDR as 0.01 at both peptide and protein levels. Protein inference was performed as described previously (25). Unless otherwise mentioned, the software parameters were kept the same for all the analyses in this study.

Subset library generation

We proposed a two-step strategy to take a subset of the spectral library. Firstly, the public library is taken to analyze the candidate DIA-MS dataset using the OpenSWATH workflow. Different FDR cutoffs were set to generate a list of identification results. Afterwards, they were matched against the public library to generate experiment-specific subset libraries.

In this study, we set the DIA Pan-Human Library (DPHL) (12) as the baseline library to analyze the colorectal cancer dataset containing 284 DIA-MS data files. FDR cutoffs were set at 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6 (n=15), to generate 15 identification results. After matching with DPHL, OpenSwathDecoyGenerator.exe in OpenMS (version2.4.0) was applied to generate equal amount of decoys in mutated fashion. The resultant subset library is a combination of DPHL subsets and decoys.

Results and Discussions

Generation of the subset library by refining DPHL

For data comprehensiveness and accessibility, DPHL built from 16 human tissue types containing 359,627 peptide precursors and 14,782 protein groups was used as the baseline spectral library. A DIA-MS dataset of 286 colorectal cancer sample cohort was analyzed to derive the initial identifications. We set the FDR cut-off for peptide precursor and protein identification to 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6 (a total of 15 tests), then retrieved the resultant subset libraries at each FDR cutoff. The four representative DIA-MS data files (sample A1-A4) within the cohort and two external colorectal cancer DIA-MS data files (sample B1 and B2) were taken to evaluate the identification performance of each subset library (Figure 1A). The number of identified peptides shows a generally decreasing trend as the FDR cutoff increases (Figure 1C), with the exceptions when FDR increases from 0.01 to 0,02, and from 0.04 to 0.05. The number of identified proteins increased as the FDR cut-off increased from 0.01 to 0.05, and gradually decreased afterward, with a drastic decline when the cutoff was beyond 0.1 (Figure 1D). This is not unexpected since the peptides identified with high FDR are more likely absent in the sample at the detection limit. As the library size increased, the negative effect prevailed. The best result was obtained from the library with a FDR cutoff of 0.05. The optimal library was composed of 85,655 peptide precursors, 62,390 peptides, and 6,448 protein groups, leading to the identification of 29,979 peptide precursors and 4,418 protein groups, respectively. This optimized library led to 44.7% and 38.1% increase of peptide precursors and protein groups, respectively, compared with the results by the unfiltered DPHL (Figure S1). The subset library with the FDR cutoff of 0.05 was the best subset library which was hence adopted for further evaluation. The DIA files used for library size optimization from samples A1-A4 led to similar data to those from independent samples (B1 and B2), suggesting that the library size optimization is generic and applicable to DIA files of the same tissue type.

Adding unidentified peptide procursors to the subset library sacrificed identification

To check if unidentified peptide precursors in a spectral library would affect the DIA-MS proteome coverage, we randomly generated nine sets of DPHL peptides that were excluded from the subset library (defined as “unidentified peptides”), with precursor number equivalent to n% of the subset library (n=10, 20, …, 90), and combined them with the subset library peptides (Figure 1B). When applying the reconstructed spectral libraries to analyze the test DIA dataset, a steady decrease of identified peptides and proteins was observed as more unidentified precursors were included (Figure 1E, F), with the highest proteome coverage coming from the library with no unidentified peptides, summing up to 29,712 peptide precursors and 4,433 protein groups.

We also replaced the unidentified peptides to *in silico* generated decoy peptides and repeated the above analyses. Peptide/protein identifications decreased as the computational peptide proportions increase from 0% to 60%. Further addition of decoys would, however, subtly increase protein identifications (Figure 1G, H). The highest proteome coverage came from the library with no decoy interferences, summing up to 19,322 peptide precursors and 3,461 protein groups. We hence concluded that any false positive interference in the library would suppress the peptide/protein identification.

Adding subset library peptides to interferences improves identification

We then conducted a backward analysis by adding increasing proportions of subset library peptides to the unidentified peptides (Figure 1B). The spectral library composed by precursors of unidentified peptides solely (n=0) could not identify any peptide or protein in the DIA-MS data. The numbers of identification of peptides and proteins exhibited almost marked increase as n increased (Figure 1I, K). Together with the above results, they validated the effectiveness of setting FDR cutoff as 0.05 to eliminate false positive targets.

Applying subLib to DIA-MS of 15 tumor sample types

We named the library generation strategy “subLib” and further applied it to the fifteen DIA datasets of different types of cancer samples, including bone, cervical, DLBCL, gallbladder, gastric, leukemia, liver, lung, myosarcoma, ovarian, pancreatic, prostate, PTC, and CCRCC (Figure 2A). Peptide/protein identifications using the subset library exceeded that from using DPHL in most cases (Figure 3A), and over 99% of the protein identifications were overlapped in every cancer type (Figure S2). We collectively found that the subLib strategy outperformed the DPHL strategy in all cancer types, with the most prominent increase from PTC carcinoma samples (19.02% increase in protein groups and 36.17% increase in peptide precursors, Figure 2B). Of note, the discrimination ability to separate the targets from decoys led to a marked increase (Figure 2C), further validating that the subLib strategy can reduce false positives in clinical proteomic data. Missing values were equivalent between DPHL and the subLib strategy (Figure 3B), and the protein quantification results were in good accordance as well with Pearson correlation ratios all over 0.92 across all the tumor tissue types (Figure 3B), suggesting that decreasing library sizes by adjusting FDR values does not impair protein identification nor quantification. Moreover, different tumor types could be well resolved using the thus generated protein matrix (Figure 2D). These results indicate that this subLib strategy could be generically used for DIA data generated from different samples.

Concluding remarks

In this study, we present a computational strategy to optimize library size for DIA data analysis. In our DIA data of human tissue specimens, setting FDR to 0.05 enabled effective spectral library subsetting. The application of this strategy to DIA data from 15 tumor types further consoidated this conclusion. This subLib strategy reduced false positive identifications, increased peptide and protein identifications, and generated protein data matrix quantitatively comparable to the DIA analysis with unfiltered library. In conclusion, the subLib strategy for DIA spectral library size optimization boosts proteome identifications of DIA-MS data.

Acknowledgements

This work is supported by grants from National Key R&D Program of China (No. 2020YFE0202200), the National Natural Science Foundation of China (81972492, 21904107), Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars (LR19C050001), Hangzhou Agriculture and Society Advancement Program (20190101A04), and Westlake Education Foundation. We thank Westlake University Supercomputer Center for assistance in data storage and computation.

Conflict of interest statement

The research group of Tiannan Guo is partly supported by Pressure Biosciences Inc, which provided access to advanced sample preparation instrumentation. T.G is shareholder of Westlake Omics Inc. W.G. is employee of Westlake Omics Inc. The remaining authors declare no competing interests.

Data Availability

The raw data and peptide/protein matrixes were deposited in ProteomeXchange Consortium (https://www.iprox.org/). Project ID: IPX0002439000 and IPX0001981000.

References

1. Yue, L., Zhang, F., Sun, R., Sun, Y., Yuan, C., Zhu, Y., and Guo, T. (2020) Generating Proteomic Big Data for Precision Medicine. *PROTEOMICS* n/a, 1900358
2. Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (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
3. Guo, T., Kouvonen, P., Koh, C. C., Gillet, L. C., Wolski, W. E., Rost, H. L., Rosenberger, G., Collins, B. C., Blum, L. C., Gillessen, S., Joerger, M., Jochum, W., and Aebersold, R. (2015) Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat Med* 21, 407-413

4. Zhang, F., Ge, W., Ruan, G., Cai, X., and Guo, T. (2020) Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. *PROTEOMICS* 20, 1900276
5. Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C., and Nesvizhskii, A. I. (2015) DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* 12, 258-+
6. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., and Ralser, M. (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 17, 41-+
7. Muntel, J., Gandhi, T., Verbeke, L., Bernhardt, O. M., Treiber, T., Bruderer, R., and Reiter, L. (2019) Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Mol Omics* 15, 348-360
8. Cho, K. C., Clark, D. J., Schnaubelt, M., Teo, G. C., Leprevost, F. D. V., Bocik, W., Boja, E. S., Hiltke, T., Nesvizhskii, A. I., and Zhang, H. (2020) Deep Proteomics Using Two Dimensional Data Independent Acquisition Mass Spectrometry. *Anal Chem* 92, 4217-4225
9. Zhong, C. Q., Wu, R., Chen, X., Wu, S., Shuai, J., and Han, J. (2020) Systematic Assessment of the Effect of Internal Library in Targeted Analysis of SWATH-MS. *J Proteome Res* 19, 477-492
10. Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., Lam, H., Amodei, D., Mallick, P., MacLean, B., and Aebersold, R. (2015) Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols* 10, 426-441
11. Rosenberger, G., Koh, C. C., Guo, T., Rost, H. L., Kouvonen, P., Collins, B. C., Heusel, M., Liu, Y., Caron, E., Vichalkovski, A., Faini, M., Schubert, O. T., Faridi, P., Ebhardt, H. A., Matondo, M., Lam, H., Bader, S. L., Campbell, D. S., Deutsch, E. W., Moritz, R. L., Tate, S., and Aebersold, R. (2014) A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci Data* 1, 140031
12. Zhu, T., Zhu, Y., Xuan, Y., Gao, H., Cai, X., Piersma, S. R., Pham, T. V., Schelfhorst, T., Haas, R. R. G. D., Bijnsdorp, I. V., Sun, R., Yue, L., Ruan, G., Zhang, Q., Hu, M., Zhou, Y., Van Houdt, W. J., Lelarge, T. Y. S., Cloos, J., Wojtuszkiewicz, A., Koppers-Lalic, D., Böttger, F., Scheepbouwer, C., Brakenhoff, R. H., van Leenders, G. J. L. H., Ijzermans, J. N. M., Martens, J. W. M., Steenbergen, R. D. M., Grieken, N. C., Selvarajan, S., Mantoo, S., Lee, S. S., Yeow, S. J. Y., Alkaff, S. M. F., Xiang, N., Sun, Y., Yi, X., Dai, S., Liu, W., Lu, T., Wu, Z., Liang, X., Wang, M., Shao, Y., Zheng, X., Xu, K., Yang, Q., Meng, Y., Lu, C., Zhu, J., Zheng, J. e., Wang, B., Lou, S., Dai, Y., Xu, C., Yu, C., Ying, H., Lim, T. K., Wu, J., Gao, X., Luan, Z., Teng, X., Wu, P., Huang, S. a., Tao, Z., Iyer, N. G., Zhou, S., Shao, W., Lam, H., Ma, D., Ji, J., Kon, O. L., Zheng, S., Aebersold, R., Jimenez, C. R., and Guo, T. (2020) DPHL: A DIA Pan-human Protein Mass Spectrometry Library for Robust Biomarker Discovery. *Genomics, Proteomics & Bioinformatics*
13. Blattmann, P., Stutz, V., Lizzo, G., Richard, J., Gut, P., and Aebersold, R. (2019) Generation of a zebrafish SWATH-MS spectral library to quantify 10,000 proteins. *Scientific Data* 6, 190011
14. Zhang, H., Liu, P., Guo, T., Zhao, H., Bensaddek, D., Aebersold, R., and Xiong, L. (2019) Arabidopsis proteome and the mass spectral assay library. *Sci Data* 6, 278
15. Parker, S. J., Venkatraman, V., and Van Eyk, J. E. (2016) Effect of peptide assay library size and composition in targeted data-independent acquisition-MS analyses. *Proteomics* 16, 2221-2237
16. Zi, J., Zhang, S., Zhou, R., Zhou, B., Xu, S., Hou, G., Tan, F., Wen, B., Wang, Q., Lin, L., and Liu, S. (2014) Expansion of the ion library for mining SWATH-MS data through fractionation proteomics. *Anal Chem* 86, 7242-7246
17. Wu, J. X., Song, X., Pascovici, D., Zaw, T., Care, N., Krisp, C., and Molloy, M. P. (2016) SWATH Mass Spectrometry Performance Using Extended Peptide MS/MS Assay Libraries. *Mol Cell Proteomics* 15, 2501-2514
18. Rosenberger, G., Bludau, I., Schmitt, U., Heusel, M., Hunter, C. L., Liu, Y., MacCoss, M. J., MacLean, B. X., Nesvizhskii, A. I., Pedrioli, P. G. A., Reiter, L., Rost, H. L., Tate, S., Ting, Y. S.,

Collins, B. C., and Aebersold, R. (2017) Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat Methods* 14, 921-927

19. Teleman, J., Hauri, S., and Malmstrom, J. (2017) Improvements in Mass Spectrometry Assay Library Generation for Targeted Proteomics. *J Proteome Res* 16, 2384-2392
20. Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M. J., and Rinner, O. (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* 12, 1111-1121
21. Barkovits, K., Pacharra, S., Pfeiffer, K., Steinbach, S., Eisenacher, M., Marcus, K., and Uszkoreit, J. (2020) Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-based Data-independent Acquisition. *Molecular & Cellular Proteomics* 19, 181-197
22. Zhu, Y., and Guo, T. (2018) High-Throughput Proteomic Analysis of Fresh-Frozen Biopsy Tissue Samples Using Pressure Cycling Technology Coupled with SWATH Mass Spectrometry. In: Sarwal, M. M., and Sigdel, T. K., eds. *Tissue Proteomics: Methods and Protocols*, pp. 279-287, Springer New York, New York, NY
23. Adusumilli, R., and Mallick, P. (2017) Data Conversion with ProteoWizard msConvert. *Methods Mol Biol* 1550, 339-368
24. Rost, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmstrom, J., Malmstrom, L., and Aebersold, R. (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* 32, 219-223
25. Gao, H., Zhang, F., Liang, S., Zhang, Q., Lyu, M., Qian, L., Liu, W., Ge, W., Chen, C., Yi, X., Zhu, J., Lu, C., Sun, P., Liu, K., Zhu, Y., and Guo, T. (2020) Accelerated Lysis and Proteolytic Digestion of Biopsy-level Fresh Frozen and FFPE Tissue Samples Using Pressure Cycling Technology. *J Proteome Res*

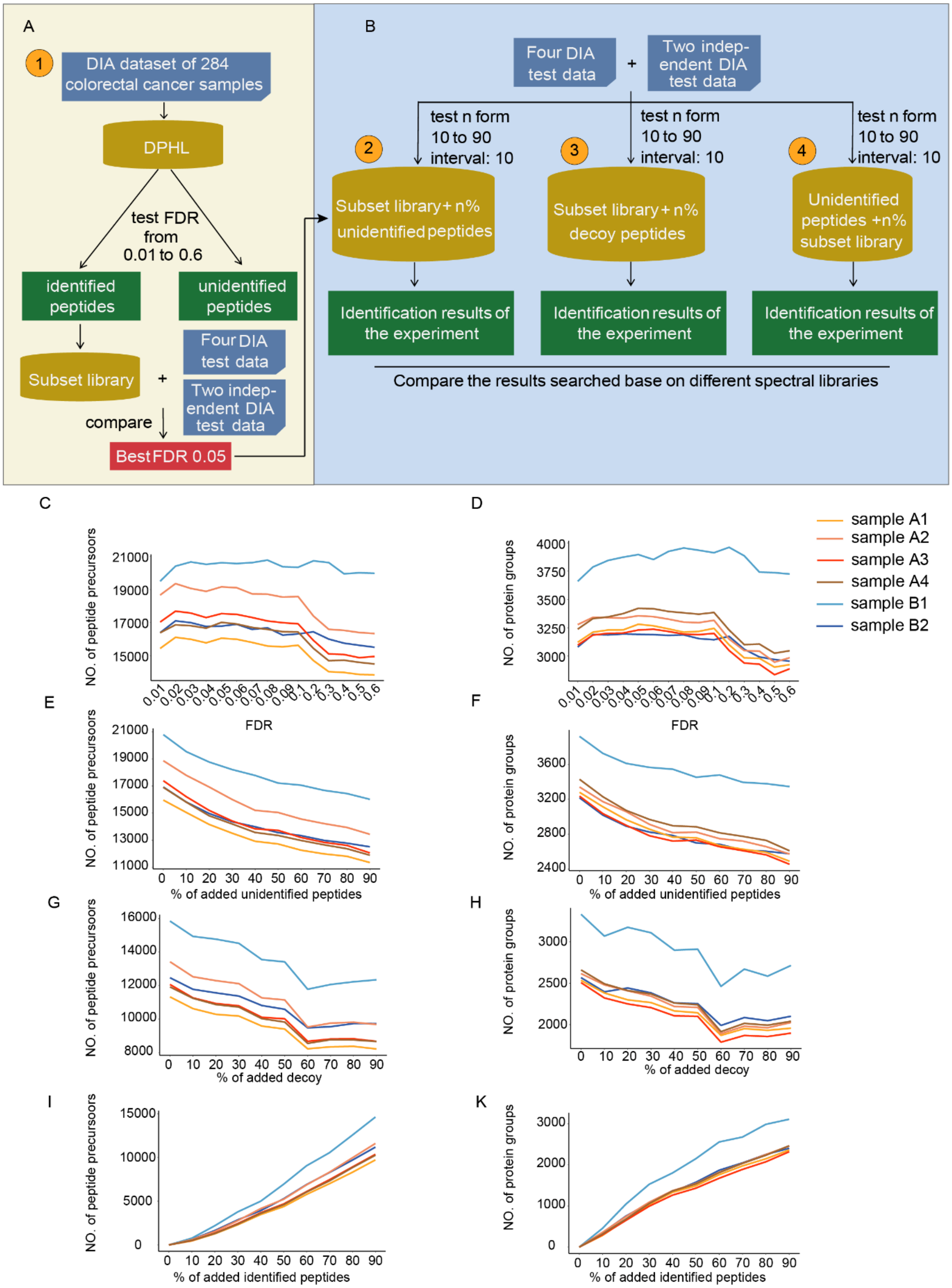

A
1
DIA dataset of 284 colorectal cancer samples
DPHL
test FDR from 0.01 to 0.6
identified peptides
unidentified peptides
Subset library
+
Four DIA test data
Two indep-endent DIA test data
compare
Best FDR 0.05
B
Four DIA test data
+
Two indep-endent DIA test data
test n form 10 to 90 interval: 10
test n form 10 to 90 interval: 10
test n form 10 to 90 interval: 10
2
Subset library+ n% unidentified peptides
3
Subset library + n% decoy peptides
4
Unidentified peptides +n% subset library
Identification results of the experiment
Identification results of the experiment
Identification results of the experiment
Compare the results searched base on different spectral libraries
C
NO. of peptide precursoors
21000
19000
17000
15000
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 0.2 0.3 0.4 0.5 0.6
FDR
D
NO. of protein groups
4000
3750
3500
3250
3000
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 0.2 0.3 0.4 0.5 0.6
FDR
sample A1
sample A2
sample A3
sample A4
sample B1
sample B2
E
NO. of peptide precursoors
21000
19000
17000
15000
13000
11000
0 10 20 30 40 50 60 70 80 90
% of added unidentified peptides
F
NO. of protein groups
3600
3200
2800
2400
0 10 20 30 40 50 60 70 80 90
% of added unidentified peptides
G
NO. of peptide precursoors
16000
14000
12000
10000
8000
0 10 20 30 40 50 60 70 80 90
% of added decoy
H
NO. of protein groups
3000
2500
2000
0 10 20 30 40 50 60 70 80 90
% of added decoy
I
NO. of peptide precursoors
15000
10000
5000
0
0 10 20 30 40 50 60 70 80 90
% of added identified peptides
K
NO. of protein groups
3000
2000
1000
0
0 10 20 30 40 50 60 70 80 90
% of added identified peptides

Figure 1. Optimizing DPHL in the DIA dataset of colorectal cancer. (A) The workflow of spectral library optimization. Step 1: Select the best FDR for refining the subset library from the public DPHL library. The subset library refined from DPHL with FDR of 0.05 is considered as the optimal subset library, which was used as a primary optimized subset spectral library in this study. Step 2: Evaluate the performance of the spectral library consisting of the subset library and n% unidentified peptides. Step 3: Evaluate the performance of spectral library consisting of subset library and n% decoy peptides. Step 4: Evaluate the performance of spectral library consisting of unidentified peptides and n% peptides from the subset library. By comparing all the identification results, the subset library refined from DPHL with FDR of 0.05 is the best experiment-specific spectral library for DIA data analysis. The numbers of identified peptides (C) and proteins (D) based on the subset libraries which were refined from DPHL at nine different FDRs. The numbers of identified peptides (E) and proteins (F) based on the spectral libraries consisting of subset library and n% unidentified peptides. The numbers of identified peptides (G) and proteins (H) based on the spectral libraries consisting of the subset library and n% decoy peptides. The numbers of identified peptides (I) and proteins (K) based on the spectral libraries consisting of unidentified peptides and n% peptides from the subset library.

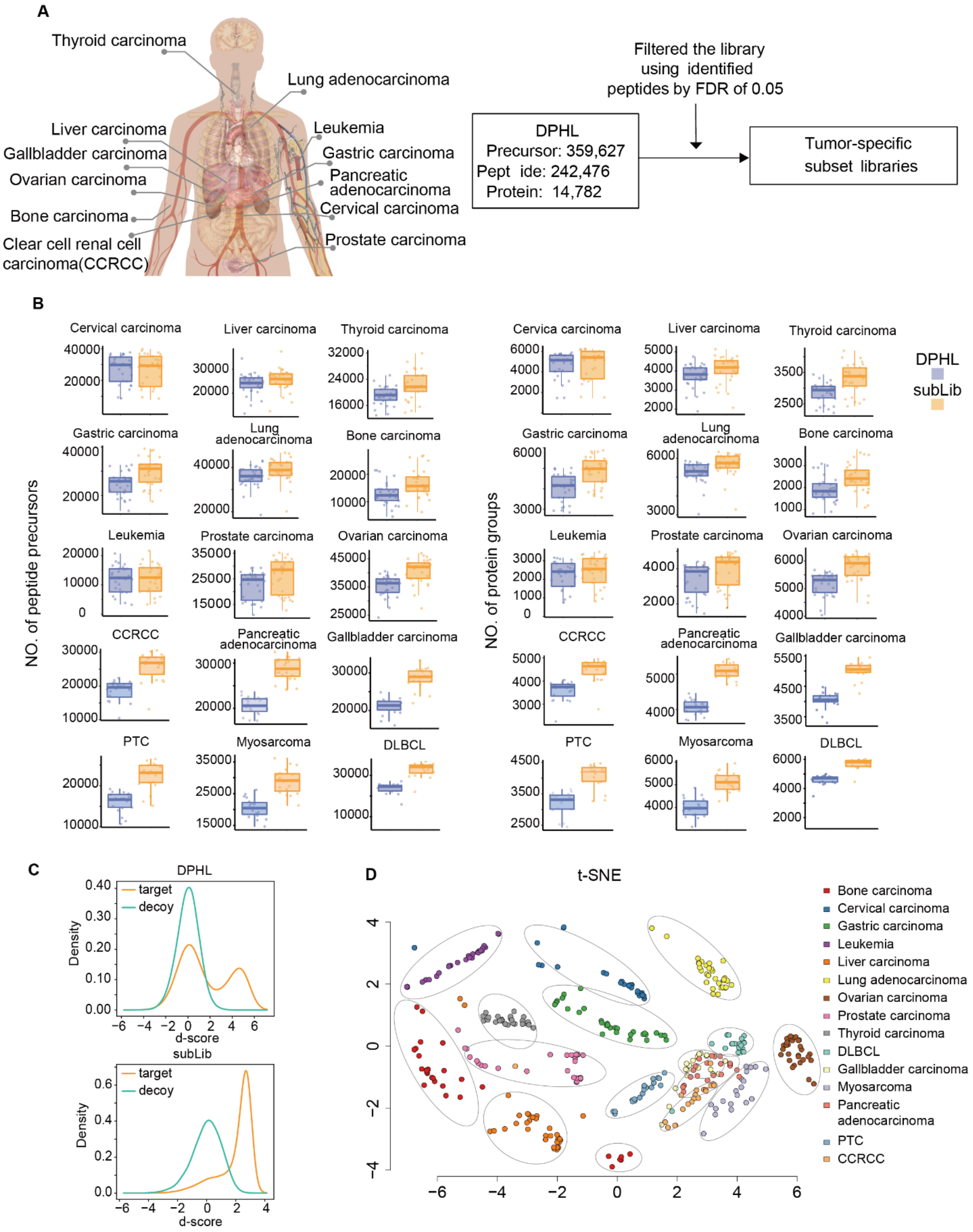
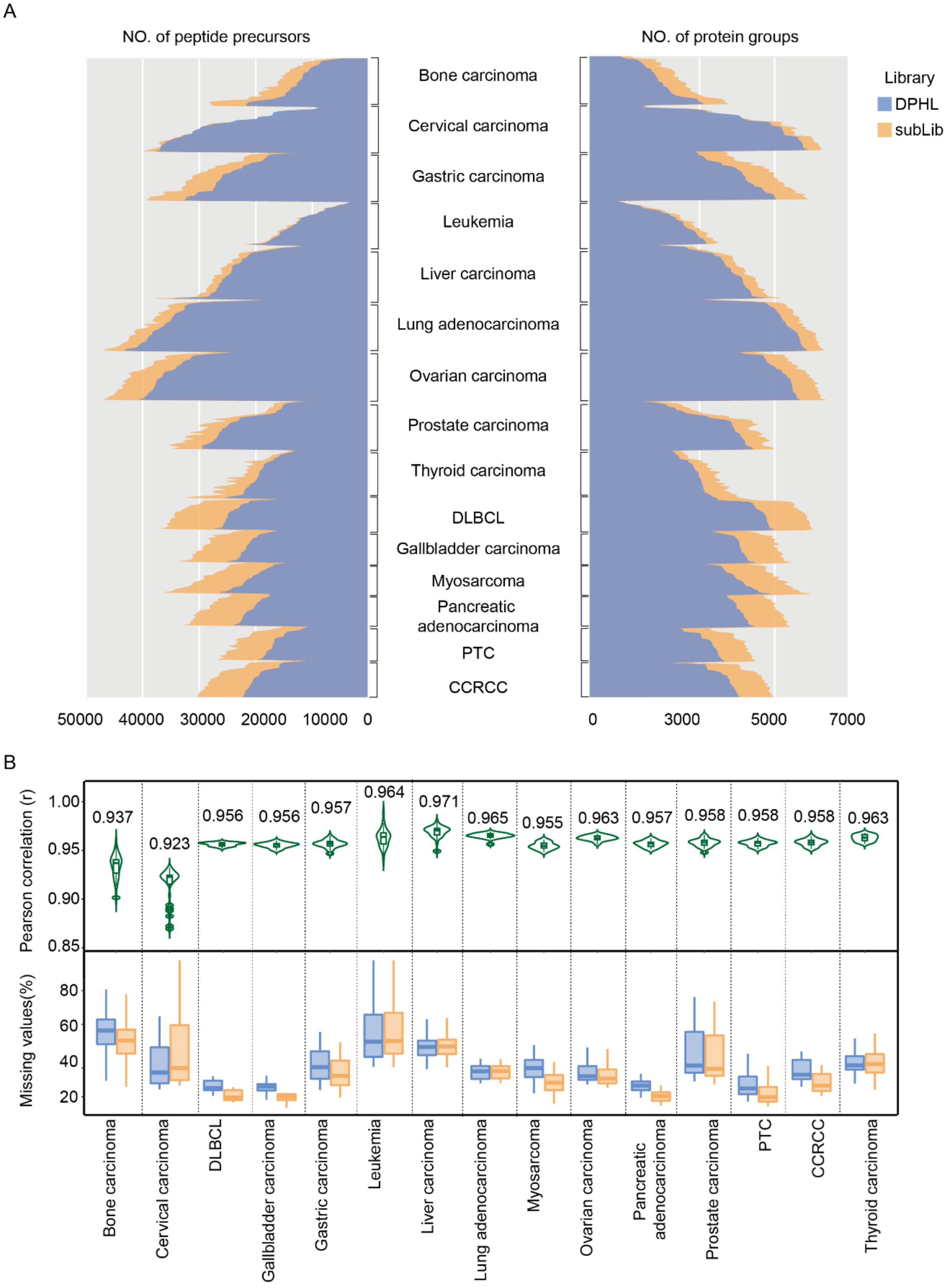

A
Thyroid carcinoma
Lung adenocarcinoma
Liver carcinoma
Gallbladder carcinoma
Ovarian carcinoma
Bone carcinoma
Clear cell renal cell carcinoma(CCRCC)
Leukemia
Gastric carcinoma
Pancreatic adenocarcinoma
Cervical carcinoma
Prostate carcinoma
DPHL
Precursor: 359,627
Pept ide: 242,476
Protein: 14,782
Filtered the library using identified peptides by FDR of 0.05
Tumor-specific subset libraries
B
NO. of peptide precursors
Cervical carcinoma
40000
20000
Liver carcinoma
30000
20000
Thyroid carcinoma
32000
24000
16000
Gastric carcinoma
40000
20000
Lung adenocarcinoma
40000
20000
Bone carcinoma
20000
10000
Leukemia
20000
10000
0
Prostate carcinoma
35000
25000
15000
Ovarian carcinoma
45000
35000
25000
CCRCC
30000
20000
10000
Pancreatic adenocarcinoma
30000
20000
Gallbladder carcinoma
30000
20000
PTC
20000
10000
Myosarcoma
35000
25000
15000
DLBCL
30000
10000
NO. of protein groups
Cervica carcinoma
6000
4000
2000
Liver carcinoma
5000
4000
3000
2000
Thyroid carcinoma
3500
2500
Gastric carcinoma
6000
3000
Lung adenocarcinoma
6000
3000
Bone carcinoma
3000
2000
1000
Leukemia
3000
2000
1000
0
Prostate carcinoma
4000
2000
Ovarian carcinoma
6000
5000
4000
CCRCC
5000
4000
3000
Pancreatic adenocarcinoma
5000
4000
Gallbladder carcinoma
5500
4500
3500
PTC
4500
3500
2500
Myosarcoma
6000
5000
4000
DLBCL
6000
4000
2000
DPHL
subLib
C
DPHL
target
decoy
Density
0.40
0.30
0.20
0.10
0.00
−6 −4 −2 0 2 4 6
d-score
subLib
target
decoy
Density
0.6
0.4
0.2
0.0
−6 −4 −2 0 2 4
d-score
D
t-SNE
4
2
0
−2
−4
−6 −4 −2 0 2 4 6
Bone carcinoma
Cervical carcinoma
Gastric carcinoma
Leukemia
Liver carcinoma
Lung adenocarcinoma
Ovarian carcinoma
Prostate carcinoma
Thyroid carcinoma
DLBCL
Gallbladder carcinoma
Myosarcoma
Pancreatic adenocarcinoma
PTC
CCRCC

Figure 2. Tumor-specific subset library improves the identifications compared with DPHL. (A) The workflow of the subLib strategy. (B) The number of peptides and proteins identified base on tumor-specific subLib and DPHL in 15 tumor types. (C) The distribution of discrimination score (d-score) of the target and decoy of the subset library and DPHL. (D) The tSNE plot shows the samples are well resolved by tissue type.


A
NO. of peptide precursors
NO. of protein groups
Bone carcinoma
Cervical carcinoma
Gastric carcinoma
Leukemia
Liver carcinoma
Lung adenocarcinoma
Ovarian carcinoma
Prostate carcinoma
Thyroid carcinoma
DLBCL
Gallbladder carcinoma
Myosarcoma
Pancreatic adenocarcinoma
PTC
CCRCC
Library
DPHL
subLib
50000 40000 30000 20000 10000 0
0 3000 5000 7000
B
Pearson correlation (r)
1.00
0.95
0.90
0.85
0.937
0.923
0.956
0.956
0.957
0.964
0.971
0.965
0.955
0.963
0.957
0.958
0.958
0.958
0.963
Missing values(%)
80
60
40
20
Bone carcinoma
Cervical carcinoma
DLBCL
Gallbladder carcinoma
Gastric carcinoma
Leukemia
Liver carcinoma
Lung adenocarcinoma
Myosarcoma
Ovarian carcinoma
Pancreatic adenocarcinoma
Prostate carcinoma
PTC
CCRCC
Thyroid carcinoma

344

Figure 3. Peptide precursor and protein identification using the optimized subset library and DPHL. (A) The number of peptide precursors and protein groups identified using the optimized subset library and DPHL for each sample of every tumor type. subLib, the optimized subset library. Protein identifications were shown on the right, and peptide precursor identifications were shown on the left. (B) The correlation values on the protein level between identification results of the optimized subset library and DPHL. The percentages of protein missing values identified base on DPHL and the optimized library of each tumor type.

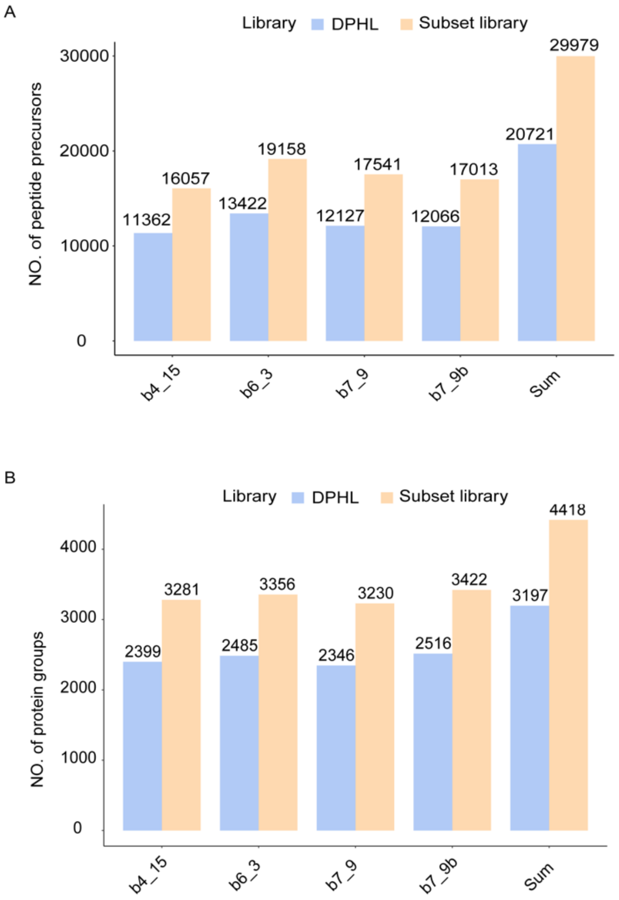


Figure S1. Identification results of the four representative DIA-MS data in the colorectal cancer cohort.

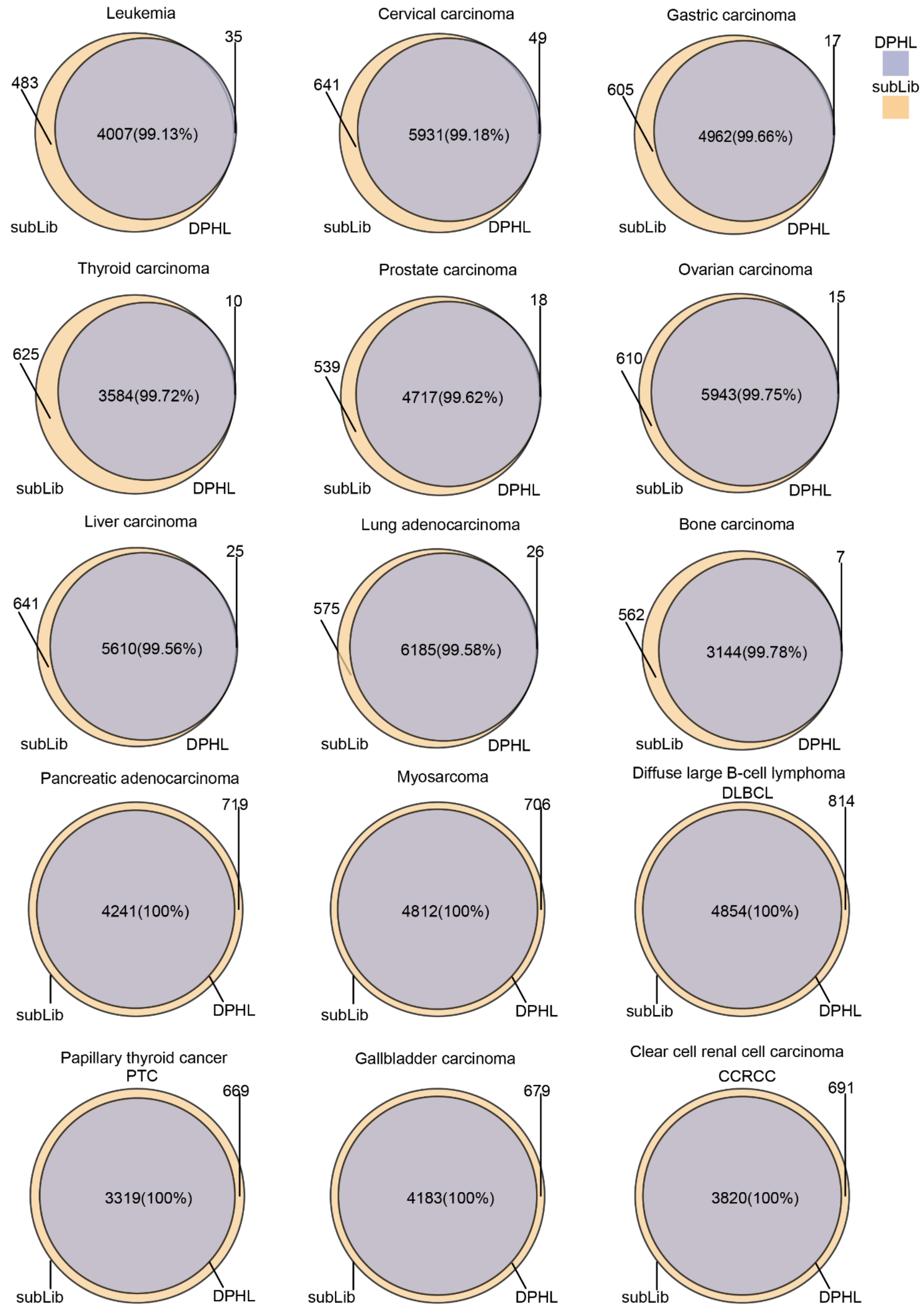

Leukemia
35
483
4007(99.13%)
subLib
DPHL
Cervical carcinoma
49
641
5931(99.18%)
subLib
DPHL
Gastric carcinoma
17
605
4962(99.66%)
subLib
DPHL
DPHL
subLib
Thyroid carcinoma
10
625
3584(99.72%)
subLib
DPHL
Prostate carcinoma
18
539
4717(99.62%)
subLib
DPHL
Ovarian carcinoma
15
610
5943(99.75%)
subLib
DPHL
Liver carcinoma
25
641
5610(99.56%)
subLib
DPHL
Lung adenocarcinoma
26
575
6185(99.58%)
subLib
DPHL
Bone carcinoma
7
562
3144(99.78%)
subLib
DPHL
Pancreatic adenocarcinoma
719
4241(100%)
subLib
DPHL
Myosarcoma
706
4812(100%)
subLib
DPHL
Diffuse large B-cell lymphoma
DLBCL
814
4854(100%)
subLib
DPHL
Papillary thyroid cancer
PTC
669
3319(100%)
subLib
DPHL
Gallbladder carcinoma
679
4183(100%)
subLib
DPHL
Clear cell renal cell carcinoma
CCRCC
691
3820(100%)
subLib
DPHL

Figure S2. Venn diagrams showing overlap of protein identifications between the optimized subset library and DPHL.