Generating Proteomic Big Data for Precision Medicine
为精准医疗产生蛋白质组学大数据

Liang Yue
岳靓

蛋白质组大数据实验室
www.guomics.com

西湖大學
WESTLAKE UNIVERSITY

Outlines

Abstract

1. Proteomics for Precision Medicine
2. Emerging Large-Scale Clinical Proteomic Data Sets Using Data-Dependent Acquisition and Data-Independent Acquisition
3. DDA or DIA for Generating Proteomic Big Data?
4. High-Throughput Sample Preparation for Proteomics
5. Considerations of Data Quality
6. Deep Learning for Proteomic Big Data

Conclusions

1.Proteomics for Precision Medicine

1. Precision medicine is to make clinical decisions regarding diagnosis, prognosis, and treatment customized to the diverse needs of individuals based on their specific phenotypic, molecular, or psychosocial characteristics.[1]
2. Complexity of disease and the therein mechanisms are hard to explained by a fixed shortlist of molecules
3. Omics offers an emerging approach to systematically profile thousands of molecular dysregulations to cover the granularity of disease pathogenesis and progression.
4. Measurement of the proteome of clinical specimens can complement their genomic data in understanding cancer biology and customizing cancer management.[2]
5. The measurement of a proteome, largely dependent on mass spectrometry [3], is inherently much more sophisticated than that of nucleic acids due to its spatiotemporal dynamics.

[1] J. L. Jameson, D. L. Longo, *N. Engl. J. Med.* 2015, **372**, 2229.
[2] H. Rodriguez, S. R. Pennington, *Cell* 2018, **173**, 535.
[3] R. Aebersold, M. Mann, *Nature* 2016, **537**, 347.

2. Emerging Large-Scale Clinical Proteomic Data Sets Using Data-Dependent Acquisition and Data-Independent Acquisition

1. Data-dependent acquisition (DDA) coupled with multi-dimensional fractionation is the most adopted method for proteomics. Analysis of 100–1000 s clinical specimens using DDA-MS becomes feasible in individual laboratories. [1]
2. In the meantime, targeted data analysis strategy, initially developed for targeted proteomics, was introduced to DIA-MS, specifically sequential window acquisition of all theoretical mass spectra (SWATH-MS), which was implemented in the TripleTOF mass spectrometers.[2]
3. DIA-MS produces a permanent digital map containing all the flyable peptide precursors of a specimen, serving as an ideal method for the proteome digitization of clinical specimens.[3]

[1] J. L. Jameson, D. L. Longo, *N. Engl. J. Med.* 2015, **372**, 2229.
[2] R. Aebersold, M. Mann, *Nature* 2016, **537**, 347.

3. DDA or DIA for Generating Proteomic Big Data?

1. Arguably, neither in-depth label-free nor label-based DDA-MS will likely generate consistent quantitative proteomic data from complex tissue samples in large clinical cohorts in high throughput due to the need of multi-dimensional fractionation.
2. Label-free quantification by DIA-MS circumvents the reproducibility problem of data-dependent sampling in DDA-MS.[1]
3. DIA records complete information of fragment ions and combined targeted approaches for more comprehensiveness and consistent acquisition on the MS level.
4. Several studies have shown that with the same LC gradient, DIA enabled deeper depth for proteomic and phosphoproteomic studies.[2,3]
5. In this viewpoint, we anticipate that increasing number of large DIA data sets from clinical cohorts will be generated in the coming years by individual laboratories equipped with high-throughput proteomics technologies and industrialized facilities

[1] K. Barkovits, S. Pacharra, K. Pfeiffer, S. Steinbach, M. Eisenacher, K. Marcus, J. Uszkoreit, Mol. Cell. Proteomics 2020, 19, 181
[2] R. Bruderer, O. M. Bernhardt, T. Gandhi, S. M. Miladinovic, L. Y. Cheng, S. Messner, T. Ehrenberger, V. Zanotelli, Y. Butscheid, C. Escher, O. Vitek, O. Rinner, L. Reiter, Mol. Cell. Proteomics 2015, 14, 1400.
[3] D. B. Bekker-Jensen, O. M. Bernhardt, A. Hogrebe, A. Martinez-Val, L. Verbeke, T. Gandhi, C. D. Kelstrup, L. Reiter, J. V. Olsen, *Nat. Commun.* 2020, **11**, 787.
[4] B. Tully, R. L. Balleine, P. G. Hains, Q. Zhong, R. R. Reddel, P. J. Robinson, Proteomics 2019, 19, 1900109.

4. High-Throughput Sample Preparation for Proteomics

1. High throughput and reproducible extraction and digestion of proteins from clinical specimens are prerequisite for generating high quality proteomic big data.
2. PCT can effectively process FFPE samples to peptides ready for LC-MS shot in about 3 hrs.[1,2]
3. Liquid handling robots can deal with samples in 96-well plates. Mann's group developed a pipeline for plasma proteome profiling, in which the entire sample preparation procedure took less than 2 h in 96-well plates.[3] Lately, they implemented it to process microdissected FFPE samples in high throughput.[4]
4. EasyPhos was developed by Mann's group to enable parallel 96-well processing to simplify and accelerate the phosphoproteomics workflows with optimized coverage and reproducibility of phosphorylation site quantification.[5]

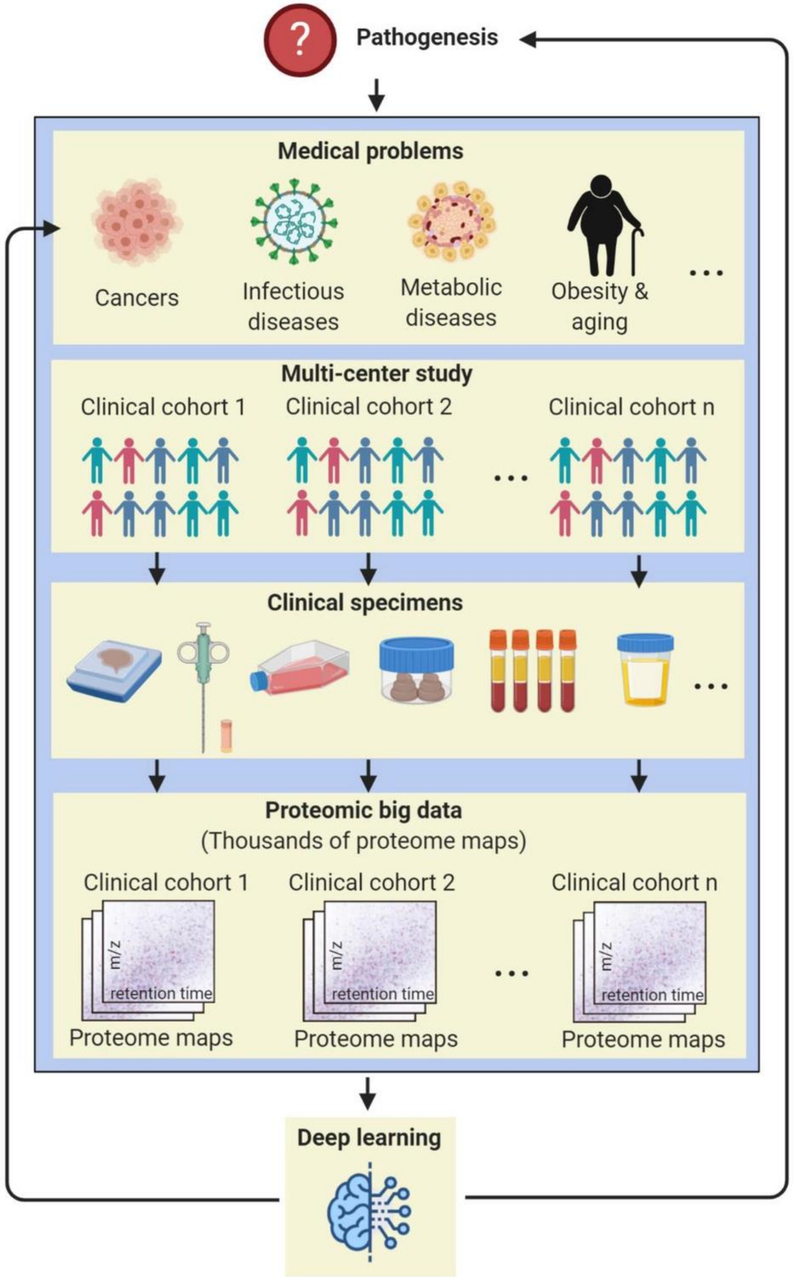
[1] T. Guo, P. Kouvonen, C. C. Koh, L. C. Gillet, W. E. Wolski, H. L. Rost, G. Rosenberger, B. C. Collins, L. C. Blum, S. Gillessen, M. Joerger, W. Jochum, R. Aebersold, Nat. Med. 2015, 21, 407.
[2] H. Gao, F. Zhang, S. Liang, Q. Zhang, M. Lyu, L. Qian, W. Liu, W. Ge, C. Chen, X. Yi, J. Zhu, C. Lu, P. Sun, K. Liu, Y. Zhu, T. Guo, J. Proteome Res. 2020, 19, 1982;
[3] P. E. Geyer, N. A. Kulak, G. Pichler, L. M. Holdt, D. Teupser, M. Mann, *Cell Syst.* 2016, **2**, 185.
[4] F. Coscia, S. Doll, J. M. Bech, A. Mund, E. Lengyel, J. Lindebjerg, G. I. Madsen, J. M. A. Moreira, M. Mann, *bioRxiv:779009*, 2019.
[5] S. J. Humphrey, O. Karayel, D. E. James, M. Mann, *Nat. Protoc.* 2018, **13**, 1897.

5. Considerations of Data Quality

1. In large cohort studies, multiple steps could introduce batch effects which may confound the biological interpretation of the data.[1]
2. The use of standard operating procedures and quality control samples for both sample preparation and MS data acquisition contributed to minimizing batch effects.[2,3]
3. Given properly designed batches, the batch effects can be evaluated and minimized.[1,2] Faster sample preparation and shorter LC gradient with microflow LC reduced batch effects too.[4]

[1] W. W. B. Goh, W. Wang, L. Wong, *Trends Biotechnol.* 2017, **35**, 498.
[2] Y. Sun, S. Selvarajan, Z. Zang, W. Liu, Y. J. Zhu, H. Zhang, H. Chen, X. Cai, H. Gao, Z. Wu, L. Chen, X. Teng, Y. Zhao, S. Mantoo, T. K.-H. Lim, B. Hariraman, S. Yeow, S. M. F. S. Abdillah, S. S. Lee, G. Ruan, Q. Zhang, T. Zhu, W. Wang, G. Wang, J. Xiao, Y. He, Z. Wang, W. Sun, Y. Qin, Q. Xiao, et al., *medRxiv* 2020,
[3] B. Shen, X. Yi, Y. Sun, X. Bi, J. Du, C. Zhang, S. Quan, F. Zhang, R. Sun, L. Qian, W. Ge, W. Liu, S. Liang, H. Chen, Y. Zhang, J. Li, J. Xu, Z. He, B. Chen, J. Wang, H. Yan, Y. Zheng, D. Wang, J. Zhu, Z. Kong, Z. Kang, X. Liang, X. Ding, G. Ruan, N. Xiang, X. Cai, H. Gao, L. Li, S. Li, Q. Xiao, T. Lu, Y. Zhu, H. Liu, H. Chen, T. Guo, Cell 2020, 182, 59.
[4] R. Sun, C. Hunter, C. Chen, W. Ge, N. Morrice, S. Liang, C. Yuan, Q. Zhang, X. Cai, X. Yu, L. Chen, S. Dai, Z. Luan, R. Aebersold, Y. Zhu, T. Guo, *bioRxiv:675348*, 2019.

5. Deep Learning for Proteomic Big Data



Many complex clinical problems have been well assisted by big data and deep learning (DL) technology. For instance, diabetic retinopathy is reported to be diagnosed by a DL model developed by learning of 128 175 retinal images with area under curve (AUC) over 99% in the validation dataset,[1] DL of 207 130 images of the retina from 4686 patients led to the diagnosis of the most common blinding retinal diseases with an AUC of 99.9%.[2]

In theory, a sizeable protein matrix of injections and proteins, if available, could be used for building a predictive model for biomarker discovery

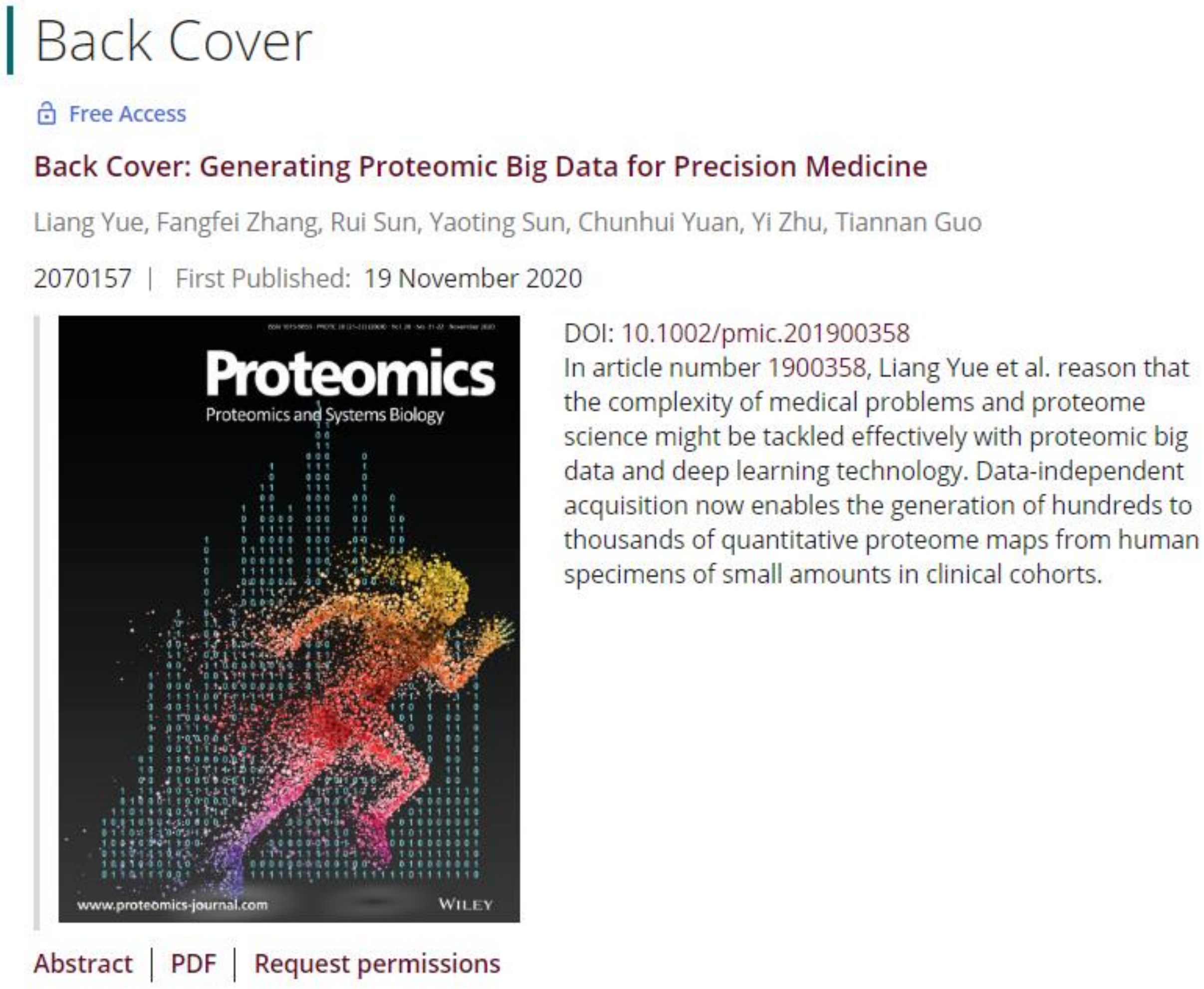
[1] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, D. R. Webster, *J. Am. Med. Assoc.* 2016, **316**, 2402.

[2] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, et al., *Cell* 2018, **172**, 1122.

Take-home messages

1. The complexity of proteomes and clinical problems necessitates the generation of large-scale proteomic data sets from multiple clinical cohorts to understand pathogenesis and to address medical problems.
2. Recent technological advances in sample preparation and LC-MS are enabling generation of proteomic big data at increasing pace. In particular, both DDA and DIA mass spectrometry exhibit increased throughput in analyzing the proteomes, while DIA seems to possess higher degree of reproducibility, throughput, and information content compared to DDA.
3. Increasing volumes of DIA-based proteome maps are being produced, approaching the minimal requirement of DL.
4. We anticipate rapid accumulation of DIA-based proteomic big data from clinical cohorts, which will build up a solid basis for DL-based modeling of diseases.
5. Future efforts are required to address technical obstacles of generating proteomic big data and customize DL

Publication



Back Cover

Free Access

Back Cover: Generating Proteomic Big Data for Precision Medicine

Liang Yue, Fangfei Zhang, Rui Sun, Yaoting Sun, Chunhui Yuan, Yi Zhu, Tiannan Guo

2070157 | First Published: 19 November 2020

DOI: 10.1002/pmic.201900358
In article number 1900358, Liang Yue et al. reason that the complexity of medical problems and proteome science might be tackled effectively with proteomic big data and deep learning technology. Data-independent acquisition now enables the generation of hundreds to thousands of quantitative proteome maps from human specimens of small amounts in clinical cohorts.

Abstract | PDF | Request permissions

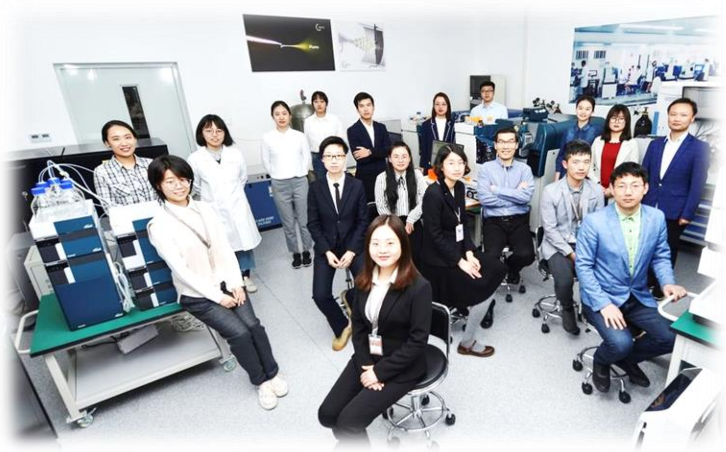
Acknowledgements

西湖大學
WESTLAKE UNIVERSITY

UOMICS -since 2017

西湖歐米
WESTLAKE OMICS

All labmates in Laboratory of Big Proteomic Data



THANK YOU

西湖大學
WESTLAKE UNIVERSITY