Current status of cancer proteogenomics: a brief introduction

Emi Hattori and Tadashi Kondo*

Division of Rare Cancer Research, National Cancer Center Research Institute

(Received January 31, 2019; Accepted April 1, 2019)

SUMMARY

Proteogenomics is a novel approach to understand the molecular backgrounds of diseases. In cancer research, proteomic studies have been conducted without using the genome data of individual samples. For example, a common public database has always been used to identify proteins by mass spectrometry. However, tumor genomes, even tumors of the same type of cancer, can differ considerably, and such differences affect the response to treatments. Thus, genomic backgrounds should be considered when identifying proteins by mass spectrometry. In cancer proteogenomics, a virtual proteome database is generated using the genome data of identical samples for the mass spectrometric identification of proteins reflecting genetic mutations, which are not common and not cited in the commonly used databases. Such proteins are candidate biomarkers and therapeutic targets. Although previous studies have reported software capable of translating genomic data to proteomic data, a standard protocol has not been established. In addition, the utility of proteogenomics has also not been established, and it is not self-evident that proteins with mutations unique to certain groups can be exploited for innovative treatments or to provide clues for the resolution of biological problems in cancers. Collaborative efforts by cancer researchers and specialists in mass spectrometry and bioinformatics are required for fruitful advancements.

Key words: proteogenomics, cancer, biomarker, therapeutic target, bioinformatics

INTRODUCTION

In cancer research, proteomics has long been used to reveal the molecular aspects of diseases. With the advent of proteomics early in this century, novel biomarkers relevant for early diagnosis were reported in ovarian cancer, which highlighted the value of proteome data to researchers who hitherto had no experience in this area. At the same time, the limitations of proteomics became obvious because the existing plasma biomarkers were not identified in the proteomics approaches. Since then, considerable efforts have been devoted to the development of more robust methods.

Mass spectrometry (MS) is a central tool of modern proteomics. Proteins cannot be amplified, and their detection limit depends on the sensitivity of the detection instrument. In this sense, the ultra-sensitive MS became a remedy of traditional protein biochemistry. MS technology has considerably improved the performance of other biochemical methods such as electrophoresis. Electrophoresis has been extensively used as a separation method in cancer proteomics. However, before MS, it was tremendously time consuming to determine the proteins in the bands of SDS-PAGE and the spots of 2D-PAGE. Using MS, the identification of proteins separated by electrophoresis became a routine experiment. As a result, we can understand the biological significance of electrophoresis data based on the sequence information. However, MS alone cannot identify proteins; MS-based protein identification requires a proteome database against which the ion peak data of barely detectable proteins can be searched. The process of protein identification is illustrated in Fig. 1. MS ionizes peptides, and the corresponding peptides are identified from a common proteome database, such as Swiss-Prot. Identical patterns revealed by the informatics search identify the proteins corresponding to the peptides from which the ion peaks are derived. Thus, modern proteomics is based on MS as well as refinements in genome research, high-performance and reduced-cost computers, the infrastructure for database utility, and social factors such as the world-wide internet connection.

Cancers are a major cause of disease-related death in developed countries. The expectation has been that pro-

* Corresponding author: Tadashi Kondo; Division of Rare Cancer Research, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
E-mail: takondo@ncc.go.jp
Tel: +81-3-3542-2511 ext. 3419
teomics can benefit cancer treatments, and a great deal of research funds have been allocated to cancer proteomics. Unfortunately, the expectations of cancer proteomics in terms of innovative medical approaches have not been forthcoming. For example, the initial enthusiasm for the utility of biomarkers in the early diagnosis of ovarian cancers was not borne out with time. Although many papers reported biomarkers discovered by proteomics approaches, none have proven to be clinically useful. It may be that global protein expression profiling alone cannot lead to the discovery of candidate biomarkers and therapeutic targets. During the last 20 years, we experienced both excitement and disappointments as cancer proteomics has matured as a research field. If the future of cancer proteomics is to be bright, a different viewpoint is required.

**INHERENT LIMITATIONS OF TRADITIONAL CANCER PROTEOMICS**

The genomes of cancer cells differ among different cells, including those of the same cancer. Thus, the amino acid sequences of proteins differ among samples. To identify proteins by MS, proteins are enzymatically digested, the generated peptides are ionized, and the ion peak data are matched with the data in a common proteome database. The problem with this approach is that the proteome database used for protein identification does not reflect the variations of samples. Instead, an identical proteome database is always used for all samples (Fig. 2). A common proteome database for protein identification, such as Swiss-Prot, contains the aberrant amino acid sequences widely observed in many types of cancers, or those extensively studied by many researchers. However, to avoid redundancy, reduce false positive identification, and increase the speed of analysis, all possible sequences are not recorded in the proteome database. Thus, the peptides including minor mutations, such as those observed only in certain sample groups or those not widely studied, cannot be detected by MS, even when they are observed as ion peaks.

This is considerably problematic in cancer research. Proteins with mutations unique to certain types of cancers or groups of patients with certain clinical and pathological features can be highly specific biomarkers or therapeutic targets. It is difficult to distinguish the origin of proteins that consist only of normal amino acid sequences. Indeed, when such proteins are detected as biomarkers for diagnosis, the differential diagnosis between cancers and normal tissues is always problematic. Those proteins will be therapeutic targets, which may cause unfavorable side effects, because the target proteins may be expressed by normal cells. Thus, the proteins with unique mutations should be candidate biomarkers and therapeutic targets. Especially, proteins with mutations unique to specific diseases, specific patient groups, and specific patients will be interesting. However, such proteins cannot be detected using traditional cancer proteomics because of the use of the common public database for protein identification without considering the genomic variations.

**PROTEOME TO PROTEOGENOMICS**

Proteogenomics may be a remedy for the tribulations of conventional cancer proteomics. By creating a proteome database individually for samples, we may be able to solve
and gastric cancer have been reported from overseas splicing variants, and post-translational modifications. In one way, considering the transcription site, it is challenging. This is because the proteome cannot be determined in a straightforward manner; obtaining proteome data from genome data remains problematic. How- ever, with the advancements in technology, obtaining proteome data from genome data is becoming increasingly feasible. The cost of proteomic studies is not as high as it once was. However, obtaining proteome data from genome data remains challenging, especially when it comes to the inherent problem of cancer proteomics (Fig. 2).

Obtaining genome data from individual samples for proteomic studies is not as costly as it once was. However, obtaining proteome data from genome data remains challenging. This is because the proteome cannot be determined in one way, considering the transcription site, splicing variants, and post-translational modifications. In contrast, if all possible proteins are deposited in the database, the database size will be exceedingly large because of the extreme redundancy. Such a database would be impractical for protein identification. We need to prioritize the possible proteins and create a proteome database. This creation process utilizes a bioinformatics approach. Table 1 summarizes the software available for the creation of proteome databases. The outlines and details of the reported software vary, and the methods to create software have not been established yet. Further software development would be advantageous.

INTERNATIONAL EFFORTS FOR CANCER PROTEOGENOMICS

The importance of cancer proteomics is internationally recognized. Recently, the International Cancer Proteogenomics (ICPC) was established (https://proteomics.cancer.gov/programs/international-cancer-proteogenome-consortium). The National Cancer Center in Tokyo participates in this consortium as a representative of Japan. The participating research groups register with the consortium by indicating the types of cancer they will investigate. Data generated by each participating center is shared with the other consortium members. Research results for lung cancer, colorectal cancer, breast cancer, ovarian cancer, and gastric cancer have been reported from overseas participants, and the aim is to validate the possible utilities of cancer proteomics in the diverse cancer types. The Japanese team is exploring the proteogenomics of sarcomas. This builds on the prior history of genome and proteome studies of sarcomas conducted at the aforementioned National Cancer Center and the establishment of a research network among Japanese researchers, clinicians, and pathologists. Consortium members from the United States and Australia have also studied sarcoma proteogenomics as well as other cancer types. The collective sarcoma proteogenomic data will be shared among the participant research groups.

PERSPECTIVE OF CANCER PROTEOGENOMICS

Proteogenomics is a potentially powerful tool for cancer research. However, there is still no guarantee that the approach will lead to the discovery of biomarkers and therapeutic targets. Three current hypotheses need to be explored and either validated or refuted.

The first hypothesis is that proteins harboring mutated amino acid sequences can stably exist. Mutations can affect protein stability. Certain mutations may influence the posttranslational modifications that might be required for the stability, functionality, and interactions with other molecules. Biomarker proteins should stably exist in the human body, and therapeutic target proteins should stably function in cancer cells. We need to examine how mutations affect the status of proteins.

The second hypothesis is that some of the proteins with mutated amino acid sequences are relevant as biomarkers. To be a useful biomarker, the expression of the protein should correlate with specific clinical or pathological phenotypes. We need to investigate if such proteins exist through well-designed experiments.

The third hypothesis is that therapeutic target proteins exist among those with mutated amino acid sequences. To be a therapeutic target, the proteins should be functionally important, and small molecules or antibodies that specifically react with them should be developed. We need to identify proteins that can be exploited as therapeutic targets and investigate the molecules that suppress the identified proteins.

One criterion for the success of proteogenomics is that the proteins with mutated amino acid sequences discovered...
by the proteogenomics approach lead to novel biomarkers and therapeutic targets. However, such success does not only rely on the performance of proteogenomics. Success also depends on the development of other technologies, and the medical requirements.

Conclusive results will likely take a long time to realize. We need to establish proper milestones and evaluate the output of research. The history of cancer proteomics has prepared the scientific community for the challenge. We should not try to run before we can walk. We may need to validate and confirm hypotheses one-by-one before trying to apply the results to clinical applications.

CONCLUSIONS

Proteogenomics seems to be promising novel approach to solve the inherent problems of traditional cancer proteomics. Proteogenomics can provide more informative data in the experiments using the traditional protein biochemistry experiments such as electrophoresis. Thus it has a considerably effects on basic science. Still, many challenges remain to be overcome before the translation of research outcomes to clinical benefits. We need to test the aforementioned hypotheses with the intent to identify limitations and validate successful options.

CONFLICT OF INTEREST

The authors have nothing to declare.

REFERENCES

1) Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet. 2002;359:572–577.
2) Anderson NL, Anderson NG. The human plasma proteome: History, character, and diagnostic prospects. Mol Cell Proteomics. 2002;1:845–867.
3) Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016;537:347–355.
4) Woo S, Cha SW, Merriniew G, et al. Proteogenomic database construction driven from large scale RNA-seq data. J Proteome Res. 2014;13:21–28.
5) Woo S, Cha SW, Na S, et al. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. Proteomics. 2014;14:2719–2730.
6) Woo S, Cha SW, Bonissone S, et al. Advanced proteogenomic analysis reveals multiple peptide mutations and complex immunoglobulin peptides in colon cancer. J Proteome Res. 2015;14:3555–3567.
7) Wang X, Zhang B. CustomProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics research. Bioinformatics. 2013;29:3233–3237.
8) Wang X, Slebos RJ, Wang D, et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res. 2012;11:1009–1017.
9) Evans VC, Barker G, Heesom KJ, et al. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods. 2012;9:1207–1211.
10) Sheynkman GM, Johnson JE, Jagtap PD, et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variants. BMC Genomics. 2014;15:703.
11) Li Y, Wang X, Cho JH, et al. JUMPg: An integrative proteogenomics pipeline identifying unannotated proteins in human brain and cancer cells. J Proteome Res. 2016;15:2309–2320.
12) Zickmann F, Renard BY. MSProGene: Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. Bioinformatics. 2015;31:i106–115.
13) Park H, Bae J, Kim H, et al. Compact variant-rich customized sequence database and a fast and sensitive database search for efficient proteogenomic analyses. Proteomics. 2014;14:2742–2749.
14) Wen B, Xu S, Zhou R, et al. PGA: An R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. BMC bioinformatics. 2016;17:244.
15) Nagaraj SH, Waddell N, Madugundu AK, et al. PGTools: A software suite for proteogenomic data analysis and visualization. J Proteome Res. 2015;14:2255–2266.
16) Krasnov GS, Dmitriev AA, Kudryavtseva AV, et al. PPLine: An automated pipeline for SNP, SAP and splice variant detection in the context of proteogenomics. J Proteome Res. 2015;14:3729–3737.
17) Crape J, Ndah E, Koch A, et al. PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res. 2015;43:e29.
18) Ruggles KV, Tang Z, Wang X, et al. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. Mol Cell Proteomics. 2016;15:1060–1071.
19) Wen B, Xu S, Sheynkman GM, et al. sapFinder: An R/Bioconductor package for detection of variant peptides in shotgun proteomics experiments. Bioinformatics. 2014;30:3136–3138.
20) Zhu Y, Engstrom PG, Tellgren-Roth C, et al. Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation in a complete genome assembly of Malassezia sympodialis. Nucleic Acids Res. 2017;45:2629–2643.
21) Aebersold R, Agar JN, Amster IJ, et al. How many human proteoforms are there? Nat Chem Biol. 2018;14:206–214.
22) Alfaro JA, Sinha A, Kislinger T, et al. Onco-proteogenomics: Cancer proteomics joins forces with genomics. Nat Methods. 2014;11:1107–1113.
23) Rodriguez H, Pennington SR. Revolutionizing precision oncology through collaborative proteogenomics and data sharing. Cell. 2018;173:535–539.
24) Sun H, Xing X, Li J, et al. Identification of gene fusions from human lung cancer mass spectrometry data. BMC Genomics. 2013;14 Suppl 8:S5.
25) Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513:382–387.
26) Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics
connects somatic mutations to signalling in breast cancer. Nature. 2016;534:55–62.

27) Zhang H, Liu T, Zhang Z, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell. 2016;166:755–765.

28) Mun DG, Bhin J, Kim S, et al. Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell. 2019;35:111–124.e110.