

Advancements in the Application of Proteogenomic in Assisting the Discovery of Novel Cancer Antigens

Di Zhao *

Department of Microbiology, University of Washington, Seattle, USA

* Corresponding author: zhaodi99@uw.edu

Abstract. Over 18.1 million patients worldwide affected by more than 100 different types of cancer. Cancer is one of the critical focuses of medical research today. Researchers have discovered numerous types of cancer, with many patients and a lack of effective unified treatment methods. Therefore, it is urgent to study the pathogenesis and treatment methods. Proteogenomic is an interdisciplinary field focusing on genomics, transcriptomics, and proteomics, which has shown significant effectiveness in tumor analysis and treatment development. Its comprehensive analysis capabilities can help scientists and physicians accurately discover, predict, and treat cancer. This article summarizes the relevant content of using proteogenomic in analyzing and treating various cancers in recent years. It explores the role of proteogenomic in assisting the discovery of new antigens for cancer immunotherapy. This review highlights the latest advancements in proteogenomic and its significant findings in cancer analysis and treatment, suggesting a promising future for its integration into clinical practice.

Keywords: Proteogenomic; Cancer; Cancer antigens; Immunotherapy.

1. Introduction

Over 18.1 million patients worldwide affected by more than 100 different types of cancer [1]. The vast number and variety of cancer types make developing a unified treatment approach challenging. Therefore, the emerging field of proteogenomic, which integrates disciplines such as proteomics, genomics, and transcriptomics, offers new insights for cancer treatment. Proteogenomic is a method that compares unknown data with existing databases using mass spectrometry. It is primarily used for identifying, correcting, and predicting genes, peptides, and proteins. Proteogenomic can directly examine gene mutations in patients or analyze RNA transcription and protein translation processes. It can even predict the consequences of specific changes caused by these processes.

Considering all aspects together, proteogenomic has shown significant effectiveness in tumor analysis and treatment development. It is particularly suitable for identifying various mutations and developing targeted therapy approaches. Proteogenomic research has discovered associations between breast cancer incidence mechanisms and Rb protein levels, CDK inhibitors, and glutaminase inhibitors relevant to lung cancer treatment. This review focuses on recent advancements in proteogenomic while summarizing its applications in identifying molecular characteristics or finding therapeutic targets for several types of cancers. The aim is to explore the scientific contributions of proteogenomic in cancer treatment.

2. Principles and Applications of Proteogenomic

2.1. Principles of Proteogenomic

Proteogenomic is an interdisciplinary field that integrates genomics, transcriptomics, and proteomics. It emerged as a new research field in the early 21st century. This new approach in genomics has rapidly developed into an essential tool for cancer research over the past two decades.

Genomics, transcriptomics, and proteomics correspond to the DNA, RNA, and proteins of organisms, respectively. DNA codes for the crucial genetic information in organisms, while proteins are the



foundation for biological activities. RNA connects DNA and proteins as a vector. The genetic information flow in all lives is called central dogma. Genomics studies the structure and function of species' genomes by reading and quantifying all genes of different organisms inside the same species through high-throughput DNA sequencing. With the help of bioinformatics, it creates a gene library to compare relationships between genes and their impact on organisms. Transcriptomics differs from genomics as it studies RNA generated through transcription from an organism's DNA. Like genomics, proteomics investigates multiple proteins that make up an organism using liquid chromatography-based mass spectrometry (LC-MS) technology to generate protein data. This data is then compared with reference protein sequence databases to study characteristics, expression levels, and interactions of proteins within organisms [1]. Focusing these three disciplines separately would result in a narrow scope and direction of research. Therefore, scholars have proposed the integration of genomics, transcriptomics, and proteomics, leading to the emergence of proteogenomic technology, identifying novel peptides in proteomic data and enhances gene models using proteomic data.

2.2. Applications of Proteogenomic

Proteogenomic, when combined with three disciplines, has various applications. The fundamental function of proteogenomic is to identify peptide types and determine the gene location encoding these peptides based on mass spectrometry data and databases. Additionally, it can locate previously undiscovered protein-coding sites or variant peptide segments to improve gene models and protein sequence databases [2]. In clinical cancer research, proteogenomic plays a significant role. One characteristic of cancer is its high specificity, resulting in different treatment methods for each type of cancer.

Furthermore, even within the same type of cancer, the response to identical treatment methods varies among patients. Therefore, by integrating proteogenomic into clinical practice, cancers are classified into different subtypes based on their carcinogenic pathways. Treatment drugs are selected according to patients' tumors' genetic characteristics and molecular targets [3].

Proteogenomic encompasses three analytical disciplines that can be used simultaneously to complement the limitations of each method. Genomic techniques cannot accurately predict the level at which they are translated into proteins. A single transcript can generate more than one protein through alternative splicing for transcriptomics. Proteomics, however, involves more significant uncertainty as many proteins only become active after modification and require specialized proteomic methods for study. Additionally, factors such as protein-RNA complexes and protein degradation rates influence proteomics. Proteogenomic fundamentally addresses these limitations by allowing simultaneous study of the same sample's genome, transcriptome, and proteome and obtaining more accurate results through cross-comparison. Therefore, proteogenomic has significant advantages in studying genomic and protein mutations.

In cancer research, proteogenomic can analyze the mechanisms of cancer from different perspectives and provide molecular-level disease information for each cancer patient, thereby offering new insights into finding treatment methods for cancer.

3. Proteogenomic in the Discovery of Novel Cancer Antigens

As of 2020, there are approximately 18.1 million cancer patients worldwide, with the most common types being breast, lung, and colorectal cancer. Different patients exhibit distinct pathological and molecular characteristics, which pose significant challenges in researching effective treatments for these cancers. However, the large number of patients also provides abundant clinical trial data for experiments using novel approaches such as proteogenomic.

This article reviews recent research to summarize the application of proteogenomic in assisting the discovery of new antigens in treating cancers.

3.1. Breast Cancer

In a study exploring the protein gene map of breast cancer incidence and targeted therapy, the authors used proteogenomic analysis to analyze 122 types of breast cancer. They utilized untreated tumor cells and extracted uniform specimens by freezing and grinding them. They identified somatic mutations, the copy number and transcription of genes, the phosphorylation and acetylation sites of proteins using STAR methods. They determined their genomic and transcriptomic characteristics using the PAM50 model before performing clusters of single- and multi-omics to discover subtypes NMFLumA and NMFLumB.

The authors also found that NMFLumB had fewer PIK3CA mutations but could not capture all biological differences between the two subtypes. In addition, they used proteogenomic metabolic analysis to identify metabolic features in four subtypes; they discovered that NMFLumA-I and NMFLumB-1 had similar metabolic changes while being opposite from those of NMFBasal-I's metabolic features. The authors also analyzed biopsy samples from breast cancer patients receiving anti-ERBB2 antibody treatment using proteogenomic techniques; they discovered pseudo-ERBB2+ resistance cases, which may be related to TOP2A amplification and protein overexpression. Analyzing ERBB2 phosphopeptide levels revealed alternative drivers of ERBB signaling associated with PAM50 HER2E [4].

Another study on the relationship between protein and gene expression of triple-negative breast cancer (TNBC) patients after treated with the chemotherapy, conducted through proteogenomic analysis, found that neoadjuvant chemotherapy response biomarkers are associated with pathologic complete response (pCR). The Molecular Signatures Database (MSigDB) Hallmark metabolic pathways are upregulated in pCR samples using Gene set enrichment analysis (GSEA), especially the level of interferon alpha and gamma in cell cycle G2-M checkpoint pathways [5].

3.2. Lung Cancer

In a research study, the biological characteristics of lung adenocarcinoma (LUAD, n=110, matched NATs, n=101) were analyzed using various proteogenomic methods. The study identified C-terminal phosphorylation of the protein DNMBP (TUBA), which contains guanine exchange factor (GEF), although its exact function remains unknown. Two adverse prognostic markers for lung cancer were discovered: Gremlin 1 (GREM1) protein and Ovarian cancer immunoreactive antigen domain containing 2 (OCIAD2). Further extensive research on LUAD revealed a therapeutic vulnerability in IDO1 inhibition within anti-CTLA4 therapy [6]. Another recent study focused on LUAD heterogeneity and survival-related factors and pathways, classifying LUAN into TRU, PI, and PP subtypes based on the expression of RNA and protein using proteogenomics. These subtypes exhibit overexpression of specific genes through regulation of translation and modification, leading to reduced immune cell activity. The overactive EGFR signaling in the TRU subtype can be considered a potential therapeutic target, while the PI subtype provides additional predictive capabilities for LUAD prognosis. The proteogenomic profile in PP subtype confirms the optimistic effects using CDK and glutaminase inhibitors treatment [7].

A study analyzing the protein genome found that the correlation between lung cancer and smoking was insignificant. By simultaneously using transcriptomics and proteomics analysis, 15 variant subtypes of cancer-driving genes were identified. The study also demonstrated varying levels of MAPK signaling activation in EGFR patients' lung cancer cells, although it could not be determined whether this was correlated with clinical outcomes. In the same study, based on protein-based classification of subtypes, the EGFR-L858R mutation was found to be more common in late-stage samples, and cancer cells with this mutation were also more prone to metastasis [8].

3.3. Colorectal Cancer

A study based on proteogenomics analysis of colon cancer found that TMT data has a significant advantage in predicting gene changes, followed by label-free proteomics data. TMT data can indicate

protein alterations in some somatic cell mutations, such as the APC mutation that leads to shortened proteins. Phosphorylation of Rb in cells affects cell proliferation and apoptosis, which can increase the risk of colorectal cancer. The same study also discovered that CT antigens are shared among multiple patients [9].

In a significant study, the tumor microenvironment characteristics in colorectal cancer was discussed through analyzing high-plex proteogenomic characterization using a novel Digital Spatial Profiler (DSP) technology. When analyzing genes in colorectal tissue samples using this technique, strong correlations were observed between epithelial cell markers and protein/RNA expression, such as EPCAM, AKT1, and Cytokeratin. T-cell markers like CD3 and CD4 showed significantly correlated with total immune cell markers like CD45, while weaker associated with B-cell marker CD20. Moreover, this study found high mRNA and protein levels of PD-L1 expression, indicating better immunogenicity response and as a prognostic biomarker for immune therapy outcomes in patients. Additionally, it was discovered that CD14 levels were lower in tumors compared to adjacent normal tissues [10].

Another study in protein genomics has discovered mutations in the CASP5, RNF43, LTN1, and BMP2 genes in microsatellite instability-high (MSI-H) tumors of human colon cancer. The glycolytic enzymes were significantly increased in the MSI subtype compared to UMS ones through analysis the profiles of protein and mRNA. Glycolysis leads to an increase in lactate production, which inhibits the function of CD8 T-cells. Therefore, inhibiting glycolysis can reduce lactate production and enhance sensitivity to immune checkpoint blockade therapy [9].

3.4. Prostate Cancer

In a study focusing on the features of protein genomics for curable prostate cancer, it was found that the accuracy is much higher using hazard ratios (HRs) of protein abundances than RNA in predicting prostate cancer recurrence. The relationship between high ACAD8 protein abundance and poor outcome was significant, making it a potential prognostic marker [11].

Another study investigating the heterogeneity of proteomics in localized human prostate cancer mentioned a significant increase in copy number alterations after analyzing the tumor samples of 39 prostate cancer patients. Correspondingly, SET Domain Bifurcated 1 (SETDB1), a gene associated with melanoma, was found to be overexpressed in prostate cancer cells. Abnormal changes in the quantities of proteins like PABL3 were more pronounced in higher-grade tumors [12].

Prostate cancer diagnosis methods are limited and invasive; therefore, a study utilizing proteomics for identifying prostate cancer through urine analysis generated great interest last year. The main objective of this research was to discover new protein targets for prostate cancer through proteomic analysis of urine samples. The dysregulated proteins related to prostate cancer patients were identified and integrated using bioinformatics after collecting relevant data from Liquid Chromatography Tandem-Mass-Spectrometry (LC-MS/MS). This study revealed natural dysregulation of proteins such as AMBP in urine samples from prostate cancer patients and an elevated level of sE-cadherin compared to healthy individuals. Furthermore, this research also investigated the impact of relevant protein mutations on biological signaling pathways, predicting that these mutations would affect the stability of protein-protein interactions (PPI) and subsequently influence the progression of prostate cancer [13].

3.5. Cervical Cancer

Proteomic analysis based on TMT labeling can identify all proteins in cervical cancer patients' cervical exfoliated cells, including 351 differentially expressed proteins, mainly involved in cell apoptosis and its related regulation. PRDX2 plays a vital role in cell proliferation and differentiation while protecting cells from apoptosis. Moreover, PRDX2 regulates the cell cycle by affecting ROS homeostasis, inducing cell death, and promoting cancer cell proliferation. The same study mentioned that abnormal expression of p53 provides a favorable environment for tumor cells' unlimited

proliferation. The anti-apoptotic protein CSTA can promote cancer cell proliferation and migration [14].

4. Current Developments and Challenges in Proteogenomic Research

New instruments have led to a qualitative leap in proteogenomic research in recent years. Furthermore, improvements in various methods have significantly increased the depth of proteogenomic studies. However, despite these advancements, the field of proteogenomic is still relatively young and needs more protein and genomic data to support its rapid development. In medical cancer research, proteogenomic has played a crucial role in further subtyping cancers, leading to a shift towards precision oncology. Researchers strive to dissect more precise mechanisms underlying different cancer subtypes to develop corresponding therapeutic drugs.

In the future, proteogenomic needs to analyze thousands of significant types of cancer genomic data while exploring interdisciplinary collaborations by integrating proteogenomic with precision oncology and other disciplines. This integration aims to improve the treatment environment for cancer patients by utilizing multiple analytical approaches on relevant data sets. Further, proteogenomic can redirect its focus toward potential therapeutic targets by applying highly prognostic and predictive biomarkers in clinical trials for developing novel treatment strategies targeting specific cancer subtypes.

5. Conclusion

Proteogenomic, as a novel discipline, can examine gene mutations, and reveal the role of proteins in the human body and their therapeutic effects on cancer. It can also provide patients with more precise treatment plans and predict patient prognosis. This highlights the importance of proteogenomic in discovering the pathogenic mechanisms of cancer, studying targeted drugs, and designing accurate treatment plans. The study of proteogenomic in cancer provides references for researching other diseases and facilitates investigations into various complex conditions. Although proteogenomic research is comprehensive, it is limited by existing databases, making it unable to fully analyze genomic data or overcome uncertainties related to peptides in the proteome. Therefore, establishing a public data repository is imperative. Further in-depth research and refinement of research directions are necessary for every study involving the use of proteogenomic to analyze diseases; simultaneously, collaboration with other disciplines should be pursued to develop new treatment strategies. Proteogenomic represents a milestone in cancer treatment through its comprehensive detection capabilities for cases and has profound clinical implications.

References

- [1] Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023. *CA Cancer J Clin.* 2023, 73 (1): 17 - 48.
- [2] Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*, 2014, 11 (11): 1114 - 25.
- [3] Rodriguez H, Zenklusen JC, Staudt LM, et al. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. *Cell*, 2021, 184 (7): 1661 - 1670.
- [4] Krug K, Jaehnig EJ, Satpathy S, et al. Clinical Proteomic Tumor Analysis Consortium. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell*, 2020, 183 (5):1436 - 1456.e31.
- [5] Anurag M, Jaehnig EJ, Krug K, et al. Proteogenomic Markers of Chemotherapy Resistance and Response in Triple-Negative Breast Cancer. *Cancer Discov*, 2022, 12 (11): 2586 - 2605.
- [6] Gillette MA, Satpathy S, Cao S, et al. Clinical Proteomic Tumor Analysis Consortium. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell*, 2020, 182 (1): 200 - 225.e35.
- [7] Soltis AR, Bateman NW, Liu J, et al. Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways. *Cell Rep Med*, 2022, 3 (11): 100819.
- [8] Chen YJ, Roumeliotis TI, Chang YH, et al. Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. *Cell*, 2020, 182 (1): 226 - 244.e17.

- [9] Vasaikar S, Huang C, Wang X, et al. Clinical Proteomic Tumor Analysis Consortium. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell*, 2019, 177 (4): 1035 - 1049.e19.
- [10] Wang N, Wang R, Li X, et al. Tumor Microenvironment Profiles Reveal Distinct Therapy-Oriented Proteogenomic Characteristics in Colorectal Cancer. *Front Bioeng Biotechnol*, 2021, 9: 757378.
- [11] Sinha A, Huang V, Livingstone J, et al. The Proteogenomic Landscape of Curable Prostate Cancer. *Cancer Cell*, 2019, 35 (3): 414 - 427.e6.
- [12] Champi K, Guo T, Zhong Q, et al. Convergent network effects along the axis of gene expression during prostate cancer progression. *Genome Biol*, 2020, 21 (1): 302.
- [13] Lima T, Barros AS, Trindade F, et al. Application of Proteogenomics to Urine Analysis towards the Identification of Novel Biomarkers of Prostate Cancer: An Exploratory Study. *Cancers (Basel)*, 2022, 14 (8): 2001.
- [14] Xu D, Zhu X, Ren J, et al. Quantitative proteomic analysis of cervical cancer based on TMT-labeled quantitative proteomics. *J Proteomics*, 2022, 252: 104453.