OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf087
https://doi.org/10.1093/bib/bbaf087
Review

Integration of proteomics profiling data to facilitate discovery of cancer neoantigens: a survey

Shifu Luo[1,2], Hui Peng[1,3], Ying Shi[1,4], Jiaxin Cai[1], Songming Zhang[1], Ningyi Shao[2,*], Jinyan Li[1,*]

[1]Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology, Shenzhen, 518107, Guangdong, China
[2]Faculty of Health Sciences, University of Macau, Taipa, Macao SAR 999078, China
[3]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore
[4]School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, China

*Corresponding authors. Ningyi Shao, Faculty of Health Sciences, University of Macau, Taipa, Macao SAR 999078, China. E-mail: nshao@um.edu.mo; Jinyan Li, Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology, Shenzhen, Guangdong, China. E-mail: lijinyan@suat-sz.edu.cn

Abstract

Cancer neoantigens are peptides that originate from alterations in the genome, transcriptome, or proteome. These peptides can elicit cancer-specific T-cell recognition, making them potential candidates for cancer vaccines. The rapid advancement of proteomics technology holds tremendous potential for identifying these neoantigens. Here, we provided an up-to-date survey about database-based search methods and de novo peptide sequencing approaches in proteomics, and we also compared these methods to recommend reliable analytical tools for neoantigen identification. Unlike previous surveys on mass spectrometry-based neoantigen discovery, this survey summarizes the key advancements in de novo peptide sequencing approaches that utilize artificial intelligence. From a comparative study on a dataset of the HepG2 cell line and nine mixed hepatocellular carcinoma proteomics samples, we demonstrated the potential of proteomics for the identification of cancer neoantigens and conducted comparisons of the existing methods to illustrate their limits. Understanding these limits, we suggested a novel workflow for neoantigen discovery as perspectives.

Keywords: cancer neoantigens; proteomics; database-based search methods; de novo peptide sequencing; deep learning

Introduction

The formation of tumors is primarily attributed to the accumulation of genomic variations within cells. Each cell accumulates approximately 15–50 mutations annually. Some advantageous clones are retained through clonal selection, eventually evolving into tumors. The diversity of mutations and clones makes tumors highly heterogeneous [1–3]. Genomic mutations in tumor cells produce self-antigens not expressed in normal cells. These antigens, known as neoantigens or neoepitopes, fall under the category of tumor-specific antigens [4, 5]. Neoantigens bind to major histocompatibility complex (MHC) molecules within the cell as peptides and are subsequently presented on the cell membrane surface for recognition by T cells [6]. Unlike tumor-associated antigens, neoantigens can evade central T-cell tolerance and do not pose a risk to normal tissues [7–10].

Neoantigens are widely utilized in T cell-based immunotherapies, including T-cell receptor-engineered T cells (TCR-T) and cancer vaccines. Both approaches aim to generate antigen-specific T cells to inhibit tumor growth [11, 12]. Notably, clinical trials for neoantigen-based cancer vaccines are being vigorously conducted [13–17]. These vaccines can be categorized into DNA vaccines, RNA vaccines, peptide vaccines, cell-based vaccines, and viral vaccines, depending on the platform [18].

One of the critical aspects in developing a cancer vaccine is the identification of immunogenic candidate neoantigens capable of eliciting a robust and specific immune response targeting tumor cells [12]. Current methodologies predominantly rely on whole genome sequencing (WGS) or whole exome sequencing (WES) to detect cancer-associated variations at the DNA level [19, 20]. Subsequently, these identified variations undergo a series of bioinformatics pipelines, including Human Leukocyte Antigen (HLA)-typing, HLA-affinity, and TCR recognition, to predict potential neoantigens [21–23]. However, the accuracy of current machine learning-based neoantigen prediction tools is typically around or below 5% when applied to the neoantigen prediction of mutations identified by WES in cancer [24–28]. It remains uncertain whether DNA-level variants can be effectively translated into proteins, which must be carefully considered in neoantigen prediction.

Proteomic technologies, primarily based on liquid chromatography coupled with tandem mass spectrometry (MS) are increasingly utilized in cancer research [29–32]. Importantly, proteomic data are systematically curated and deposited in public databases such as ProteomeXchange and PRIDE, significantly enhancing the utilization and accessibility of proteomic information [33, 34]. Since proteins are the executors of biological functions, proteomics offers distinct advantages over DNA and RNA sequencing.

Received: October 17, 2024. **Revised:** December 29, 2024. **Accepted:** February 19, 2025

Proteomics can capture successfully translated mutations, RNA alternative splicing products, gene fusion products, and proteins translated from non-coding regions and circular RNAs [8, 35]. Neoantigen identification based solely on DNA or RNA sequencing data may lead to the omission or misidentification of neoantigens. Therefore, integrating proteomics into the pipeline for cancer neoantigen screening is expected to enhance both the quantity and accuracy of neoantigen detection.

Polyakova et al. have previously reviewed the unique advantages of proteomics in identifying neoantigens and suggested incorporating proteomics into the pipeline for neoantigen discovery [36]. Verma *et al.* then combined multi-omics and bioinformatics tools to create a framework for using proteomics in neoantigen identification [37]. The framework utilizes the patient's paired WES and RNA-seq data as a reference database for MS searches, thereby enabling the identification of neoantigens. A recent review also discussed this framework, highlighting the potential and challenges of using proteomics in neoantigen vaccine design [38].

However, no review currently provides a detailed analysis or comparison of the practical applications of this theoretical framework. It remains theoretical, lacking comprehensive practical solutions and tool recommendations. Additionally, they focus only on database-based search methods, neglecting comparisons with de novo peptide sequencing approaches in neoantigen identification. The rapid evolution of artificial intelligence (AI) has led to the creation of many advanced proteomics analysis tools for de novo peptide sequencing. These tools overcome the limitations of database-based search methods regarding reference databases, making them more suitable for neoantigen identification [39]. Nevertheless, no one has yet analyzed or summarized the potential of these tools in cancer neoantigen identification, nor explored how to integrate these tools into pipelines for cancer neoantigen screening.

In this paper, we provided an updated survey of AI-assisted de novo peptide sequencing methods and their applications in neoantigen discovery. We also examined methods for implementing and recommending tools for neoantigen identification using both proteomics database search techniques and de novo peptide sequencing approaches. Importantly, we conducted a case study analysis on liver cancer samples to compare existing methods and showcase the potential of proteomics in neoantigen identification. This analysis also revealed the limitations of current methods. Finally, we proposed a novel proteomics-based workflow for neoantigen identification, which holds promise for accelerating the development of clinical cancer vaccines.

Materials and methods

We can calculate the H-index in Web of Science (https://webofscience.clarivate.cn/):

1) Access Web of Science and select "Title" in the search bar dropdown menu.
2) Enter the article title you are looking for and initiate the search.
3) Click on the article title from the search results to open its detailed record.
4) Click on the "Citations" link to view the citations for the article.
5) On the citations page, click on the "Citation Report" button to generate a comprehensive report.
6) In the citation report, set the publication years filter to 2020–24 to focus on the most recent citations.
7) The H-index for the selected period will be displayed in the citation report.

Case study:

1) Construct the variant reference database of HCC

First, we downloaded all mutation information related to liver cancer from the Cosmic database (https://cancer.sanger.ac.uk/cosmic). Subsequently, the human standard protein reference sequence was downloaded from the Ensembl database (https://www.ensembl.org/) and modified the protein sequence according to the mutation information. Finally, the reference sequences of mutant proteins related to liver cancer were obtained, totaling 262,654 mutant protein sequences. Meanwhile, we retained the original Ensembl protein sequences as control.

2) Peptide identification

Download the HepG2 cell line and nine mixed samples data from HCC patients from the PRIDE database (https://www.ebi.ac.uk/pride/archive/projects/PXD036643). Analyze using MaxQuant (2.4.2.0), with parameters consistent with those used in the original paper [40] for the dataset, and use the variant reference database of HCC as the reference database.

Use Casanovo with the default weight model to directly infer each MS file. Summarize and deduplicate the results. Match the results with the variant reference database of HCC to obtain candidate neoantigens.

Cancer neoantigen screening pipeline in silico

Classical methods for screening immunogenic neoantigens, such as cDNA library screening, are time-consuming and costly [41–44]. The advent of next-generation sequencing and advancements in bioinformatics analysis has laid a solid foundation for rapid and high-throughput cancer neoantigen screening [45].

The in silico neoantigen screening process includes key steps such as mutation calling, HLA typing, HLA affinity prediction, and T cell recognition [23]. These steps and algorithms are designed to mimic *in vivo* cellular immune processes (Fig. 1A). Initially, the target protein is degraded into peptides by the proteasome within antigen-presenting cells (APCs). Subsequently, MHC class I molecules bind with the peptides in the endoplasmic reticulum to form peptide–MHC (p-MHC) complexes. These complexes are transported through the Golgi apparatus and presented on the cell membrane. Finally, the p-MHC on the cell membrane is recognized by CD8+ T-cell receptors, activating a cytotoxic T lymphocyte (CTL) response. For exogenous proteins, APCs ingest and degrade them into peptides, which are recognized by MHC-II molecules and presented on the cell surface. These peptides are recognized by CD4+ T-cells, which help enhance the cytotoxic effects of CTLs [46]. The primary mechanism by which cancer vaccines inhibit tumors relies on target antigens inducing strong and sustained responses from CD4+ T helper cells and CTLs [47].

The HLA typing algorithms, such as OptiType [48], Polysolver [49], and PHLAT [50], can achieve over 95% accuracy in HLA class I typing. Polysolver and OptiType perform comparably and are superior to PHLAT. Additionally, Polysolver can be applied at a higher resolution (eight digits) [49, 51]. Both OptiType and PHLAT can analyze RNA-seq data, with PHLAT also capable of performing HLA class II typing (Table 1).

After determining the HLA typing, tools like MixMHCpred [52], NetMHCpan 4.1 [53], and MHCflurry [54] are used to predict the

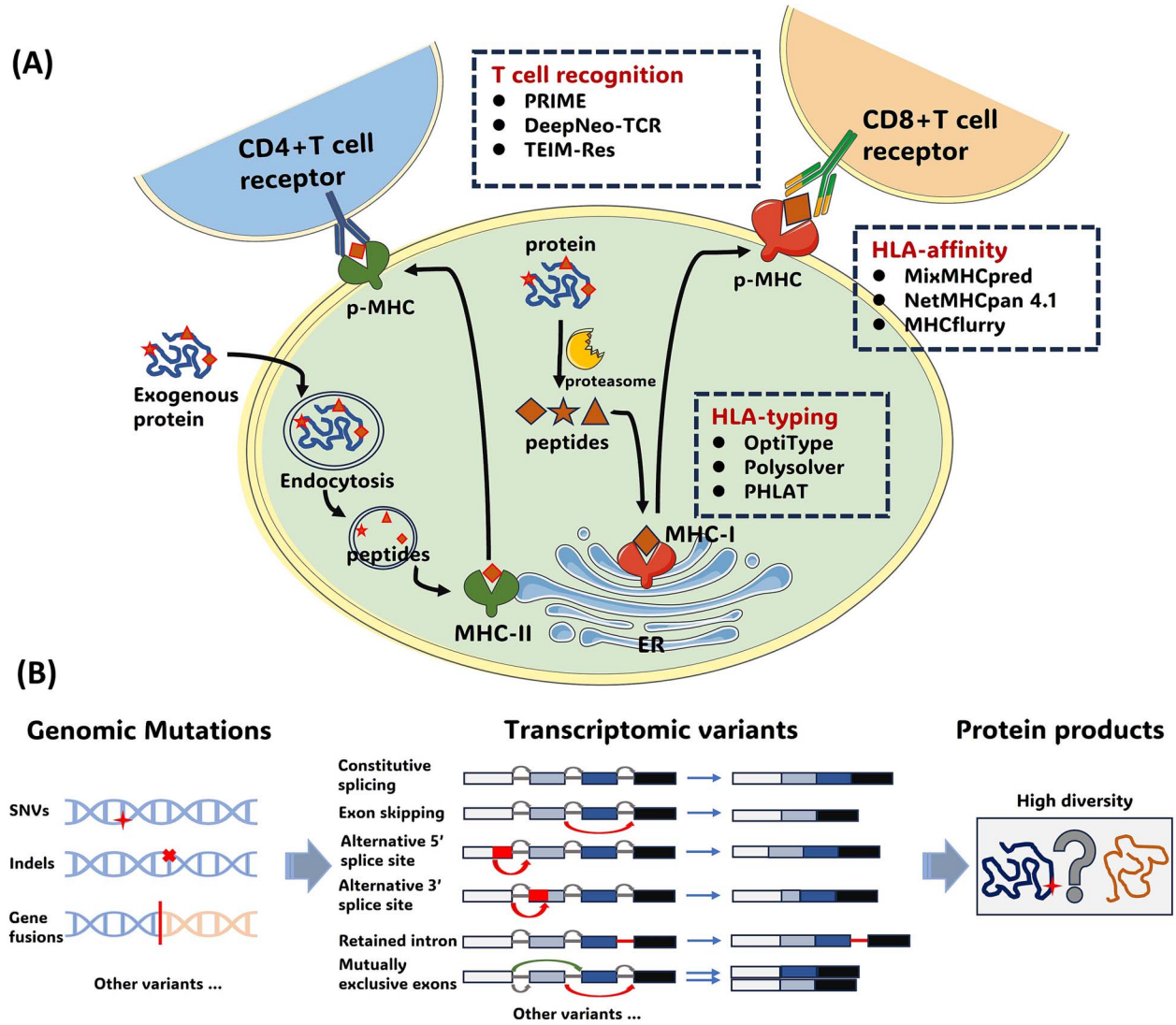

Figure 1. The pipeline of cancer neoantigen screening in silico and the pathways of neoantigen generation. (A) The process of antigen presentation by MHC class I/II molecules in APCs and the activation of T cell responses, along with the corresponding neoantigen screening tools for each step. MHC class I molecules are responsible for endogenous antigen presentation and CD8+ T cell activation. MHC class II molecules are responsible for exogenous antigen presentation and CD4+ T cell activation. (B) The mechanism of neoantigen generation includes genomic variations (SNVs, indels, and gene mutations) and transcriptome alternative splicing variants, which make protein products diversified.

affinity between peptides and MHC-I molecules, identifying which peptides are likely to be presented on the cell surface. In a benchmark experiment for MS MHC class I eluted peptides, NetMHCpan 4.1 demonstrated superior performance compared to other tools. The median positive predictive values are as follows: NetMHCpan-4.1: 0.8291, MixMHCpred: 0.7911, and MHCFlurry: 0.7256 [53]. Yet, MHCflurry has a significant speed advantage, exceeding 7000 predictions per second, which is 396 times faster than NetMHCpan 4.0 [54]. For MHC-II molecule predictions, NetMHCIIpan-4.0 [53] is available. Additionally, both NetMHCpan 4.1 and NetMHCIIpan 4.0 offer user-friendly web servers.

However, not all p-MHC complexes can be recognized by T cells. Hence, researchers have further developed T-cell recognition tools to predict the immunogenicity of peptides, such as PRIME [55], DeepNeo-TCR [56], and TEIM-Res [57]. Additionally, integrated analysis pipelines like pVACtools [21], MuPeXI [58], and OpenVax [59] facilitate user operations by directly providing candidate neoantigens based on somatic variant calling. Nevertheless, the positive rate of neoantigen prediction through these pipelines remains low [28].

In addition to improving the performance of these tools, the source of neoantigens is also a crucial factor in neoantigen screening. Neoantigens mainly originate from single nucleotide variants (SNVs), insertions and deletions (indels), gene fusions, RNA alternative splicing, and mutations in non-coding regions (some of which have translation functions), among others (Fig. 1B). SNVs refer to single nucleotide variations, which are the most common type of variation [63]. Indels refer to insertions or deletions of bases ranging from 1 to 10 000 bp in length and are the second most common type of genetic variation [64, 65]. Gene fusions occur when two or more genes hybridize, often due to chromosomal rearrangements or transcription-induced expression of chimeric genes [66]. Alternative splicing is a regulated process in which specific exons of a gene may be excluded from the premature mRNA, or certain introns may be retained, thereby increasing protein diversity [67]. Notably, up to 75% of

Table 1. The tools of cancer neoantigen screening pipeline in silico

Name	Application	Input	References
OptiType	HLA-I typing	WES/WGS/RNA-seq	Ref. [48]
Polysolver	HLA-I typing	WES	Ref. [49]
PHLAT	HLA-I/II typing	WES/RNA-seq	Ref. [50]
MixMHCpred	HLA-I affinity	HLA type & peptide sequence	Ref. [52]
NetMHCpan 4.1	HLA-I affinity	HLA type & peptide sequence	Ref. [53]
NetMHCIIpan 4.0	HLA-II affinity	HLA type & peptide sequence	Ref. [53]
MHCflurry	HLA-I affinity	HLA type & peptide sequence	Ref. [54]
PRIME	T cell recognition	Peptide sequence	Ref. [55]
DeepNeo-TCR	T cell recognition	Peptide sequence	Ref. [56]
TEIM-Res	T cell recognition	CDR3 & peptide sequence	Ref. [57]
pVACtools	Comprehensive tool	VCF file	Ref. [21]
MuPeXI	Comprehensive tool	VCF file	Ref. [58]
OpenVax	Comprehensive tool	VCF file	Ref. [59]
NeoPredPipe	Comprehensive tool	VCF file	Ref. [60]
TruNeo	Comprehensive tool	WES&RNA-seq fastq	Ref. [61]
Seq2Neo	Comprehensive tool	WES&RNA-seq fastq	Ref. [62]

the genome can be transcribed and potentially translated into proteins, and 99% of cancer mutations occur in noncoding regions [68, 69].

However, WES is limited to detecting variations at the DNA level and cannot ascertain whether these variations are expressed at the protein level. Similarly, RNA-seq data cannot confirm the translation of transcripts or predict protein modifications and variations. In contrast, proteomics directly examines the end products of DNA and RNA variations, making it more suitable for neoantigen detection.

Database-based search methods for protein identification from proteomics data

Proteomics workflow and software recommendation

Traditional proteomics strategies primarily include "top-down" and "bottom-up" approaches. The bottom-up approach analyzes proteolytic peptides, while the top-down method measures intact proteins. Among these, the bottom-up method, also known as shotgun proteomics, is widely used due to its high throughput and sensitivity [31, 70–72].

Shotgun proteomics is an indirect measurement of proteins through peptides derived from the proteolytic digestion of intact proteins (Fig. 2). The critical steps in proteomics analysis are the identification and quantification of proteins [73]. Initially, a reference database can be constructed based on the species, utilizing resources such as the human protein reference sequences from UniProt, Ensembl, and other relevant databases [74, 75]. These reference sequences are then fragmented according to specific rules to generate theoretical spectra. Next, candidate peptides are identified by matching and scoring the actual spectra against the theoretical spectra. Finally, the candidate peptides are assembled for protein identification and quantification [76–78]. In certain experimental methods, shotgun proteomics usually utilizes either labeled or label-free quantification techniques. The most used labeled method is tandem mass tag (TMT), a type of chemical labeling. TMT offers high stability and good reproducibility but has lower throughput and is expensive, making it suitable for small sample cohorts [79, 80].

On the other hand, label-free quantification primarily includes data-dependent acquisition (DDA) and data-independent acquisition (DIA) techniques. Both are high-throughput methods suitable for large-sample cohort detection [81, 82]. However, DDA is gradually being replaced by DIA due to its low coverage and poor reproducibility. DDA only performs secondary fragmentation and MS2 detection on high-abundance precursor ions, whereas DIA performs a full scan of all precursor ions and their fragmented product ions [83, 84]. Due to the high heterogeneity of tumors, multiple clones exist, and neoantigens vary in abundance. DDA struggles to capture those low-abundance neoantigens. In a cohort of 195 prostate cancer patients, DIA identified more peptides and proteins than DDA, adding 17.3% to 57.3% of proteins per patient [85]. This suggests that, in theory, the DIA technique may be more effective for detecting cancer neoantigens.

Table 2 lists the commonly used software based on the database search method. We used the Web of Science tool to calculate the H-index for each software over the past 5 years (2020–24). Among these, MaxQuant [86] stands out with the highest H-index of 91. MaxQuant is favored by many researchers due to its strong performance and open-source nature [87, 88]. In an analysis identifying HLA Class I allele-specific peptides, four mainstream tools were evaluated. The true-positive rate of peptide identifications made by each engine was: Comet = 41%, MaxQuant = 59%, MS-GF + = 58%, and Peaks = 68%, with Peaks leading the other software [89]. However, another study demonstrated that MaxQuant is better suited for identifying low-abundance proteins, aligning more closely with the requirements for neoantigen identification [90]. Therefore, we recommend using MaxQuant for examples and analysis.

Besides MaxQuant, widely used tools such as Mascot [91] and MS-GF+ [92] are frequently employed for comparison analysis [93–96]. Notably, MSFragger [97] utilizes a fragment-ion indexing method, which increases its speed by over 100 times compared to most existing database search-based tools. With an H-index of 45, it surpasses many classic software tools, demonstrating significant potential. In recent years, to accommodate the development of DIA technology, MaxQuant, and MSFragger have extended their capabilities to include components specifically for analyzing DIA data, such as MaxDIA [98] and MSFragger-DIA [99].

Methods for the establishment of reference databases

For MS-based neoantigen identification, it is essential to establish a specific reference database. For instance, Li *et al.* developed the Cancer Proteome Variation (CanProVar) database,

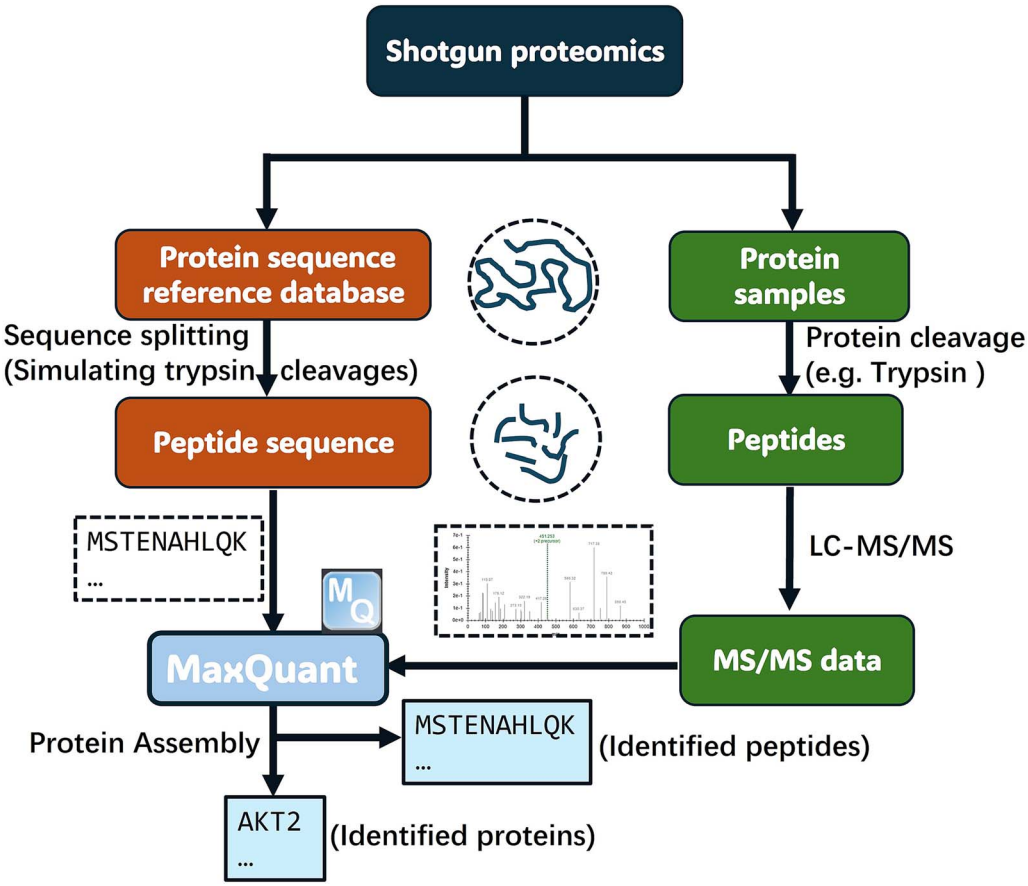

Figure 2. Shotgun proteomics workflow. The actual spectra are obtained by protein cleavage and mass spectrometry, while the theoretical spectra are obtained by interrupting the reference protein sequence according to the theoretical site of the corresponding cleavage method. Peptide identification is achieved by matching and scoring theoretical spectra against actual spectra, and peptides are assembled to achieve protein identification.

Table 2. Database search software for proteomic analysis

Name	Published year	H-index in the past five years	True-positive peptide identifications rates [89]	Running time in *Escherichia coli* dataset (min) [100]	Open source	References
SEQUEST	1994	15	NA	NA	Yes	Ref. [101]
Mascot	1999	43	NA	NA	Yes	Ref. [91]
X!Tandem	2004	30	NA	NA	Yes	Ref. [102]
pFind	2005	11	NA	57.9	Yes	Ref. [103]
MaxQuant	2008	91	59%	464.5	Yes	Ref. [86]
Peaks DB	2012	29	68%	NA	No	Ref. [104]
MODa	2012	14	NA	NA	Yes	Ref. [105]
Comet	2013	47	41%	236.0	Yes	Ref. [106]
MS-GF+	2014	36	58%	259.5	Yes	Ref. [92]
MSFragger	2017	45	NA	NA	Yes	Ref. [97]

which integrates information on human cancer-related variations from multiple databases [107, 108]. This CanProVar database is used instead of traditional reference sequence databases for peptide identification. While this strategy can identify variant peptides, it faces significant challenges, such as a restricted search space. It is important to note that these cancer variant-related databases only include DNA-level variants.

At the RNA level, alternative splicing events are frequent in cancer and play a crucial role in tumor development and progression [109–112]. Variants generated by alternative splicing, as important components of neoantigens, should be considered for inclusion in the variant reference database [113]. For example, the OncoSplicing database integrated all alternative splicing events of cancers derived from the TCGA dataset, which can be considered for inclusion [114]. Unfortunately, there is no available reference sequence library (FASTA format) for alternative splicing variants.

An alternative method involves constructing a personalized database for each patient using proteomeGenerator [115, 116]. This approach is viable only if the patient's tumor samples undergo both RNA sequencing and proteomic analysis concurrently. The tool generates the reference database from RNA sequences, following a two-step process: first, inferring

Table 3. Different methods for de novo peptide sequencing

Name	Algorithm	Published year	Focus on missing peaks	Still available	Reference
Peaks	Dynamic Programming	2003	No	Yes, software	Ref. [117]
NovoHMM	Hidden Markov	2005	No	No	Ref. [118]
PepNovo	Spectrum Graph Theory	2005	No	No	Ref. [119]
pNovo	Spectrum Graph Theory	2010	No	No	Ref. [120]
Novor	Decision Tree	2015	No	Yes, software	Ref. [121]
DeepNovo	CNN + RNN	2017	No	Yes, code	Ref. [122]
PointNovo	RNN + PointNet	2021	No	Yes, code	Ref. [123]
Casanovo	Transformer	2022	No	Yes, code	Ref. [124]
GraphNovo	Graphormer	2023	Yes	Yes, code	Ref. [125]
Spectralis	CNN	2024	Yes	Yes, code	Ref. [126]

protein sequences from the RNA data, and second, employing protein identification tools (e.g. MaxQuant) for peptide matching and protein identification. The findings indicate that the self-reference database can identify a greater number of non-classical peptides compared to traditional reference databases (e.g. Uniprot). However, this method must also discern cancer-specific sequences, necessitating a control design in experiments. Consequently, this approach is associated with higher economic and computational costs.

Overall, the primary limitation of database search methods is the completeness of the variation reference database. Given the significant individual differences, relying on a single variant database is impractical. Developing a comprehensive database that includes both DNA and RNA variant information is essential. A more effective approach involves using paired RNA-seq data to construct a personalized self-reference database, which can account for individual differences and include transcriptome variant information. That allows for more precise detection of neoantigens.

De novo peptide sequencing methods based on AI for proteomics: state-of-the-art

With the rapid evolution of AI, de novo peptide sequencing methods have also seen significant growth, introducing new concepts to proteomics. Unlike traditional methods, de novo peptide sequencing does not depend on a reference database but generates sequences directly from the spectrum [39]. In this section, we will provide a detailed discussion of several representative tools (Table 3).

Classic machine learning used for de novo peptide sequencing

The innovation in AI technology is closely linked to the development of de novo peptide sequencing tools, with both fields complementing each other. In the early stages of de novo peptide sequencing methods, traditional machine learning or statistical models were employed to analyze MS data (Fig. 3A). Representative tools include Peaks, NovoHMM, PepNovo, and pNovo [117–120]. These methods typically use a spectrum graph or a modified approach and generally consist of two main steps. First, the original spectrum is transformed into a directed acyclic graph. Second, dynamic programming algorithms are applied to identify the optimal paths.

Another classical approach, Novor [121], employs two large decision trees for de novo peptide sequencing. Specifically, it designs a new scoring function to evaluate the quality of the match between a peptide sequence and the input spectrum. This scoring function utilizes one decision tree to learn its thousands of parameters from a large peptide spectral library containing over 300 000 spectra.

The results showed that Novor achieved the highest recall (0.548, 0.569, 0.411, and 0.635) in a test on four test datasets, outperforming the commercial software Peaks. Detailed reviews of machine learning-based de novo peptide sequencing methods have been extensively covered elsewhere, so we will not reiterate them here [127–130]. However, it should be noted that only Peaks and Novor offer available software currently, while the other tools are no longer in service.

Deep learning methods for de novo peptide sequencing

Due to the inherent complexity of MS data, traditional machine learning methods often fail to capture more nuanced features effectively. In contrast, deep learning offers more sophisticated networks that are better suited to handle such complex tasks [131].

For instance, DeepNovo [122] demonstrated significant improvements in peptide sequencing precision, surpassing traditional machine learning approaches like Novor and Peaks. The area under the curve of DeepNovo was 18.8%–50.0% higher than Peaks, 7.7%–34.4% higher than Novor. The model is primarily composed of two components: a convolutional neural network (CNN) [132] module for extracting MS features and a recurrent neural network (RNN) [133] for decoding these features into sequences. This combination of CNN and RNN mirrors the integration of image recognition (e.g. spectra) with natural language processing (e.g. protein sequences). Given the superior performance of long short-term memory (LSTM) networks in sequence processing, Qiao et al. introduced PointNovo, which integrates LSTM and PointNet [123, 134, 135]. They evaluated the performance of DeepNovo and PointNovo using MS data from nine species. The results indicated that PointNovo consistently outperformed DeepNovo at the peptide level by a margin ranging from 13.01% to 23.95%. Furthermore, their experiments showed a notable decrease in performance when LSTM was omitted, underscoring its critical role in the model.

In 2017, Google introduced the transformer architecture, which revolutionized traditional sequence-processing methods [136]. The transformer model consists of an encoder and a decoder, incorporating a self-attention mechanism to compute correlations between sequences in parallel. By integrating positional encoding, it can simultaneously calculate the attention between each position and all other positions, thereby capturing global dependencies without information loss. Unlike LSTM, which processes sequentially and may lose information in

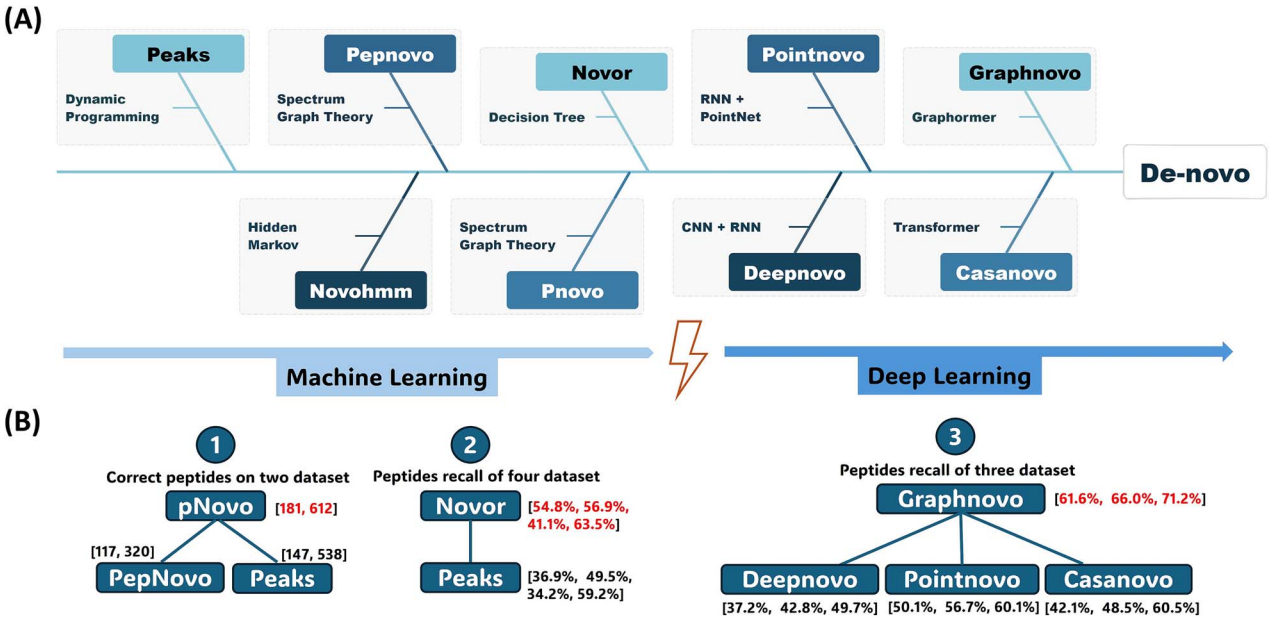

Figure 3. The co-evolution trajectory in the development of both the peptide de novo sequencing methods and AI algorithms. (A) the peptide de novo sequencing methods have undergone a revolution from machine learning algorithms to deep learning algorithms (from left to right). (B) Performance of the peptide de novo sequencing methods on different datasets: (i) the number of correct peptides identified by pNovo [120] was 181 and 612 in the two datasets, which was better than PepNovo and peaks. (ii) Novor [121] outperforms peaks in recall on four datasets; (iii) Graphnovo [125] outperforms other methods in recall on all three datasets.

long sequences, transformers excel in capturing long-range dependencies. Consequently, the transformer has become the dominant method for sequence tasks, with numerous successful applications in biological sequence transformation [137–139]. For instance, Yilmaz et al. developed a powerful de novo peptide sequencing method, Casanovo [124], using the transformer framework, trained on approximately 300,000 unique peptide sequences. On nine test datasets, Casanovo outperformed previous models, with a mean improvement of 0.373 and 0.310 in precision relative to DeepNovo and PointNovo, respectively. Notably, Casanovo exhibits faster inference speed, processing 119 spectra per second on an RTX 2080, compared to DeepNovo's reported rate of 36 spectra per second and PointNovo's 20 spectra per second on the more powerful RTX 2080 Ti. This advantage is attributed to the advanced architecture of the transformer.

Missing fragmentation and contamination are common issues in MS data generation. To address missing peaks, Mao et al. proposed GraphNovo [125], a two-stage de novo peptide sequencing algorithm based on a graph neural network. This algorithm comprises two components: GraphNovo-PathSearcher and GraphNovo-SeqFiller. It focuses on finding the optimal path (PathSearcher) in the first stage to guide the sequence prediction (SeqFiller) in the second stage. PathSearcher generates a node sequence (the source and target nodes), while SeqFiller outputs an amino acid (AA) sequence. GraphNovo mitigates the missing fragmentation problem primarily through the PathSearcher component. The results showed that GraphNovo achieved an overall recall of 0.786, compared to 0.61 for DeepNovo and 0.68 for PointNovo in missing peak data. When PathSearcher was used to modify DeepNovo and PointNovo with the optimal path, their recall improved to 0.742 and 0.760, respectively, approaching the performance of GraphNovo.

A recently published tool, Spectralis [126], also addresses the problem of missing peak fragments but employs a different approach. It uses bin class predictions to improve peptide-spectrum matches. Specifically, it creates discrete bins of 1 Dalton (Da) for sliding windows and introduces the AA-gapped convolutional layer to recover the peptide sequence by reading out the m/z differences of either series. The bin class prediction was used to generate the Spectralis-score, which was then applied to rank the existing predicted peptides. Additionally, the bin reclassification model can correct predictions from Novor and Casanovo, expanding the utility of these tools.

Method comparisons in a case study on liver cancer neoantigen identification from mass spectrometry data

To demonstrate the potential of proteomics in neoantigen identification, we conducted a peptide identification analysis using MS data from the liver cancer cell line HepG2. The analysis was performed with MaxQuant software, following parameter settings consistent with those reported in the original publication [40]. Notably, we utilized a self-constructed reference database containing 262,654 mutant protein sequences derived from all liver cancer-related mutations in the Cosmic database (see Methods and materials).

If a sequence is not matched in the original reference but successfully matches in the mutated sequence database, it is considered a neoantigen candidate. Our analysis identified a total of 237 candidates.

We examined the mutation information of HepG2 (DepMap ID: ACH-000739) in the CCLE database and found 152 mutation entries [140]. Surprisingly, none of the identified neoantigen candidates overlapped with the known mutation sites in HepG2. Further examination revealed that only two mutation entries (NRAS_ p.Q61L and PREX2_ p.L50V) were included in our mutant protein sequences database. Even the world's largest and most comprehensive databases of somatic mutations in cancer cannot meet the need for personalized neoantigen identification. Nevertheless, we observed 237 "new sequences" at the protein level

that could not be captured by WES, underscoring the potential of proteomics in neoantigen discovery.

To explore the differences between the database retrieval approach and the de novo sequencing approach, we analyzed the same data using both methods. We collected pooled samples from nine HCC patients, each subjected to three technical replicates. The MS data were divided into eight parts each time, resulting in 24 files. [40].

For the database search method, we used MaxQuant with parameters consistent with the original study [40], employing a mutant protein sequences database generated from the Cosmic (see Methods and materials). For de novo peptide sequencing, we used Casanovo. The final number of neoantigen candidates identified was 329 by MaxQuant and 252 by Casanovo (Fig. 4A). Only 13 candidates were common between the two methods (Fig. 4B). This result exceeded our expectations: the de novo method identified fewer candidates than the database search method, and their consistency was poor.

Further analysis of Casanovo results showed that only six mutant genes were detected consistently across all three technical replicates (Fig. 4C). The results indicate poor consistency between technical replicate samples. Based on these findings, we have concerns about the accuracy of Casanovo's results. Additionally, we compiled a list of high-frequency mutated genes in HCC from the Cosmic database and multiple study cohorts [141–144]. The high-frequency mutated genes represent that they are more likely to be detected in the population of HCC. The mutation genes in the list are sorted by the frequency of detection in the population from high to low. Using MaxQuant, we detected only two high-frequency mutated genes (CPS1 and TNN) in nine HCC patients, which seems unreasonable given the expected mutation frequency. In contrast, Casanovo identified nine high-frequency mutated genes, more consistent with the observed mutation frequency in HCC populations (Fig. 4D). This demonstrates the advantage of de novo peptide sequencing in identifying neoantigens without relying on reference databases.

Whether using a de novo or database search method, we checked their performance using the same liver cancer-related variant reference database. MaxQuant, after strict false discovery rate (FDR) control and filtering, identified 94 775 peptides. FDR is typically controlled by incorporating decoy sequences into the database, created by reversing target sequences and using their scores to model probabilistic functions [145]. In contrast, Casanovo lacks any evaluation mechanism to assess its accuracy, resulting in 1 325 364 peptides. For the de novo method, even a single amino acid error in the predicted peptide is fatal, and such a peptide will not be considered. Only 252 neoantigen candidates were matched in the reference database. Despite its limited search space, MaxQuant identified 329 candidates with higher quality. Therefore, it is essential to develop and evaluate new proteomic pipelines to improve neoantigen identification accuracy.

Future perspective

Based on the results of the above comparative analysis, we propose an improved workflow to enhance the application of proteomics in neoantigen identification. Database-based search methods generally offer high confidence but rely heavily on a comprehensive reference sequence library. This can be achieved by integrating large cancer-related mutation databases, such as Cosmic, Cbioportal, CanproVar, and OncoSplicing [107, 114, 141, 146]. When only MS data are available, neoantigen identification depends on a comprehensive reference sequence library of cancer variants, which may limit the number of identified neoantigens. However, if both MS and RNA-seq paired data are available for the same patient, a self-reference database can be constructed, allowing for more accurate neoantigen identification. In this scenario, while the theoretical spectrum step limits the peptide space, a moderate number of neoantigens can still be obtained.

In contrast, de novo peptide sequencing methods do not rely on reference databases and generate sequences directly from MS data. However, the accuracy of these generated sequences requires further evaluation. Therefore, we propose using both MS and RNA-seq paired data for neoantigen identification in de novo peptide sequencing. This approach allows self-reference sequences derived from RNA-seq data to serve as ground truth, thereby improving prediction confidence. Consequently, more high-confidence neoantigens can be identified (Fig. 5). Specifically, peptides predicted by de novo peptide sequencing tools are reverse-translated into codon sequences. Based on the patient's paired RNA-seq data, a cancer-specific sequence reference library is generated. The codon sequences are then aligned with the reference sequences to identify candidate neoantigens. The advantage of this method is that it bypasses open reading frame prediction and other steps, saving time and enhancing accuracy.

Nanopore sequencing is an advanced technique that enables real-time sequencing by inferring molecular composition based on changes in electrical current as single molecules pass through biological nanopores [147, 148]. It has been successfully applied in DNA and RNA sequencing and has achieved commercialization [149]. However, proteins, composed of 20 different types of amino acids, have much higher structural complexity, making nanopore protein sequencing significantly more challenging than nucleic acid sequencing [150].

In 2021, Brinkerhoff *et al.* developed a method to pull a DNA-peptide conjugate through a biological nanopore using a helicase, successfully sequencing a synthetic peptide of 26 amino acids with an average accuracy of 87% [151]. This method can repeatedly measure the same peptide segment and distinguish single amino acid changes in the protein sequence. Recent research by Motone *et al.* has demonstrated the feasibility of full-length sequencing of complete proteins using nanopore technology, pushing the boundaries of this field [152]. In the near future, we anticipate that this technology will be applied to production practices, potentially revolutionizing traditional proteomics and enhancing its role in neoantigen detection.

Discussion

Cancer vaccines, as emerging treatment methods, have garnered significant attention in recent years, with numerous clinical trials currently underway [153, 154]. However, the current screening process for neoantigens has limitations, resulting in only a small number of candidates demonstrating clinical efficacy [28]. The emergence of proteomics offers new hope for neoantigen screening in cancer vaccines. Here, we propose a novel and feasible workflow for identifying neoantigens through proteomics, grounded in both theoretical considerations and practical applications.

For database search methods, constructing a reference database is a crucial step. Two approaches can be used: a comprehensive cancer variant reference database and a self-reference database. A comprehensive cancer variant reference database requires prior knowledge of mutation sequences relevant to human cancers. Given the vastness of the human genome, it

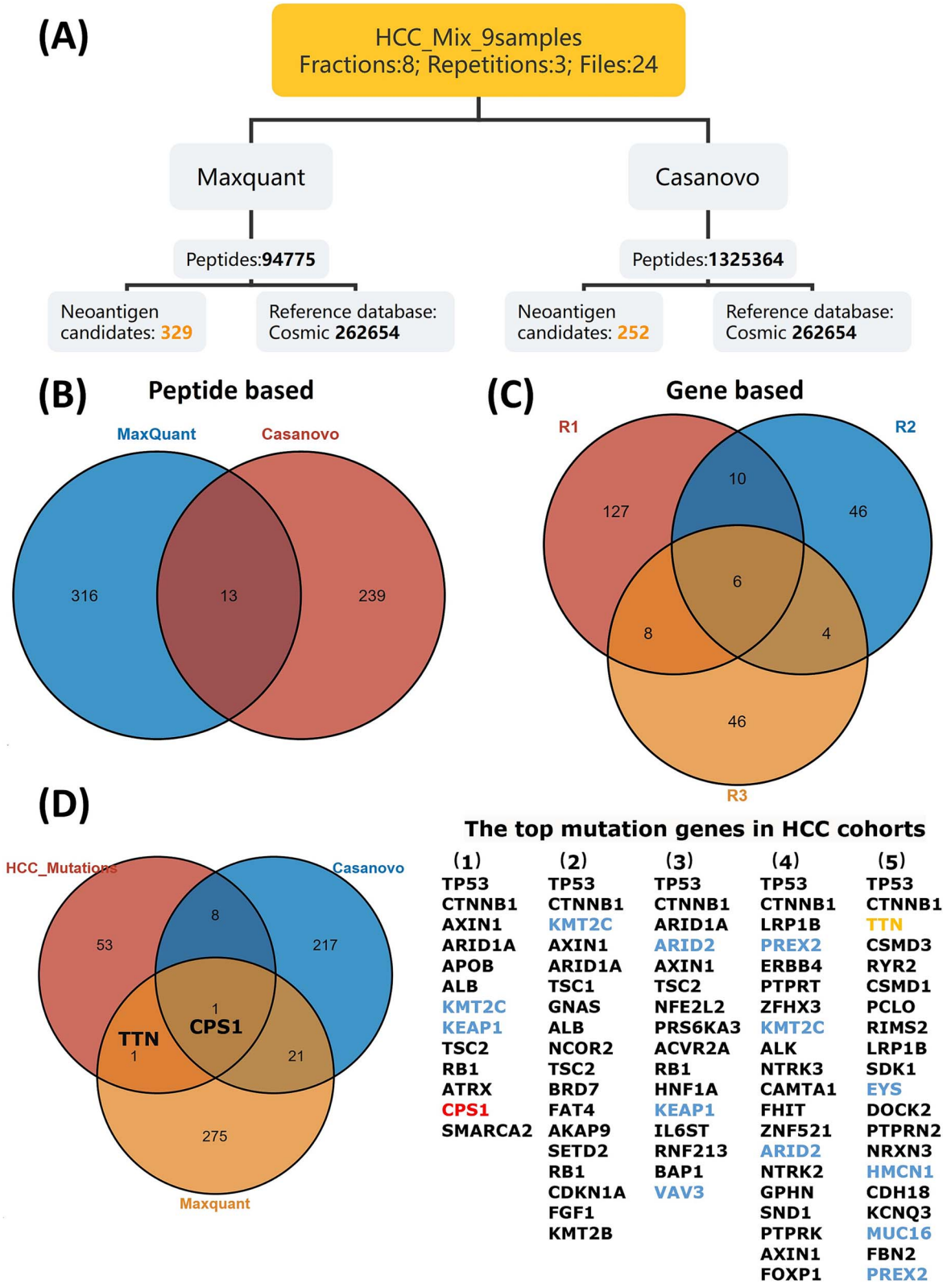

Figure 4. A case study of neoantigen identification in HCC patients. (A) Sample information on liver cancer and the workflow of neoantigen identification. (B) Venn diagram of identified neoantigen candidates under two methods. (C) Venn diagram of identified genes in three technical replicates by Casanovo. (D) The detection of high-frequency mutant genes in HCC under two methods. The list on the right shows the highly mutated genes in different HCC cohorts.

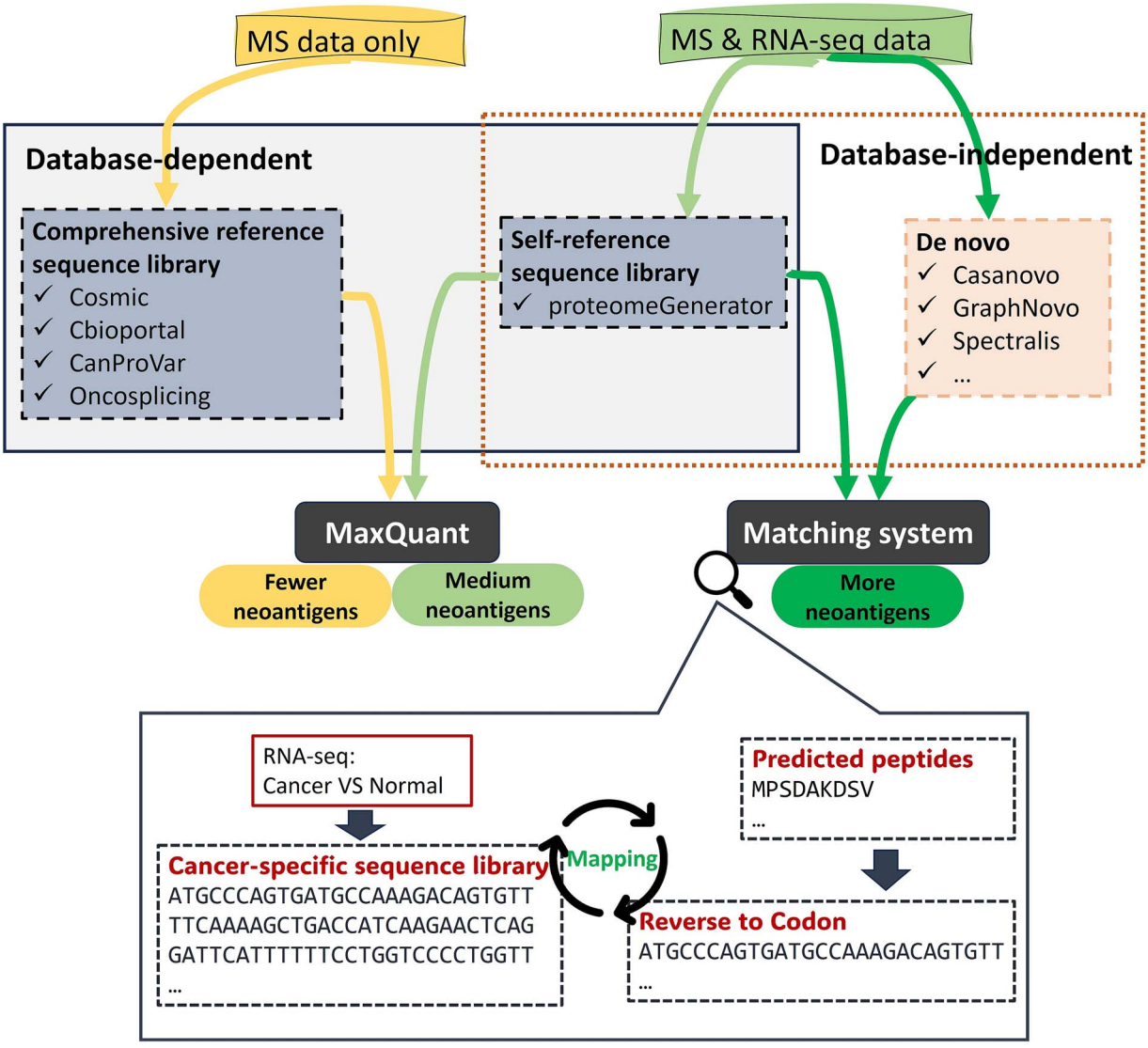

Figure 5. The proteomics-based neoantigen identification workflow. The process of neoantigen identification based on MS includes both database-dependent and database-independent methods. Depending on the type of data input, different analysis processes are selected, resulting in varying numbers of identified neoantigen candidates. The method of database search depends on a comprehensive reference sequence library or self-reference sequence library (RNA-seq data). Database-independent methods require a self-reference sequence library as validation: The sequences obtained by reverse transcription of the predicted peptides were matched with the cancer-specific sequence library obtained from RNA-seq data to obtain neoantigen candidates.

is nearly impossible to establish an accurate and comprehensive mutation-related information database. This limitation hinders personalized neoantigen screening for patients. For example, when identifying neoantigens using MS on HepG2 cells, only two mutation entries for HepG2 are recorded in the COSMIC database. Therefore, using the patient's paired RNA-seq data to establish a self-reference database is more relevant and feasible.

De novo peptide sequencing methods not only address the issue of database construction but also have the potential to identify peptide sequences beyond prior knowledge, making them more promising for neoantigen identification tasks [39]. As AI evolves, various de novo peptide sequencing methods have been developed alongside the emergence of different algorithms. In the early stages, traditional machine learning algorithms were prominent, with tools such as PEAKS, NovoHMM, PepNovo, pNovo, and Novor.

With the rise of deep learning, researchers have found that deep learning methods are more capable of handling complex tasks compared to traditional machine learning methods [131, 155, 156]. The advent of various deep neural network architectures, such as CNN, RNN, and transformer, has led to the development of corresponding tools, including DeepNovo, PointNovo, and Casanovo. In addition to iterative algorithm updates, it is crucial to address specific real-world problems. For instance, tools like GraphNovo and Spectralis have been designed to tackle the issue of missing fragments in MS data, significantly enhancing the accuracy of de novo peptide sequencing predictions. Recent deep learning-based de novo peptide sequencing tools offer not only high prediction accuracy but also trainable pre-trained models. This flexibility allows users to adjust the model for different scenarios and data characteristics, leading to better prediction results. However, compared to the commercial software, these

deep learning-based de novo peptide tools are not user-friendly. They require substantial computing resources, and users must have a solid understanding of deep learning to apply them effectively.

Meanwhile, an analysis of liver cancer cases using both database search and de novo peptide sequencing methods reveals that the de novo peptide sequencing approach holds a higher potential for cancer neoantigen identification. This approach addresses the issue of mutation database incompleteness and can generate peptide segments beyond existing knowledge. However, it exhibited poor reproducibility in the three technical replicate samples. That could be attributed to the inherent limitations of the DDA technique or issues with the accuracy of the de novo peptide sequencing method. At present, there is no systematic study assessing the performance differences between DDA and DIA in detecting neoantigens, which represents a promising area for future research. Additionally, these de novo peptide sequencing models exhibit significant performance variations across different species' MS data (Fig. 3B). For instance, Casanovo performs the worst on the human dataset in a test set of nine species [124]. Therefore, it is necessary to fine-tune these models using high-quality human datasets to enhance their performance in human MS data. Since the analyzed samples do not provide ground truth, we cannot assess their accuracy. Therefore, we propose constructing a self-reference database using the patient's paired RNA-seq data as a baseline to aid in neoantigen identification (Fig. 5). Nonetheless, we remain optimistic that the accuracy of de novo peptide sequencing methods will overcome current challenges, enabling better application in neoantigen identification.

Conclusion

To elucidate the potential applications of proteomics in neoantigen identification, this paper reviewed the representative methods and tools for proteomic analysis, and their development trends, and recommended reliable analytical strategies and tools for neoantigen identification. Additionally, we compared these methods in a case study analysis on HepG2 cell line and nine mixed liver cancer proteomics samples to demonstrate the potential of proteomics in neoantigen identification. This analysis also uncovered some limitations of existing methods, so we proposed an improved, feasible analytical workflow for neoantigen identification. In the future, proteomics is suggested to be integrated into standard neoantigen identification pipelines to enhance the efficiency of neoantigen screening, thereby facilitating the maturation of clinical cancer vaccines.

Key Points

- We offer a comprehensive overview of proteomics in the discovery of neoantigens, along with detailed identification methodologies and tool recommendations.
- We survey recent developments in de novo peptide sequencing methods, highlighting their potential in neoantigen identification.
- We provide a comparative analysis of the use of proteomics for neoantigen identification, demonstrate the potential and existing drawbacks of proteomics in neoantigen identification, and propose a novel workflow.

Acknowledgements

I would like to thank ServierMedicalArt for providing the materials used to create the figures.

Conflict of interest: None declared.

Funding

This work was supported by the start-up funding grants for new staff (J.L.) at the Shenzhen University of Advanced Technology.

Data availability

None declared.

References

1. Moore L, Cagan A, Coorens THH. *et al.* The mutational landscape of human somatic and germline cells. *Nature* 2021;**597**:381–6. https://doi.org/10.1038/s41586-021-03822-7
2. Seferbekova Z, Lomakin A, Yates LR. *et al.* Spatial biology of cancer evolution. *Nat Rev Genet* 2023;**24**:295–313. https://doi.org/10.1038/s41576-022-00553-x
3. Harrington KJ, Nenclares P. The biology of cancer. *Medicine* 2023;**51**:1–6. https://doi.org/10.1016/j.mpmed.2022.10.001
4. Yarchoan M, Johnson BA 3rd, Lutz ER. *et al.* Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer* 2017;**17**:209–22. https://doi.org/10.1038/nrc.2016.154
5. Hacohen N, Fritsch EF, Carter TA. *et al.* Getting personal with neoantigen-based therapeutic cancer vaccines. *Cancer Immunol Res* 2013;**1**:11–5. https://doi.org/10.1158/2326-6066.Cir-13-0022
6. Jhunjhunwala S, Hammer C, Delamarre L. Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion. *Nat Rev Cancer* 2021;**21**:298–312. https://doi.org/10.1038/s41568-021-00339-z
7. Chen F, Zou Z, Du J. *et al.* Neoantigen identification strategies enable personalized immunotherapy in refractory solid tumors. *J Clin Invest* 2019;**129**:2056–70. https://doi.org/10.1172/jci99538
8. Xie N, Shen G, Gao W. *et al.* Neoantigens: promising targets for cancer therapy. *Signal Transduction Targeted Ther* 2023;**8**:1–38. https://doi.org/10.1038/s41392-022-01270-x
9. Smith CC, Selitsky SR, Chai S. *et al.* Alternative tumour-specific antigens. *Nat Rev Cancer* 2019;**19**:465–78. https://doi.org/10.1038/s41568-019-0162-4
10. Ward JP, Gubin MM, Schreiber RD. The role of neoantigens in naturally occurring and therapeutically induced immune responses to cancer. *Adv Immunol* 2016;**130**:25–74. https://doi.org/10.1016/bs.ai.2016.01.001
11. Ping Y, Liu C, Zhang Y. T-cell receptor-engineered T cells for cancer treatment: current status and future directions. *Protein Cell* 2018;**9**:254–66. https://doi.org/10.1007/s13238-016-0367-1
12. Lybaert L, Lefever S, Fant B. *et al.* Challenges in neoantigen-directed therapeutics. *Cancer Cell* 2023;**41**:15–40. https://doi.org/10.1016/j.ccell.2022.10.013
13. Awad MM, Govindan R, Balogh KN. *et al.* Personalized neoantigen vaccine NEO-PV-01 with chemotherapy and anti-PD-1 as first-line treatment for non-squamous non-small cell lung cancer. *Cancer Cell* 2022;**40**:1010–1026.e11. https://doi.org/10.1016/j.ccell.2022.08.003
14. Lorentzen CL, Haanen JB, Met Ö. *et al.* Clinical advances and ongoing trials on mRNA vaccines for cancer

treatment. *Lancet Oncol* 2022;**23**:e450–8. https://doi.org/10.1016/s1470-2045(22)00372-2
15. Lin MJ, Svensson-Arvelund J, Lubitz GS. *et al.* Cancer vaccines: the next immunotherapy Frontier. *Nat Cancer* 2022;**3**:911–26. https://doi.org/10.1038/s43018-022-00418-6
16. Yarchoan M, Gane EJ, Marron TU. *et al.* Personalized neoantigen vaccine and pembrolizumab in advanced hepatocellular carcinoma: a phase 1/2 trial. *Nat Med* 2024;**30**:1044–53. https://doi.org/10.1038/s41591-024-02894-y
17. Chen H, Li Z, Qiu L. *et al.* Personalized neoantigen vaccine combined with PD-1 blockade increases CD8(+) tissue-resident memory T-cell infiltration in preclinical hepatocellular carcinoma models. *J Immunother Cancer* 2022;**10**:e004389. https://doi.org/10.1136/jitc-2021-004389
18. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nat Rev Clin Oncol* 2021;**18**:215–29. https://doi.org/10.1038/s41571-020-00460-2
19. Olson ND, Wagner J, Dwarshuis N. *et al.* Variant calling and benchmarking in an era of complete human genome sequences. *Nat Rev Genet* 2023;**24**:464–83. https://doi.org/10.1038/s41576-023-00590-0
20. Pei S, Liu T, Ren X. *et al.* Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Brief Bioinform* 2021;**22**:bbaa148. https://doi.org/10.1093/bib/bbaa148
21. Hundal J, Kiwala S, McMichael J. *et al.* pVACtools: a computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol Res* 2020;**8**:409–20. https://doi.org/10.1158/2326-6066.Cir-19-0401
22. Zhou Z, Wu J, Ren J. *et al.* TSNAD v2.0: a one-stop software solution for tumor-specific neoantigen detection. *Comput Struct Biotechnol J* 2021;**19**:4510–6. https://doi.org/10.1016/j.csbj.2021.08.016
23. Richters MM, Xia H, Campbell KM. *et al.* Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome Med* 2019;**11**:56. https://doi.org/10.1186/s13073-019-0666-2
24. Robbins PF, Lu YC, El-Gamil M. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med* 2013;**19**:747–52. https://doi.org/10.1038/nm.3161
25. Yadav M, Jhunjhunwala S, Phung QT. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* 2014;**515**:572–6. https://doi.org/10.1038/nature14001
26. McGranahan N, Furness AJ, Rosenthal R. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 2016;**351**:1463–9. https://doi.org/10.1126/science.aaf1490
27. Bobisse S, Genolet R, Roberti A. *et al.* Sensitive and frequent identification of high avidity neo-epitope specific CD8 (+) T cells in immunotherapy-naive ovarian cancer. *Nat Commun* 2018;**9**:1092. https://doi.org/10.1038/s41467-018-03301-0
28. Wells DK, van Buuren MM, Dang KK. *et al.* Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* 2020;**183**:818–834.e13. https://doi.org/10.1016/j.cell.2020.09.015
29. Chen L, Qin D, Guo X. *et al.* Putting proteomics into immunotherapy for glioblastoma. *Front Immunol* 2021;**12**:593255. https://doi.org/10.3389/fimmu.2021.593255
30. Sethi MK, Hancock WS, Fanayan S. Identifying N-glycan biomarkers in colorectal cancer by mass spectrometry. *Acc Chem Res* 2016;**49**:2099–106. https://doi.org/10.1021/acs.accounts.6b00193
31. Mani DR, Krug K, Zhang B. *et al.* Cancer proteogenomics: current impact and future prospects. *Nat Rev Cancer* 2022;**22**:298–313. https://doi.org/10.1038/s41568-022-00446-5
32. Cheung CHY, Juan HF. Quantitative proteomics in lung cancer. *J Biomed Sci* 2017;**24**:37. https://doi.org/10.1186/s12929-017-0343-y
33. Vizcaíno JA, Deutsch EW, Wang R. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 2014;**32**:223–6. https://doi.org/10.1038/nbt.2839
34. Vizcaíno JA, Csordas A. del-Toro N et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* 2016;**44**:D447–56. https://doi.org/10.1093/nar/gkv1145
35. Zhou C, Zhu C, Liu Q. Toward in silico identification of tumor neoantigens in immunotherapy. *Trends Mol Med* 2019;**25**:980–92. https://doi.org/10.1016/j.molmed.2019.08.001
36. Polyakova A, Kuznetsova K, Moshkovskii S. Proteogenomics meets cancer immunology: mass spectrometric discovery and analysis of neoantigens. *Expert Rev Proteomics* 2015;**12**:533–41. https://doi.org/10.1586/14789450.2015.1070100
37. Verma A, Halder A, Marathe S. *et al.* A proteogenomic approach to target neoantigens in solid tumors. *Expert Rev Proteomics* 2020;**17**:797–812. https://doi.org/10.1080/14789450.2020.1881889
38. Ren Y, Yue Y, Li X. *et al.* Proteogenomics offers a novel avenue in neoantigen identification for cancer immunotherapy. *Int Immunopharmacol* 2024;**142**:113147. https://doi.org/10.1016/j.intimp.2024.113147
39. Karunratanakul K, Tang HY, Speicher DW. *et al.* Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Mol Cell Proteomics* 2019;**18**:2478–91. https://doi.org/10.1074/mcp.TIR119.001656
40. Wang M, Weng S, Li C. *et al.* Proteomic overview of hepatocellular carcinoma cell lines and generation of the spectral library. *Sci Data* 2022;**9**:732. https://doi.org/10.1038/s41597-022-01845-x
41. Wölfel T, Hauer M, Schneider J. *et al.* A p16INK4a-insensitive CDK4 mutant targeted by cytolytic T lymphocytes in a human melanoma. *Science* 1995;**269**:1281–4. https://doi.org/10.1126/science.7652577
42. Coulie PG, Lehmann F, Lethé B. *et al.* A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma. *Proc Natl Acad Sci U S A* 1995;**92**:7976–80. https://doi.org/10.1073/pnas.92.17.7976
43. Brändle D, Brasseur F, Weynants P. *et al.* A mutated HLA-A2 molecule recognized by autologous cytotoxic T lymphocytes on a human renal cell carcinoma. *J Exp Med* 1996;**183**:2501–8. https://doi.org/10.1084/jem.183.6.2501
44. Hogan KT, Eisinger DP, Cupp SB 3rd. *et al.* The peptide recognized by HLA-A68.2-restricted, squamous cell carcinoma of the lung-specific cytotoxic T lymphocytes is derived from a mutated elongation factor 2 gene. *Cancer Res* 1998;**58**:5144–50.
45. Satam H, Joshi K, Mangrolia U. *et al.* Next-generation sequencing technology: current trends and advancements. *Biology (Basel)* 2023;**12**:997. https://doi.org/10.3390/biology12070997
46. Sellars MC, Wu CJ, Fritsch EF. Cancer vaccines: building a bridge over troubled waters. *Cell* 2022;**185**:2770–88. https://doi.org/10.1016/j.cell.2022.06.035
47. Saxena M, van der Burg SH, Melief CJM. *et al.* Therapeutic cancer vaccines. *Nat Rev Cancer* 2021;**21**:360–78. https://doi.org/10.1038/s41568-021-00346-0

48. Szolek A, Schubert B, Mohr C. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 2014;**30**:3310–6. https://doi.org/10.1093/bioinformatics/btu548
49. Shukla SA, Rooney MS, Rajasagi M. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol* 2015;**33**:1152–8. https://doi.org/10.1038/nbt.3344
50. Bai Y, Wang D, Fury W. PHLAT: inference of high-resolution HLA types from RNA and whole exome sequencing. *Methods Mol Biol* 2018;**1802**:193–201. https://doi.org/10.1007/978-1-4939-8546-3_13
51. Matey-Hernandez ML, Brunak S, Izarzugaza JMG. Benchmarking the HLA typing performance of Polysolver and Optitype in 50 Danish parental trios. *BMC Bioinf* 2018;**19**:239. https://doi.org/10.1186/s12859-018-2239-6
52. Bassani-Sternberg M, Chong C, Guillaume P. *et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol* 2017;**13**:e1005725. https://doi.org/10.1371/journal.pcbi.1005725
53. Reynisson B, Alvarez B, Paul S. *et al.* NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* 2020;**48**:W449–w454. https://doi.org/10.1093/nar/gkaa379
54. O'Donnell T, Rubinsteyn A. High-throughput MHC I ligand prediction using MHCflurry. *Methods Mol Biol* 2020;**2120**:113–27. https://doi.org/10.1007/978-1-0716-0327-7_8
55. Schmidt J, Smith AR, Magnin M. *et al.* Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep Med* 2021;**2**:100194. https://doi.org/10.1016/j.xcrm.2021.100194
56. Kim JY, Cha H, Kim K. *et al.* MHC II immunogenicity shapes the neoepitope landscape in human tumors. *Nat Genet* 2023;**55**:221–31. https://doi.org/10.1038/s41588-022-01273-y
57. Peng X, Lei Y, Feng P. *et al.* Characterizing the interaction conformation between T-cell receptors and epitopes with deep learning. *Nat Mach Intell* 2023;**5**:395–407. https://doi.org/10.1038/s42256-023-00634-4
58. Bjerregaard AM, Nielsen M, Hadrup SR. *et al.* MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol Immunother* 2017;**66**:1123–30. https://doi.org/10.1007/s00262-017-2001-3
59. Kodysh J, Rubinsteyn A. OpenVax: an open-source computational pipeline for cancer neoantigen prediction. *Methods Mol Biol* 2020;**2120**:147–60. https://doi.org/10.1007/978-1-0716-0327-7_10
60. Schenck RO, Lakatos E, Gatenbee C. *et al.* NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline. *BMC Bioinf* 2019;**20**:264. https://doi.org/10.1186/s12859-019-2876-4
61. Tang Y, Wang Y, Wang J. *et al.* TruNeo: an integrated pipeline improves personalized true tumor neoantigen identification. *BMC Bioinf* 2020;**21**:532. https://doi.org/10.1186/s12859-020-03869-9
62. Diao K, Chen J, Wu T. *et al.* Seq2Neo: a comprehensive pipeline for cancer neoantigen immunogenicity prediction. *Int J Mol Sci* 2022;**23**:11624. https://doi.org/10.3390/ijms231911624
63. Katsonis P, Koire A, Wilson SJ. *et al.* Single nucleotide variations: biological impact and theoretical interpretation. *Protein Sci* 2014;**23**:1650–66. https://doi.org/10.1002/pro.2552
64. Lin M, Whitmire S, Chen J. *et al.* Effects of short indels on protein structure and function in human genomes. *Sci Rep* 2017;**7**:9313. https://doi.org/10.1038/s41598-017-09287-x
65. Mullaney JM, Mills RE, Pittard WS. *et al.* Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet* 2010;**19**:R131–6. https://doi.org/10.1093/hmg/ddq400
66. Latysheva NS, Babu MM. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res* 2016;**44**:4487–503. https://doi.org/10.1093/nar/gkw282
67. Ule J, Blencowe BJ. Alternative splicing regulatory networks: functions, mechanisms, and evolution. *Mol Cell* 2019;**76**:329–45. https://doi.org/10.1016/j.molcel.2019.09.017
68. Djebali S, Davis CA, Merkel A. *et al.* Landscape of transcription in human cells. *Nature* 2012;**489**:101–8. https://doi.org/10.1038/nature11233
69. Khurana E, Fu Y, Chakravarty D. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* 2016;**17**:93–108. https://doi.org/10.1038/nrg.2015.17
70. Wolters DA, Washburn MP, Yates JR 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 2001;**73**:5683–90. https://doi.org/10.1021/ac010617e
71. Link AJ, Eng J, Schieltz DM. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* 1999;**17**:676–82. https://doi.org/10.1038/10890
72. Duong VA, Lee H. Bottom-up proteomics: advancements in sample preparation. *Int J Mol Sci* 2023;**24**:5350. https://doi.org/10.3390/ijms24065350
73. Eng JK, Searle BC, Clauser KR. *et al.* A face in the crowd: recognizing peptides through database search. *Mol Cell Proteomics* 2011;**10**:R111.009522. https://doi.org/10.1074/mcp.R111.009522
74. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–d531. https://doi.org/10.1093/nar/gkac1052
75. Harrison PW, Amode MR, Austine-Orimoloye O. *et al.* Ensembl 2024. *Nucleic Acids Res* 2024;**52**:D891–d899. https://doi.org/10.1038/s41587-025-02590-3
76. Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteomics data. *BMC Syst Biol* 2014;**8**, **Suppl 2(Suppl 2)**:S3. https://doi.org/10.1186/1752-0509-8-s2-s3
77. Li X, Wang W, Chen J. Recent progress in mass spectrometry proteomics for biomedical research. *Sci China Life Sci* 2017;**60**:1093–113. https://doi.org/10.1007/s11427-017-9175-2
78. Gu Y, Guo Y, Gao N. *et al.* The proteomic characterization of the peritumor microenvironment in human hepatocellular carcinoma. *Oncogene* 2022;**41**:2480–91. https://doi.org/10.1038/s41388-022-02264-3
79. Wühr M, Haas W, McAlister GC. *et al.* Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal Chem* 2012;**84**:9214–21. https://doi.org/10.1021/ac301962s
80. Ow SY, Cardona T, Taton A. *et al.* Quantitative shotgun proteomics of enriched heterocysts from Nostoc sp. PCC 7120 using 8-plex isobaric peptide tags. *J Proteome Res* 2008;**7**:1615–28. https://doi.org/10.1021/pr700604v
81. Stahl DC, Swiderek KM, Davis MT. *et al.* Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J Am Soc Mass Spectrom* 1996;**7**:532–40. https://doi.org/10.1016/1044-0305(96)00057-8

82. Venable JD, Dong MQ, Wohlschlegel J. *et al.* Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* 2004;**1**:39–45. https://doi.org/10.1038/nmeth705
83. Rosenberger G, Bludau I, Schmitt U. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat Methods* 2017;**14**:921–7. https://doi.org/10.1038/nmeth.4398
84. Pappireddi N, Martin L, Wühr M. A review on quantitative multiplexed proteomics. *Chembiochem* 2019;**20**:1210–24. https://doi.org/10.1002/cbic.201800650
85. Ha A, Khoo A, Ignatchenko V. *et al.* Comprehensive prostate fluid-based spectral libraries for enhanced protein detection in urine. *J Proteome Res* 2024;**23**:1768–78. https://doi.org/10.1021/acs.jproteome.4c00009
86. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008;**26**:1367–72. https://doi.org/10.1038/nbt.1511
87. Cox J, Neuhauser N, Michalski A. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 2011;**10**:1794–805. https://doi.org/10.1021/pr101065j
88. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 2016;**11**:2301–19. https://doi.org/10.1038/nprot.2016.136
89. Parker R, Tailor A, Peng X. *et al.* The choice of search engine affects sequencing depth and HLA class I allele-specific peptide repertoires. *Mol Cell Proteomics* 2021;**20**:100124. https://doi.org/10.1016/j.mcpro.2021.100124
90. Peng J, Chan C, Meng F. *et al.* Comparison of database searching programs for the analysis of single-cell proteomics data. *J Proteome Res* 2023;**22**:1298–308. https://doi.org/10.1021/acs.jproteome.2c00821
91. Perkins DN, Pappin DJ, Creasy DM. *et al.* Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;**20**:3551–67. https://doi.org/10.1002/(sici)1522-2683(19991201)20:18<3551::Aid-elps3551>3.0.Co;2-2
92. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 2014;**5**:5277. https://doi.org/10.1038/ncomms6277
93. Yu F, Li N, Yu W. PIPI: PTM-invariant peptide identification using coding method. *J Proteome Res* 2016;**15**:4423–35. https://doi.org/10.1021/acs.jproteome.6b00485
94. Uszkoreit J, Barkovits K, Pacharra S. *et al.* Dataset containing physiological amounts of spike-in proteins into murine C2C12 background as a ground truth quantitative LC-MS/MS reference. *Data Brief* 2022;**43**:108435. https://doi.org/10.1016/j.dib.2022.108435
95. Qiang J, Xu Z, Li Y. *et al.* Carboxypeptidase Y assisted Disulfide-bond identification with linearized database search. *Anal Chem* 2021;**93**:14940–5. https://doi.org/10.1021/acs.analchem.1c03932
96. Schoor C, Brocke-Ahmadinejad N, Gieselmann V. *et al.* Investigation of oligodendrocyte precursor cell differentiation by quantitative proteomics. *Proteomics* 2019;**19**:e1900057. https://doi.org/10.1002/pmic.201900057
97. Kong AT, Leprevost FV, Avtonomov DM. *et al.* MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* 2017;**14**:513–20. https://doi.org/10.1038/nmeth.4256
98. Sinitcyn P, Hamzeiy H, Salinas Soto F. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat Biotechnol* 2021;**39**:1563–73. https://doi.org/10.1038/s41587-021-00968-7
99. Yu F, Teo GC, Kong AT. *et al.* Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat Commun* 2023;**14**:4154. https://doi.org/10.1038/s41467-023-39869-5
100. Wang K-F, Wu Y-Z, Chi H. A universal database reduction method based on the sequence tag strategy to facilitate large-scale database search in proteomics. *Int J Mass Spectrom* 2023;**483**:116966. https://doi.org/10.1016/j.ijms.2022.116966
101. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;**5**:976–89. https://doi.org/10.1016/1044-0305(94)80016-2
102. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;**20**:1466–7. https://doi.org/10.1093/bioinformatics/bth092
103. Li D, Fu Y, Sun R. *et al.* pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* 2005;**21**:3049–50. https://doi.org/10.1093/bioinformatics/bti439
104. Zhang J, Xin L, Shan B. *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* 2012;**11**:M111.010587. https://doi.org/10.1074/mcp.M111.010587
105. Na S, Bandeira N, Paek E. Fast multi-blind modification search through tandem mass spectrometry. *Mol Cell Proteomics* 2012;**11**:M111.010199. https://doi.org/10.1074/mcp.M111.010199
106. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics* 2013;**13**:22–4. https://doi.org/10.1002/pmic.201200439
107. Li J, Duncan DT, Zhang B. CanProVar: a human cancer proteome variation database. *Hum Mutat* 2010;**31**:219–28. https://doi.org/10.1002/humu.21176
108. Zhang M, Wang B, Xu J. *et al.* CanProVar 2.0: an updated database of human cancer proteome variation. *J Proteome Res* 2017;**16**:421–32. https://doi.org/10.1021/acs.jproteome.6b00505
109. Pan Q, Shai O, Lee LJ. *et al.* Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;**40**:1413–5. https://doi.org/10.1038/ng.259
110. Yang X, Coulombe-Huntington J, Kang S. *et al.* Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 2016;**164**:805–17. https://doi.org/10.1016/j.cell.2016.01.029
111. Fu XD, Ares M Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* 2014;**15**:689–701. https://doi.org/10.1038/nrg3778
112. Climente-González H, Porta-Pardo E, Godzik A. *et al.* The functional impact of alternative splicing in cancer. *Cell Rep* 2017;**20**:2215–26. https://doi.org/10.1016/j.celrep.2017.08.012
113. Zhang Y, Qian J, Gu C. *et al.* Alternative splicing and cancer: a systematic review. *Signal Transduct Target Ther* 2021;**6**:78. https://doi.org/10.1038/s41392-021-00486-7
114. Zhang Y, Yao X, Zhou H. *et al.* OncoSplicing: an updated database for clinically relevant alternative splicing in 33 human cancers. *Nucleic Acids Res* 2022;**50**:D1340–d1347. https://doi.org/10.1093/nar/gkab851

115. Cifani P, Dhabaria A, Chen Z. *et al.* ProteomeGenerator: a framework for comprehensive proteomics based on de novo transcriptome assembly and high-accuracy peptide mass spectral matching. *J Proteome Res* 2018;**17**:3681–92. https://doi.org/10.1021/acs.jproteome.8b00295
116. Kwok N, Aretz Z, Takao S. *et al.* Integrative Proteogenomics using ProteomeGenerator2. *J Proteome Res* 2023;**22**. https://doi.org/10.1021/acs.jproteome.3c00005
117. Ma B, Zhang K, Hendrie C. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003;**17**:2337–42. https://doi.org/10.1002/rcm.1196
118. Fischer B, Roth V, Roos F. *et al.* NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal Chem* 2005;**77**:7265–73. https://doi.org/10.1021/ac0508853
119. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005;**77**:964–73. https://doi.org/10.1021/ac048788h
120. Chi H, Sun RX, Yang B. *et al.* pNovo: de novo peptide sequencing and identification using HCD spectra. *J Proteome Res* 2010;**9**:2713–24. https://doi.org/10.1021/pr100182k
121. Ma B. Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom* 2015;**26**:1885–94. https://doi.org/10.1007/s13361-015-1204-0
122. Tran NH, Zhang X, Xin L. *et al.* De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* 2017;**114**:8247–52. https://doi.org/10.1073/pnas.1705691114
123. Qiao R, Tran NH, Xin L. *et al.* Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nat Mach Intell* 2021;**3**:420–5. https://doi.org/10.1038/s42256-021-00304-3
124. Yilmaz M, Fondrie W, Bittremieux W. *et al.* Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nature Communications*. 2024/07/30 2024;**15**:6427. https://doi.org/10.1038/s41467-024-49731-x
125. Mao Z, Zhang R, Xin L. *et al.* Mitigating the missing-fragmentation problem in de novo peptide sequencing with a two-stage graph-based deep learning model. *Nat Mach Intell* 2023;**5**:1250–60. https://doi.org/10.1038/s42256-023-00738-x
126. Klaproth-Andrade D, Hingerl J, Bruns Y. *et al.* Deep learning-driven fragment ion series classification enables highly precise and sensitive de novo peptide sequencing. *Nat Commun* 2024;**15**:151. https://doi.org/10.1038/s41467-023-44323-7
127. Ng CCA, Zhou Y, Yao ZP. Algorithms for de-novo sequencing of peptides by tandem mass spectrometry: a review. *Anal Chim Acta* 2023;**1268**:341330. https://doi.org/10.1016/j.aca.2023.341330
128. Allmer J. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev Proteomics* 2011;**8**:645–57. https://doi.org/10.1586/epr.11.54
129. Vitorino R, Guedes S, Trindade F. *et al.* De novo sequencing of proteins by mass spectrometry. *Expert Rev Proteomics* 2020;**17**:595–607. https://doi.org/10.1080/14789450.2020.1831387
130. Muth T, Renard BY. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform* 2018;**19**:954–70. https://doi.org/10.1093/bib/bbx033
131. Chauhan NK, Singh K. A review on conventional machine learning vs deep learning In *Proceedings of the International Conference on Computing, Power and Communication Technologies (GUCON)* 2018;347–52. https://doi.org/10.1109/GUCON.2018.8675097
132. Lecun Y, Bottou L, Bengio Y. *et al.* Gradient-based learning applied to document recognition. *Proc IEEE* 1998;**86**:2278–324. https://doi.org/10.1109/5.726791
133. Kaur M, Mohta A. A review of deep learning with recurrent neural network In *Proceedings of the International Conference on Smart Systems and Inventive Technology (ICSSIT)* 2019;460–5. https://doi.org/10.1109/ICSSIT46314.2019.8987837
134. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735
135. Charles RQ, Su H, Kaichun M. *et al.* PointNet: deep learning on point sets for 3D classification and segmentation. 2017;77–85. https://doi.org/10.1126/science.adw6805
136. Ashish Vaswani NS, Parmar N, Uszkoreit J. *et al.* Illia Polosukhin. *Attention Is All You Need archivePrefix* 2017;**abs:1706.03762**. https://doi.org/10.48550/arXiv.1706.03762
137. Rives A, Meier J, Sercu T. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**:e2016239118. https://doi.org/10.1073/pnas.2016239118
138. Avsec Ž, Agarwal V, Visentin D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021;**18**:1196–203. https://doi.org/10.1038/s41592-021-01252-x
139. De Waele G, Clauwaert J, Menschaert G. *et al.* CpG transformer for imputation of single-cell methylomes. *Bioinformatics* 2022;**38**:597–603. https://doi.org/10.1093/bioinformatics/btab746
140. Barretina J, Caponigro G, Stransky N. *et al.* The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7. https://doi.org/10.1038/nature11003
141. Tate JG, Bamford S, Jubb HC. *et al.* COSMIC: the catalogue of somatic mutations In cancer. *Nucleic Acids Res* 2019;**47**:D941–d947. https://doi.org/10.1093/nar/gky1015
142. Jiang Y, Sun A, Zhao Y. *et al.* Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* 2019;**567**:257–61. https://doi.org/10.1038/s41586-019-0987-8
143. Gao Q, Zhu H, Dong L. *et al.* Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. *Cell* 2019;**179**:561–577.e22. https://doi.org/10.1016/j.cell.2019.08.052
144. Li X, Xu W, Kang W. *et al.* Genomic analysis of liver cancer unveils novel driver genes and distinct prognostic features. *Theranostics* 2018;**8**:1740–51. https://doi.org/10.7150/thno.22010
145. Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* 2010;**604**:55–71. https://doi.org/10.1007/978-1-60761-444-9_5
146. Cerami E, Gao J, Dogrusoz U. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4. https://doi.org/10.1158/2159-8290.Cd-12-0095
147. Rhee M, Burns MA. Nanopore sequencing technology: nanopore preparations. *Trends Biotechnol* 2007;**25**:174–81. https://doi.org/10.1016/j.tibtech.2007.02.008
148. Rhee M, Burns MA. Nanopore sequencing technology: research trends and applications. *Trends Biotechnol* 2006;**24**:580–6. https://doi.org/10.1016/j.tibtech.2006.10.005
149. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol* 2016;**34**:518–24. https://doi.org/10.1038/nbt.3423
150. Dorey A, Howorka S. Nanopore DNA sequencing technologies and their applications towards single-molecule

proteomics. *Nat Chem* 2024;**16**:314–34. https://doi.org/10.1038/s41557-023-01322-x

151. Brinkerhoff H, Kang ASW, Liu J. *et al.* Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science* 2021;**374**:1509–13. https://doi.org/10.1126/science.abl4381
152. Motone K, Kontogiorgos-Heintz D, Wee J. *et al.* Multi-pass, single-molecule nanopore reading of long protein strands. *Nature* 2024;**633**:662–9. https://doi.org/10.1038/s41586-024-07935-7
153. Fan T, Zhang M, Yang J. *et al.* Therapeutic cancer vaccines: advancements, challenges, and prospects. *Signal Transduct Target Ther* 2023;**8**:450. https://doi.org/10.1038/s41392-023-01674-3
154. Lang F, Schrörs B, Löwer M. *et al.* Identification of neoantigens for individualized therapeutic cancer vaccines. *Nat Rev Drug Discov* 2022;**21**:261–82. https://doi.org/10.1038/s41573-021-00387-y
155. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44. https://doi.org/10.1038/nature14539
156. Dong S, Wang P, Abbas K. A survey on deep learning and its applications. *Comput Sci Rev* 2021;**40**:100379. https://doi.org/10.1016/j.cosrev.2021.100379

Briefings in Bioinformatics, 2025, **26(2)**, bbaf087
https://doi.org/10.1093/bib/bbaf087
Review