Computational and Structural Biotechnology Journal 23 (2024) 2057–2066

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

DIPAN: Detecting personalized intronic polyadenylation derived neoantigens from RNA sequencing data

Xiaochuan Liu [a,1], Wen Jin [a,b,1], Dengyi Bao [a], Tongxin He [a], Wenhui Wang [a,b], Zekun Li [a], Xiaoxiao Yang [a,b], Yang Tong [a], Meng Shu [a], Yuting Wang [a], Jiapei Yuan [c], Yang Yang [a,b,*]

[a] The Province and Ministry Co-sponsored Collaborative Innovation Center for Medical Epigenetics, Tianjin Key Laboratory of Inflammatory Biology, Tianjin Geriatrics Institute, Tianjin Medical University General Hospital, The Second Hospital of Tianjin Medical University, Department of Bioinformatics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China
[b] Department of Pharmacology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China
[c] State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Haihe Laboratory of Cell Ecosystem, Institute of Hematology and Blood Diseases Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Tianjin, China

ARTICLE INFO

Keywords:
Intronic polyadenylation
Cancer
Neoantigen
Mass spectrometry

ABSTRACT

Intronic polyadenylation (IPA) refers to a particular type of alternative polyadenylation where a gene makes use of a polyadenylation site located within its introns. Aberrant IPA events have been observed in various types of cancer. IPA can produce noncoding transcripts or truncated protein-coding transcripts with altered coding sequences in the resulting protein product. Therefore, IPA events hold the potential to act as a reservoir of tumor neoantigens. Here, we developed a computational method termed DIPAN, which incorporates IPA detection, protein fragmentation, and MHC binding prediction to predict IPA-derived neoantigens. Utilizing RNA-seq from breast cancer cell lines and ovarian cancer clinical samples, we demonstrated the significant contribution of IPA events to the neoantigen repertoire. Through mass spectrometry immunopeptidome analysis, we further illustrated the processing and presentation of IPA-derived neoantigens on the surface of cancer cells. While most IPA-derived neoantigens are sample-specific, shared neoantigens were identified in both cancer cell lines and clinical samples. Furthermore, we demonstrated an association between IPA-derived neoantigen burden and overall survival in cancer patients.

1. Introduction

Personalized tumor vaccines, based on tumor-specific neoantigens, have emerged as a promising and beneficial treatment strategy for individual patients. These vaccines can elicit anti-tumor T cell responses that specifically target the identified neoantigens [1]. Clinical studies have already indicated that personalized peptide vaccines can significantly improve progression-free survival and overall survival rates in patients with castration-resistant prostate cancer [2]. The initial phase in the development of personalized tumor vaccines entails the identification of tumor-specific neoantigens. These neoantigens can originate from diverse sources, encompassing single nucleotide variants (SNV), insertion, deletion, gene fusion, and post-translational modifications [3–8]. Exploring novel sources of neoantigens would not only broaden our understanding of the neoantigen landscape but also unlock new possibilities for tumor vaccine development.

Two key processes involved in the maturation of mRNA, known as alternative splicing and alternative polyadenylation, are now recognized to occur in most human genes. These processes can change the identity of the encoded protein, thereby serving as a reservoir of neoantigens. Alternative splicing (AS), which happens during RNA processing, holds the potential to alter the encoded protein through cassette exon, 5′ or 3′ splice site selection, mutually exclusive exons, and intron retention [9]. Alternative polyadenylation (APA) involves the cleavage and polyadenylation at different locations on a transcript, affecting the length of the final mRNA [10]. Specially, intronic polyadenylation

* Corresponding author at: The Province and Ministry Co-sponsored Collaborative Innovation Center for Medical Epigenetics, Tianjin Key Laboratory of Inflammatory Biology, Tianjin Geriatrics Institute, Tianjin Medical University General Hospital, The Second Hospital of Tianjin Medical University, Department of Bioinformatics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China.
E-mail address: yy@tmu.edu.cn (Y. Yang).
[1] Xiaochuan Liu and Wen Jin contributed equally to this work.

https://doi.org/10.1016/j.csbj.2024.05.008
Received 6 December 2023; Received in revised form 5 May 2024; Accepted 6 May 2024
Available online 9 May 2024

(IPA), a specific type of APA occurring within intronic regions, has the potential to result in the truncation of the protein product [10]. Recent evidence highlights the crucial role of IPA in various biological and pathological processes, particularly in malignant transformation and tumor progression [11–13]. IPA events are globally upregulated across multiple cancer types [14], although their precise impact on carcinogenesis remains incompletely understood. In B cell leukemia, aberrant IPA events have been shown to produce truncated proteins, deactivating tumor suppressor genes, such as *DICER*, *FOXN3*, and *MGA* [12]. In lung cancer, the IPA isoform of *TSC1* is upregulated, which results in a truncated protein that fails to form a functional TSC1/2 complex, leading to the aberrant activation of mTOR kinase [14]. In breast cancer, the IPA isoform of *MAGI3* generates a truncated protein that inhibits the interaction between full-length MAGI3 and the YAP oncoprotein. This leads to the release of YAP suppression and promotes the malignant transformation of human mammary epithelial cells [11].

Several computational methods have been developed to detect neoantigens originating from aberrant splicing events in RNA-seq data, such as ASNEO, NeoSplice and IRIS [15–17]. However, there is a lack of tools to detect neoantigens derived from IPA events. In light of the pervasive upregulation of aberrant IPA events observed in tumors and the concurrent alterations in coding regions, we propose a hypothesis that IPA could serve as a potential reservoir of tumor neoantigens. The existing computational tools for characterizing IPA from conventional RNA-seq data are limited, which impedes a comprehensive understanding of IPA [14,18]. In this study, we introduce a computational pipeline, designated as DIPAN (**D**etecting **IPA**-derived **N**eoantigens), devised for the identification of tumor-specific neoantigens originating from IPA events, leveraging RNA-seq data. DIPAN uses a tool called IPAFinder to identify IPA events from RNA-seq data [14]. In addition, mass spectrometry data was employed to verify the IPA-derived neoantigens bound to MHC class I molecules. Our framework unveils the prevalence of IPA-derived neoantigens in both cancer cell lines and tumor tissues. These peptides can be presented by MHC class I molecules and are highly immunogenic, providing valuable resources for advancing personalized neoantigen-based immunotherapy strategies.

2. Materials and methods

2.1. DIPAN

The raw sequencing data from tumor and normal samples were collected and aligned to the reference human genome hg38 using HISAT2 with default parameters [19]. The resulting BAM files were sorted and indexed using Samtools [20], serving as input for IPA events prediction. The Reference Sequence (RefSeq) annotation of hg38 was obtained from the UCSC website to generate an annotation file containing information on introns and exons. IPAFinder, a suite of Python scripts, was then utilized to identify IPA events in both tumor and normal samples [14]. Ideal neoantigens should exhibit specific expression in tumor samples while remaining absent in normal samples. To ensure tumor specificity, IPA events from tumor samples that overlapped with any normal sample were excluded. For the tumor-specific IPA events, novel IPA isoforms were assembled based on reference annotation. By analyzing the upstream 5′-untranslated regions and coding sequences, novel IPA isoforms were constructed, with a focus on identifying the first in-frame stop codon. Detection of the stop codon within the IPA isoforms led to the categorization of the transcripts as a protein-coding transcript. The coding sequences of these isoforms were subsequently translated into amino acid sequences using the gffread tool [21]. Peptides of 8–11 amino acids in length, with at least one intronic amino acid, were considered candidate targets, as it is generally recognized that MHC class I molecules bind peptides within this length range [22]. HLA alleles in tumor samples were inferred using OptiType algorithm [23]. Binding affinities of the peptides to MHC class I alleles were predicted using NetMHCpan v4.0, which employs artificial neural networks and achieves high specificity (98.5%) with a percentile rank threshold of 2% [24]. Potential IPA-derived neoantigens were filtered based on a rank threshold of $< 2\%$. Additionally, IPA-derived neoantigens with amino acid sequences present in the normal human proteome, obtained from the UniProt human reference proteome [25] and RefSeq annotation, were excluded due to the likelihood of host immune tolerance. Furthermore, we have also gathered 730 normal samples from The Cancer Genome Atlas (TCGA) and compiled a catalog of IPA-derived peptides identified in normal samples. By using this catalog alongside the existing normal human proteome, researchers can use it as a reference control. Finally, a set of tumor-specific IPA-derived neoantigens with high affinity for MHC class I molecules could be identified.

2.2. Evaluation of IPA-derived neoantigens using mass spectrometry data

The predicted IPA-derived neoantigens required confirmation through mass spectrometry in complex with MHC class I. Paired RNA-seq and mass spectrometry data were collected from two sets of samples. The first set included ten breast cancer cell lines (RNA-seq data PRJNA210428; mass spectrometry data PXD006406) [26], while the second set was from patients with ovarian cancer (RNA-seq data PRJNA398141; mass spectrometry data PXD007635) [27]. The breast cancer cell line dataset comprised ten cell lines (BT549, CAMA1, HCC1395, HCC1419, HCC1428, HCC1569, HCC1806, HCC70, LY2, and MCF7), consisting of 5 luminal, 3 basal, and 2 claudin-low cell lines. The matched normal RNA-seq data was absent for the breast cancer cell line dataset, six samples from the human mammary epithelial cell line (MCF10A cell line, RNA-seq data PRJNA827109) were utilized [28]. DIPAN was applied to each dataset, followed by the processing of the corresponding mass spectrometry data for peptide identification. Raw mass spectrometry data were converted to mzML format using msConvert, facilitating data archiving and downstream analysis [29]. Peptide identification was performed using the Comet search engine, a crucial tool in proteomics research [30]. Peptide segments with a mass tolerance of 20 ppm or less for precursor ions and a mass tolerance of 0.5 Da or less for fragment ions would be selected. Trypsin was selected to fully-digest peptide fragments, resulting in the cleavage of the peptide sequences into N-terminal and C-terminal fragment ions, without specifying any cleavage specificity. The identified neoantigens were visualized using xiSPEC (https://spectrumviewer.org/), a web-based spectrum viewer designed for mass spectrometry data [31].

2.3. Evaluation of MHC II neoantigens derived from IPA in ovarian cancer data

We examined the ovarian cancer dataset containing mass spectrometry data of MHC II restricted peptides to identify potential MHC II neoantigens derived from IPA in the dataset [27]. We adapted DIPAN to identify MHC II neoantigens derived from IPA, culminating in the identification of tumor-specific IPA events and IPA isoforms. MHC II typically bind peptides with 13–18 amino acid residues [32]. Therefore, candidate targets were selected from peptides ranging from 13 to 18 amino acids, with at least one intronic amino acid. The NetMHCIIpan v4.0 was applied to predict peptide binding to MHC class II molecules through artificial neural networks [33]. Peptides ranking below 2% were considered as potential neoantigens. Subsequently, neoantigens matching amino acid sequences found in the normal human proteome were excluded. Finally, mass spectrometry data from MHC II restricted peptides was used to evaluate the predicted MHC II neoantigens derived from IPA.

2.4. Evaluation of immunogenicity

The DeepImmuno-CNN model demonstrates strong predictive performance for the immunogenicity of MHC–peptide pairs [34]. However, it is important to note that the Deepimmuno-CNN only supports peptides

with lengths of 9 and 10 amino acids. As a result, only 9-mer and 10-mer peptides were extracted for the prediction of immunogenicity.

2.5. Collection of $CD8^+$ T cell abundance and SNV neoantigen

We employed two reliable methods, namely CIBERSORT and quanTIseq, to calculate the abundance of $CD8^+$ T cells. CIBERSORT infers cell proportions relative to the total immune cell population, while the quanTIseq estimates absolute proportions relative to the whole cell mixture [35,36]. We leveraged previously computed immune cell abundance of samples in the TCGA cohort. Specifically, we gathered CIBERSORT results from Shmulevich et al. study [37], and quanTIseq results from Liu et al. study [38]. The Cancer Immunome Database (TCIA, https://tcia.at/) provides comprehensive immunogenomic analyses of sequencing data for 20 types of solid cancers from TCGA. The SNV neoantigen data was downloaded from the TCIA database [39].

2.6. Survival analysis

To comprehensively understand the correlation between characteristics of interest and survival, patients were stratified into low and high groups based on the median values of variable metrics. These metrics included IPA-derived neoantigen load, SNV neoantigen load, $CD8^+$ T cell abundance, SNV neoantigen load * $CD8^+$ T cell abundance, and IPA-derived neoantigen load * $CD8^+$ T cell abundance. Subsequently, the log-rank test and Kaplan-Meier survival analysis were performed for the low and high groups. Additionally, univariate and multivariate Cox regression analysis was employed to discern independent prognostic factors. A significance level of $P < 0.05$ was considered statistically significant.

3. Results

3.1. DIPAN design

We have developed a computational approach termed DIPAN for the identification of tumor-specific neoantigens derived from intronic polyadenylation (IPA) events utilizing conventional RNA-seq data (Fig. 1). DIPAN initially detects novel IPA events and their associated novel terminal exons, and subsequently constructs the IPA transcripts from the RNA-seq data [14]. The approach considers two types of novel terminal exons derived from IPA events: composite terminal exons, which encompass the entire sequence from the upstream donor splice site to the IPA site, and skipped terminal exons, which involve the introduction of a novel exon terminating at the IPA site, necessitating the recognition of both an upstream donor splice site and a new acceptor splice site [40]. Subsequently, DIPAN identifies tumor-specific IPA events by excluding any IPA events that overlap with those identified in normal tissues. DIPAN then translates the tumor-specific IPA events into proteins in silico. By utilizing a sliding window (n = 8, 9, 10, 11) on the proteins, DIPAN generates all possible short peptides. Only peptides derived from the intronic sequences introduced by IPA events are retained to identify tumor-specific neo-peptides. Furthermore, by integrating the sample-specific MHC class I alleles predicted from the RNA-seq data, DIPAN can predict binding affinities between these neo-peptides and the corresponding MHC class I molecules using NetMHCpan [24]. Finally, DIPAN filters IPA-derived neoantigens by checking if they are found in the normal human proteome and can bind with MHC class I molecules.

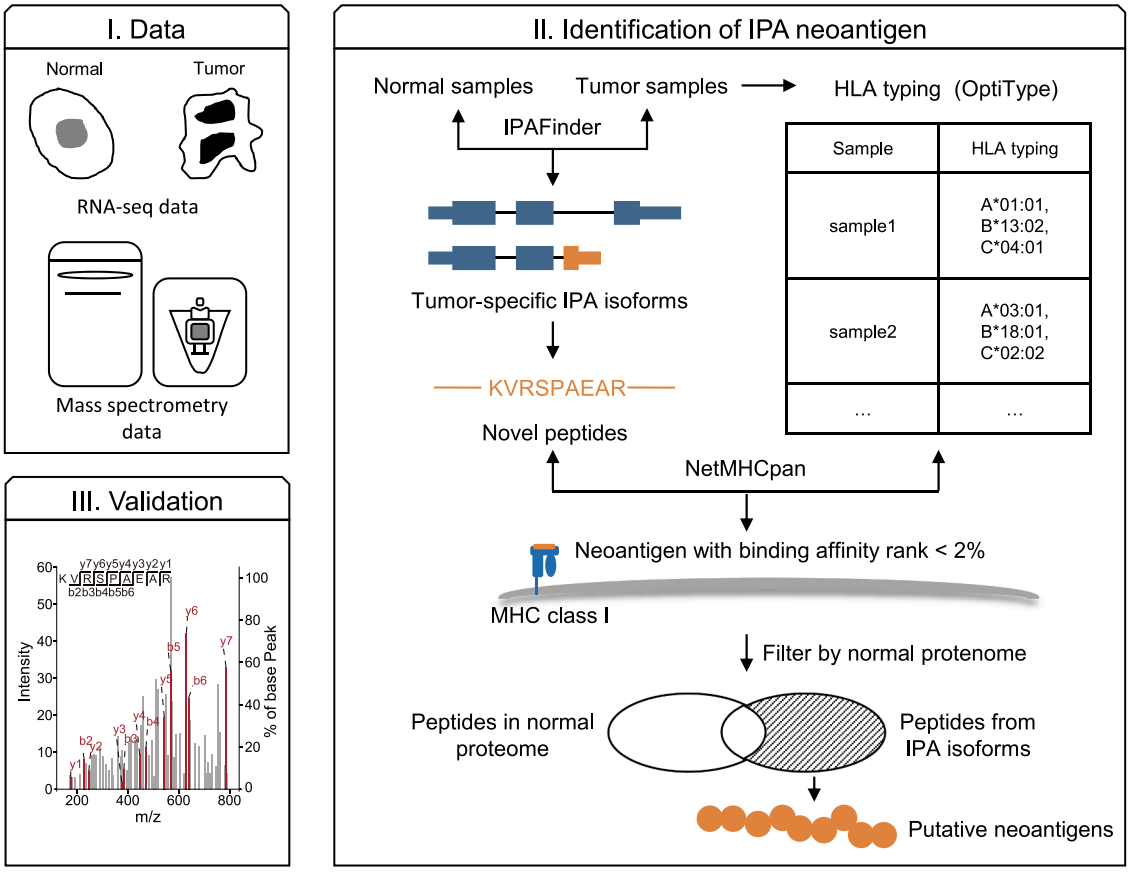


Fig. 1. DIPAN is designed to identify IPA-derived neoantigens from RNA-seq data. Part I represents the data employed in this analysis, including RNA-seq and mass spectrometry data. Part II outlines the steps involved in the DIPAN pipeline. Tumor-specific IPA events are filtered, and their coding sequences are translated into amino acid sequences. Peptides with at least one intronic amino acid are evaluated for binding affinities using NetMHCpan. DIPAN incorporates a filtering and thresholding process to eliminate artifacts. Part III demonstrates the validation of IPA-derived neoantigens using mass spectrometry data.

DIPAN sets a threshold of rank < 2% to identify strong MHC-I binders and removes neoantigens with amino acid sequences matching those in normal human proteome. This filtering process yields a reliable list of potential IPA-derived neoantigens.

3.2. Identification of IPA-derived neoantigens in breast cancer cell lines

To demonstrate DIPAN's efficacy, DIPAN was employed on an RNA-seq dataset encompassing 10 distinct breast cancer cell lines [26]. The comprehensive analysis revealed 1336 tumor-specific IPA events in the breast cancer cell lines compared to the normal mammary epithelial cell line (Fig. 2**A**). IPA transcripts with coding potential can generate novel proteins, serving as a source of neoantigens following degradation. By assessing the binding affinities between IPA derived neo-peptides and sample-specific MHC class I alleles, putative IPA-derived neoantigens were identified. The IPA-derived neoantigen load exhibited a range of 3004 to 8529 across cell lines, with a median value of 4924 (Fig. 2**B**).

The high-affinity interaction between MHC class I molecules and these peptides is crucial for the cytotoxic T cell response against the tumor, which is also pertinent to the design of peptide vaccines. To validate the processing and presentation of IPA-derived neoantigens on MHC class I, mass spectrometry immunopeptidome analysis was employed to directly identify peptides complexed with MHC class I. Encouragingly, mass spectrometry immunopeptidome analysis confirmed a total of 1883 IPA-derived neoantigens (Supplementary Table S1), ranging from 154 to 309 in each sample (Supplementary Table S2). The majority of validated IPA-derived neoantigens (87.1%, n = 1641) were uniquely detected in individual samples, highlighting the scarcity of shared neoantigens (Fig. 2C). A tumor-specific composite IPA isoform of the *PFKL* gene was detected in the HCC1428, HCC1419, and MCF7 cell lines (Fig. 2D). Mass spectrometry immunopeptidome analysis further confirmed the presence of the resulting IPA-derived neoantigen (ADACCTWTRR) in complex with MHC class I (Fig. 2E-G). Additionally, a tumor-specific skipped IPA isoform of the *NCOA3* gene exclusive in the LY2 cell lines was discovered, and its resulting IPA-derived neoantigen (KQFGLQAR) was confidently identified in complex with MHC class I via mass spectrometry (Supplementary Fig. S1A, B). Moreover, we discovered the presence of IPA-derived neoantigens LAPSVTYPR from the *MDFI* gene and MTSGAHMR from the *CDC42BPB* gene in the HCC1806 cell line (Supplementary Fig. S1C-F). Subsequently, the IPA-derived neoantigens were categorized into high-affinity and low-affinity groups based on their estimated binding affinity to MHC class I molecules in each sample. The data indicated a significantly higher proportion of neoantigens validated by mass spectrometry in the high-affinity group compared to the low-affinity group (Supplementary Fig. S2A). Furthermore, the relative abundances of IPA isoforms were found to be significantly higher in the high-affinity group relative to the low-affinity group (Supplementary Fig. S2B). These results collectively suggest that IPA-derived neoantigens are more likely to be supported by immunopeptidome data when originating from IPA isoforms that are abundant and exhibit strong binding to MHC molecules, consistent with the anticipated MHC-peptide binding pattern.

3.3. Identification of IPA-derived neoantigens in ovarian cancer patients

In addition to the cell lines, we also applied it to an RNA-seq dataset of ovarian cancer samples [27]. The analysis identified 2494 tumor-specific IPA events in 14 cancer samples compared to benign fallopian tube tissue, suggesting a potential role of widespread IPA events in tumor progression (Fig. 3A). A wide range of IPA-derived neoantigens, comprising 8–11 amino acids, were evaluated for their binding affinity with MHC class I for each sample, ranging from 3389 to 14,877 (Fig. 3B). Notably, the mass spectrometry immunopeptidome analysis confirmed that a total of 5291 IPA-derived neoantigens were indeed bound to MHC class I molecules (Supplementary Table S3). The validated IPA-derived neoantigen load ranged from 243 to 915, with a median value of 544.5 (Supplementary Table S4). Similarly, out of the validated IPA-derived neoantigens, the majority (76.9%, n = 4071) were exclusively detected in a single sample, while the remaining peptides were identified in multiple samples (Fig. 3C). For instance, we identified a tumor-specific composite IPA isoform of the *SMPDL3B* gene in nine patients (Fig. 3D), and the predicted IPA-derived neoantigen (KVRSPAEAR) was confidently detected in complex with MHC class I by mass spectrometry (Fig. 3E-G; Supplementary Fig. S3). Furthermore, we discovered a tumor-specific skipped IPA isoform of the *PHIP* gene exclusive to one sample (Supplementary Fig. S4A), and its resulting IPA-derived neoantigen (RTFVILGRR) was also confidently identified in complex with MHC class I through mass spectrometry (Supplementary Fig. S4B). Additionally, we showed the presence of IPA-derived neoantigens YIGTGRTTR and QCTRYRRPTR derived from *SPATA6L* and *WASHC5*, respectively (Supplementary Fig. S4C-F). Consistent with observations in the breast cancer cell line, the proportions of neoantigens validated by mass spectrometry were significantly higher in the high-affinity group compared to the low-affinity group in ovarian cancer (Supplementary Fig. S5A). Additionally, the data indicated that the relative abundances of the IPA isoforms were significantly higher in the high-affinity group compared to the low-affinity group (Supplementary Fig. S5B). These findings collectively provide compelling evidence regarding the processing and presentation of IPA-derived neoantigens through the MHC class I pathway in both cell lines and tumor samples, thereby underscoring the significance of IPA as a substantial source of neoantigens in cancer.

MHC class II molecules on the surface of antigen-presenting cells display a range of peptides for recognition by the T cell receptors of $CD4^+$ T helper cells. MHC class II restricted neoantigens also have the potential to be a promising target of tumor immunotherapy [41–43]. We have also tried to identify MHC II neoantigens derived from IPA in the ovarian cancer data and validated by available mass spectrometry data that captured MHC II restricted peptides. The number of predicted MHC II IPA-derived neoantigens ranged from 46 to 615 in the ovarian cancer samples (Supplementary Fig. S6). The predicted neoantigens validated by mass spectrometry data ranged from 1 to 44. Given the limited number of MHC-II neoantigens and lower performance of existing tools compared to MHC-I-driven models, DIPAN focused exclusively on MHC-I neoantigens derived from IPA events.

3.4. The immunogenicity of IPA-derived neoantigens identified by DIPAN

The IPA-derived neoantigens, identified by DIPAN, have the potential to create peptide-MHC complexes on cell surface, which may or may not be recognized by T cell receptors. To understand the immunogenicity of these predicted neoantigens. It is important to determine if these peptides can be recognized by T cells and trigger an immune response. The DeepImmuno was used to estimated immunogenicity in peptide-MHC pairs [34]. In the breast cancer cell line data, 49.7% to 98.0% of peptides were found to be immunogenic, with a median of 87.8% (Fig. 4A). In ovarian cancer data, the range was from 64.3% to 98.2%, with a median of 85.6% (Fig. 4B). These results indicate that a significant portion of predicted IPA-derived neoantigens can activate $CD8^+$ T cells, suggesting their potential as tumor-specific neoantigens. Further investigations into the immunogenicity of IPA-derived neoantigens and their impact on antitumor immune responses are warranted to fully exploit their potential for clinical use.

3.5. IPA-derived neoantigens correlate with patients' overall survival

Neoantigens possess the potential to shape the tumor microenvironment and enhance antitumor immune responses [16]. Our study has demonstrated that IPA-derived neoantigens serve as a crucial supplement to the neoantigen repertoire, suggesting their potential as biomarkers for prognosis. By employing RNA-seq data from the TCGA ovarian cancer (TCGA-OV) cohort, we probed the relationship between

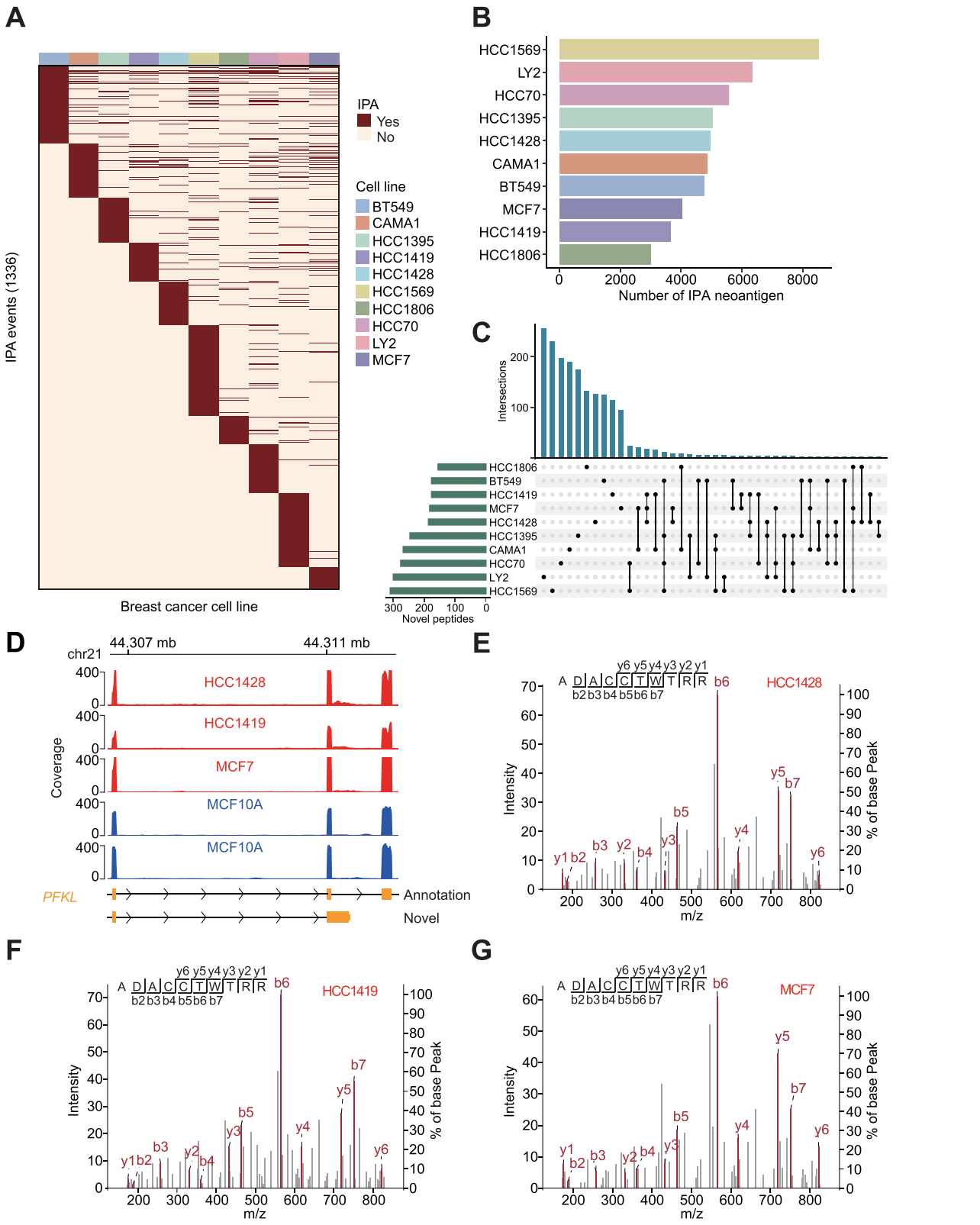


Fig. 2. Identification of IPA-derived neoantigens in breast cancer cell lines. (A) The heatmap visualizes the occurrence of tumor-specific IPA events in 10 breast cancer cell lines. The darker color indicates the occurrence of IPA events, while the lighter color indicates the absence of IPA events. **(B)** The bar plot shows the number of predicted IPA-derived neoantigens in each cell line using DIPAN. (C) The upset plot indicates the number of IPA-derived neoantigens that have been successfully verified through mass spectrometry. The overlapping regions indicate the neoantigens that are shared in different cell lines. **(D)** The plot illustrates the read coverage of RNA-seq data for the *PFKL* gene, highlighting a specific example of tumor-specific IPA events. (E-G) The IPA-derived neoantigen (ADACCTWTRR) identified from the *PFKL* gene is both predicted in silico and found by mass spectrometry in HCC1428 **(E)**, HCC1419 **(F)**, MCF7 **(G)** cell lines.

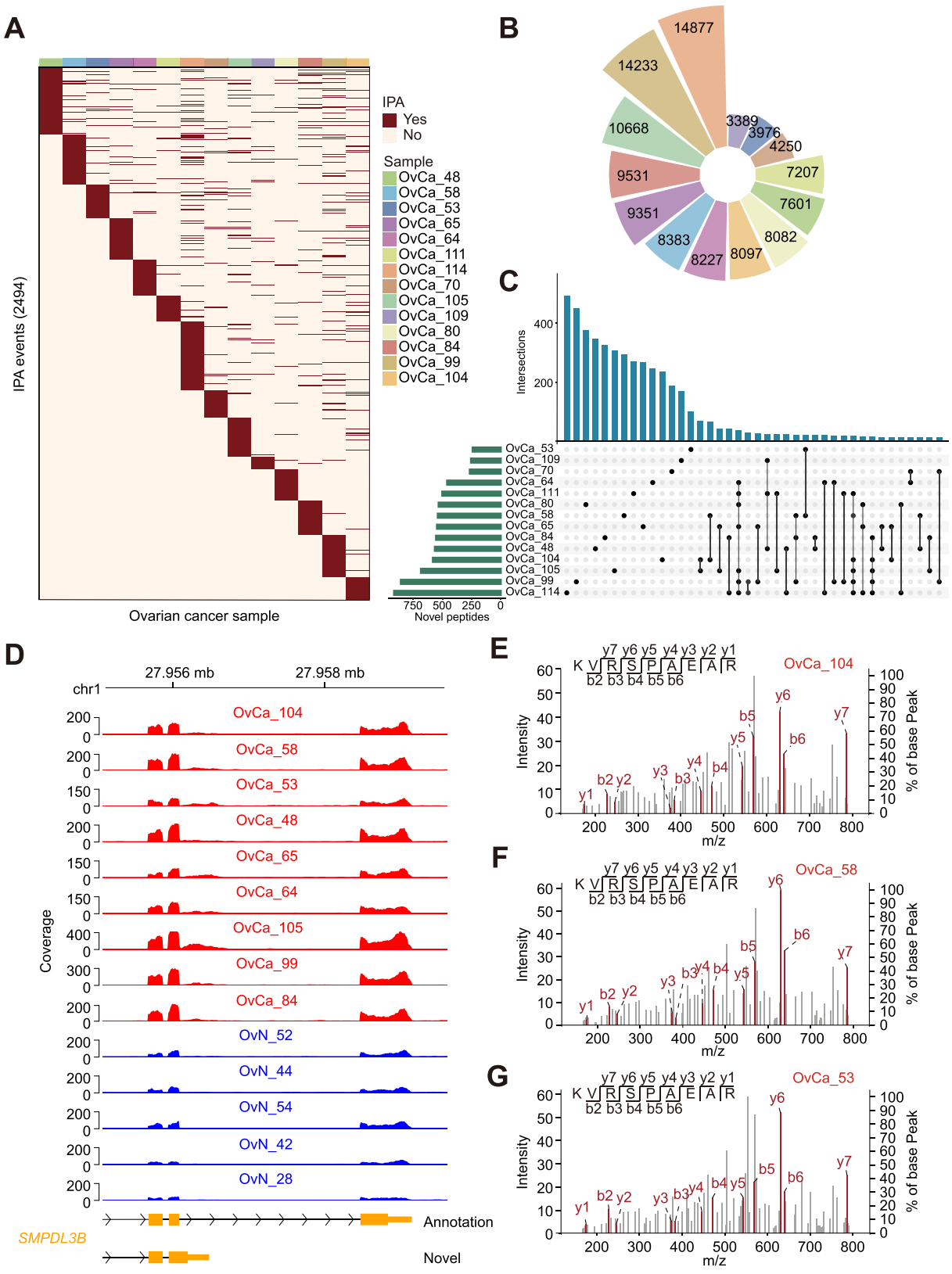


Fig. 3. IPA-dervied neoantigens in ovarian cancer patients. (A) The heatmap visualizes the occurrence of tumor-specific IPA events in 14 ovarian cancer patients. The darker color indicates the occurrence of IPA events, while the lighter color indicates the absence of IPA events. (B) Nightingale rose chart illustrates the count of predicted IPA-derived neoantigens per patient utilizing DIPAN. Samples are indicated by color, as depicted in Fig. 3A. (C) The upset plot visualizes the number of IPA-derived neoantigens that have been successfully verified through mass spectrometry. The overlapping regions indicate the neoantigens that are shared in different samples. (D) The plot illustrates the read coverage of RNA-seq data for the *SMPDL3B* gene, highlighting a specific example of tumor-specific IPA events. (E-G) The IPA-derived neoantigen (KVRSPAEAR) identified from the *SMPDL3B* gene is both predicted in silico and found by mass spectrometry in OvCa_104 **(E)**, OvCa_58 **(F)** and OvCa_53 **(G)** samples.

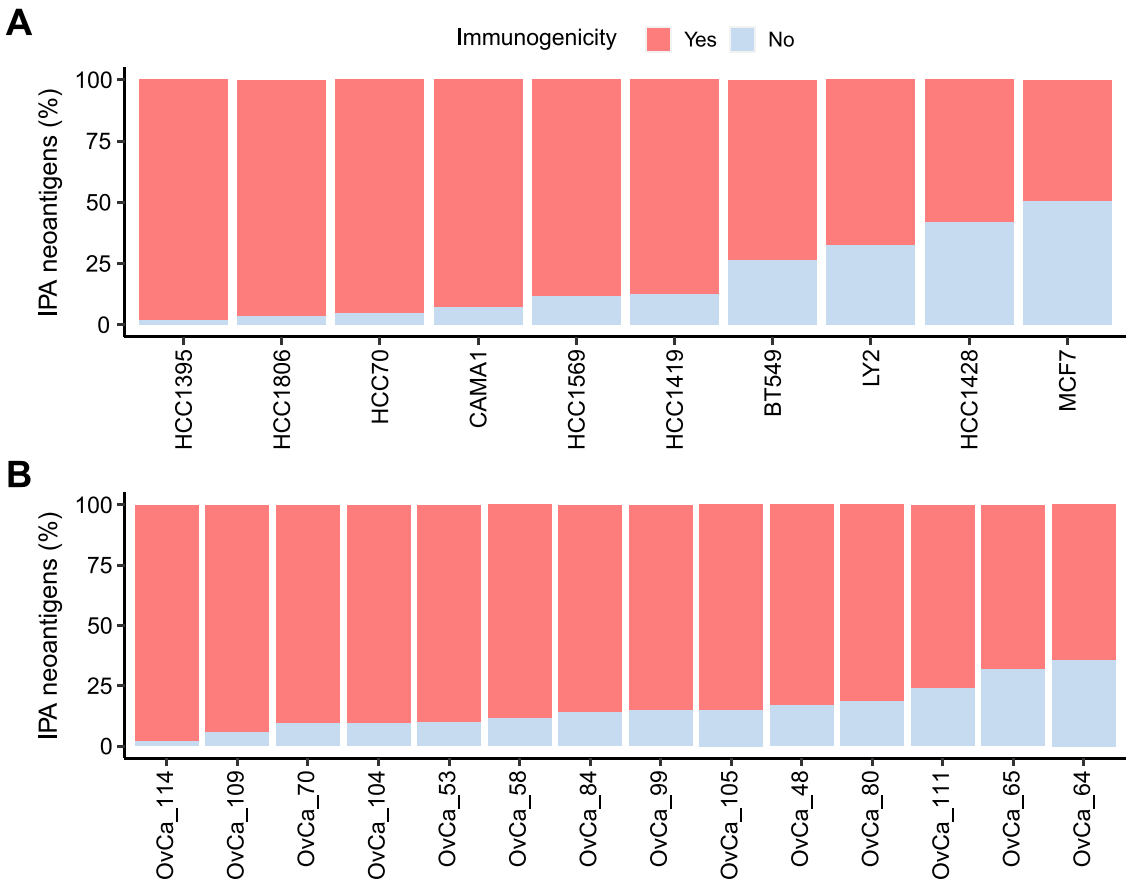


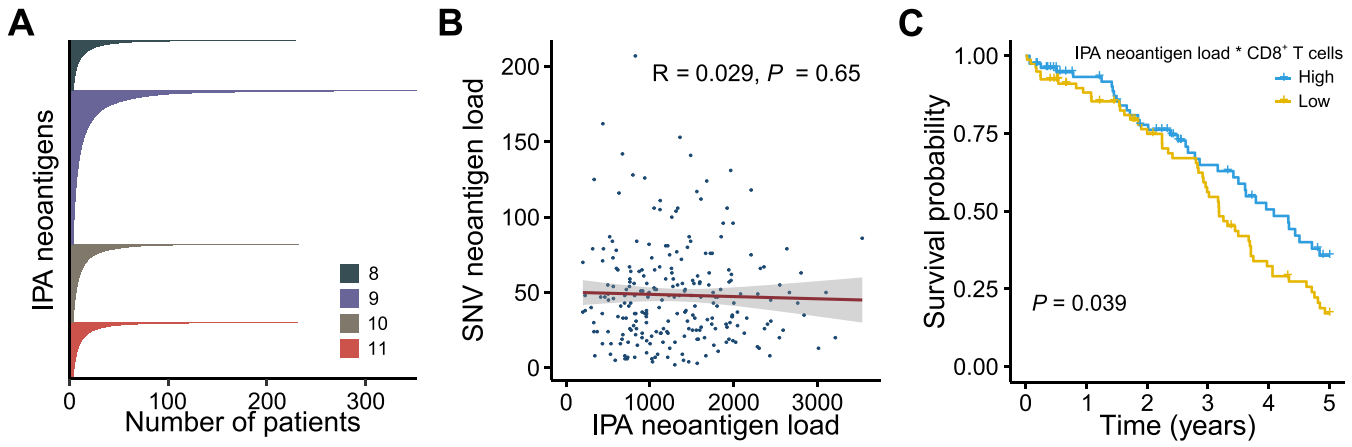
Fig. 4. Immunogenicity of IPA-derived neoantigens (A, B) Proportions of immunogenic IPA-derived neoantigens were evaluated using the DeepImmuno model in both the breast cancer cell lines **(A)** and ovarian cancer samples (**B**).

IPA-derived neoantigens and patients' prognoses. Our analysis unveiled a total of 155,251 tumor-specific IPA-derived neoantigens, with 28,006 peptides detected in at least 5 samples (Fig. 5A). These neoantigens were utilized to compute the IPA-derived neoantigen load (IPA neoantigen load). Interestingly, our findings revealed no correlation between the IPA neoantigen load and the SNV neoantigen load (Fig. 5B), indicating that IPA-derived neoantigens represent a distinct source of neoantigens compared to those originating from SNVs.

The survival analysis showed that the individual neoantigen loads of IPA and SNV were not independently associated with the overall survival of TCGA-OV patients (Supplementary Fig. S7A-B). Given the crucial role of tumor-specific peptide-MHC complexes in activating T cell-mediated immune responses, the combined effect of neoantigen load and $CD8^+$ T cell abundance was investigated. Two computational tools, quanTIseq and CIBERSORT, were employed to quantify $CD8^+$ T cells. While quanTIseq estimates absolute fractions, CIBERSORT infers cell fractions relative to the total immune cell population [35,36]. Despite the recognized importance of $CD8^+$ T cells in the anticancer immunity, their abundance alone did not exhibit a significant correlation with patients' overall survival (Supplementary Fig. S7C, Supplementary Fig. S8A). Intriguingly, only the combined effect of IPA neoantigen load and $CD8^+$ T cell abundance (IPA neoantigen load * $CD8^+$ T cells) demonstrated a statistically significant association with overall survival. This association remained significant for both absolute and relative fractions of $CD8^+$ T cells (Fig. 5C, Supplementary Fig. S8C). Conversely, the combined effect of SNV neoantigen load and $CD8^+$ T cell abundance did not yield significant results (Supplementary Fig. S7D, Supplementary Fig. S8B). Univariate and multivariate Cox regression analyses confirmed that the combined effect of IPA neoantigen load and $CD8^+$ T cell abundance was predictive of OV patients' survival, independent of age and stage (Fig. 5D, Supplementary Fig. S8D). However, in the TCGA breast cancer (TCGA-BRCA) cohort, no significant association was observed between IPA-derived neoantigens and survival (Supplementary Fig. S9-S11). When extending the analysis to the TCGA lung squamous cell carcinoma (TCGA-LUSC) cohort, a significant association with survival was observed for the combination of IPA-derived neoantigens and $CD8^+$ T cell abundance (Supplementary Fig. S12-S14). Furthermore, we have conducted comparative analyses between IPA neoantigens and SNV neoantigens within the TCGA-OV, TCGA-BRCA and TCGA-LUSC cohorts. The findings revealed that substantial disparities in the neoantigen loads between IPA and SNV categories in TCGA-OV and TCGA-LUSC cohorts, with only minor differences observed in TCGA-BRCA (Supplementary Fig. S15). Additionally, a higher proportion of immunogenic neoantigens derived from IPA was found compared to immunogenic neoantigens derived from SNV in TCGA-OV and TCGA-LUSC cohorts, while no significant contrast was observed in TCGA-BRCA (Supplementary Fig. S15). These differences may provide insights into why IPA neoantigen * $CD8^+$ T cells serves as a better predictor of patient survival than SNV neoantigen * $CD8^+$ T cells, as well as the absence of a significant correlation between IPA-derived neoantigens and survival in TCGA-BRCA cohort.

4. Discussion

In this study, we presented DIPAN, a computational pipeline designed for predicting neoantigens derived from IPA events based on conventional RNA-seq. DIPAN demonstrated its ability to identify neoantigens derived from IPA events within RNA-seq data from both cancer cell lines and tissue samples. Validation through mass spectrometry immunopeptidome data confirmed that a subset of these predicted



D

Variable	Univariate analysis Hazard ratio (95% CI)	*P*–value	Multivariate analysis (Age + Stage) Hazard ratio (95% CI)	*P*–value
CD8⁺ T cells				
Low	Reference		Reference	
High	0.82 (0.54, 1.24)	0.345	0.78 (0.51, 1.19)	0.253
SNV neoantigen				
Low	Reference		Reference	
High	0.88 (0.58, 1.33)	0.536	0.86 (0.57, 1.31)	0.496
IPA neoantigen				
Low	Reference		Reference	
High	0.77 (0.51, 1.17)	0.224	0.78 (0.51, 1.18)	0.238
SNV neoantigen load * CD8⁺ T cells				
Low	Reference		Reference	
High	0.77 (0.51, 1.17)	0.224	0.75 (0.50, 1.14)	0.185
IPA neoantigen load * CD8⁺ T cells				
Low	Reference		Reference	
High	0.64 (0.42, 0.98)	0.041*	0.61 (0.40, 0.94)	0.025*

Fig. 5. Associations between IPA-derived neoantigen load and patients' overall survival in ovarian cancer patients. (A) The barplot shows the distribution of shared neoantigens from at least 5 samples. **(B)** Relationship between IPA-derived neoantigen loads (x-axis) and SNV neoantigen loads (y-axis). The Pearson correlation coefficient and *P* value are shown in the plot. **(C)** Kaplan-Meier curves for overall survival of groups stratified by IPA-derived neoantigen load * $CD8^+$ T cell abundance, with the median value as the cutoff. Significance is assessed through a two-sided log-rank test. **(D)** Results of univariate and multivariate Cox regression analysis for overall survival, with patients stratified by median values.

neoantigens were indeed presented and complexed with MHC class I. By employing DeepImmuno to assess the immunogenicity of the IPA-derived neoantigens identified by DIPAN, it was revealed that the majority of these neoantigens are indeed immunogenic. The presence of MHC-neoantigen complexes and their immunogenicity suggested that aberrant IPA events in human cancers not only compromise protein function but also possess the potential to modulate tumor immunity. Throughout tumor progression, splicing perturbations contribute to the gradual accumulation of abnormal IPA events in tumor samples. A fraction of these tumor-specific IPA isoforms could undergo translation and degradation, yielding immunogenic peptides. Subsequently, these peptides are bound by MHC class I and presented on the cell surface as targets for T lymphocytes [44]. This source of neoantigens derived from tumor-specific IPA events, previously overlooked, opens new avenues for the design of neoantigen-based personalized cancer vaccines. Moreover, our analysis demonstrated that the IPA-derived neoantigen load, when combined with $CD8^+$ T cell abundance, served as a predictor for patient survival. Our analysis underscores the significance of IPA as a crucial source of neoantigens, thereby presenting additional opportunities for the development of neoantigen-based treatments.

As a proof-of-concept analysis, we have successfully corroborated the presence of IPA-derived neoantigens as identified by DIPAN within the MHC I immunopeptidome of both human cancer cell lines and tissue specimens. Although only a fraction of the DIPAN identified neoantigens can be supported by immunopeptidome data, it revealed a validation rate that is comparable to those reported in existing literature. For example, Alicia et al. proposed a method for identifying neoantigens derived from intron retention, yet only a minimal number of neoantigens were detected via mass spectrometry data amidst a vast pool of computationally predicted neoantigens across six human cancer cell lines [7]. Similarly, Pan et al. introduced IRIS, a computational tool for the identification of neoantigens derived from aberrant alternative splicing, with only a limited subset being validated within the immunopeptidome of three cancer cell lines [17]. Various factors contribute to the observed low validation rate, including the sensitivity of mass spectrometry-based detection influenced by peptide abundance, MHC binding affinity and sample quality. Noteworthy is the possibility that the current computational methods have inherent limitations, with many neoantigen prediction algorithms focusing primarily on presentation criteria rather than recognition criteria, potentially explaining the

low validation rates encountered.

The immunogenicity of a peptide relies on two fundamental criteria: MHC presentation and T cell receptor recognition. DIPAN solely focused on the affinity of IPA-derived neoantigens with MHC class I, overlooking the recognition of MHC-neoantigen complexes by T cell receptors. By employing DeepImmuno to assess the immunogenicity of the IPA-derived neoantigens identified by DIPAN, it was revealed that the majority of these neoantigens are indeed immunogenic. Our future research will focus on conducting functional immunological assays to further validate these findings. Furthermore, considering the existence of multiple sources of neoantigens, it is worthwhile to investigate whether neoantigens derived from various sources share common pathways for stimulating cytotoxic T cells, particularly in terms of the processing, presentation, and immunogenicity of neoantigens.

Tumor neoantigens show promise for cancer treatment by triggering personalized immune response against cancer cells. IPA-derived neoantigens serve as a valuable supplement to the neoantigen repertoire, enhancing our understanding of the tumor microenvironment. Cancer vaccines targeting neoantigens have the potential to activate the immune system and eliminate cancer cells. Although the development of cancer vaccines encounters various challenges. Initial studies on personalized vaccines for melanoma [45] and glioblastoma [46] patients have shown positive results. Neoantigen-based adoptive cell transfer is also a promising approach in the field of immunotherapy, exhibiting remarkable efficacy in hematologic tumors while encountering challenges in solid tumors [47]. Tumor-specific IPA-derived neoantigens significantly expand the neoantigen repertoire, thereby enhancing the array of potential targets for immunotherapeutic interventions. DIPAN, applied to conventional RNA-seq data, can identify potential tumor-specific IPA-derived neoantigens. These predicted IPA-derived neoantigens are linked to the tumor microenvironment and contribute to expanding the pool of neoantigen peptides, providing valuable insights for advancing personalized neoantigen-based immunotherapy strategies.

5. Conclusions

In this study, the tumor-specific neoantigens derived from IPA events were investigated in both breast cancer cell lines and ovarian cancer patients. The presentation of these neoantigens via MHC class I confirms the significance of IPA as a substantial source of cancer neoantigens. Additionally, the IPA-derived neoantigen load was shown to be a potential predictor for patient survival, highlighting the intricate interplay between the tumor microenvironment and neoantigens. These findings offer promise for the development of personalized tumor vaccines. The study also introduces DIPAN as a robust and reliable pipeline for the precise identification of neoantigens from IPA events. The potential of DIPAN in discerning tumor-specific IPA-derived neoantigens from RNA-seq data provides valuable insights into their immunogenicity and implications for cancer immunotherapy.

Author contributions

YY conceived and designed the study. YY and JY supervised the study and data analysis. XL and WJ developed the method, wrote the original code, and analyzed the data. DB, TH, WW, ZL, XY, YT, and MS helped the data analysis, YW helped the method development, YY, XL and WJ wrote the manuscript with input from all authors. All authors approved the final manuscript submitted.

CRediT authorship contribution statement

Jiapei Yuan: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision. **Yuting Wang:** Formal analysis, Funding acquisition, Investigation, Methodology. **Yang Yang:** Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Dengyi Bao:** Data curation, Formal analysis, Investigation. **Wenhui Wang:** Data curation, Formal analysis, Investigation. **Wen Jin:** Data curation, Formal analysis, Investigation, Methodology, Software, Visualization. **Tongxin He:** Data curation, Formal analysis, Investigation. **Xiaochuan Liu:** Data curation, Formal analysis, Investigation, Methodology, Software, Visualization. **Xiaoxiao Yang:** Data curation, Formal analysis, Investigation. **Zekun Li:** Data curation, Formal analysis, Investigation. **Meng Shu:** Data curation, Formal analysis, Investigation. **Yang Tong:** Data curation, Formal analysis, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

A comprehensive overview of the datasets employed in this study is presented in Supplementary Table S5. RNA-seq data for ten breast cancer cell lines were obtained from the National Center for Biotechnology Information BioProject database under accession number PRJNA210428, while RNA-seq data for a normal breast epithelial cell line originated from PRJNA827109. RNA-seq data for ovarian cancer samples were sourced from the BioProject database under accession number PRJNA398141. Mass spectrometry data for breast cancer cell lines and ovarian cancer samples were obtained from the ProteomeXchange Consortium through the PRIDE partner repository, with dataset identifiers PXD006406 and PXD007635, respectively. DIPAN is an open-source tool available on the GitHub repository at https://github.com/YY-TMU/DIPAN.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 32100534, to YY), Talent Excellence Program from Tianjin Medical University (to YY), National Natural Science Foundation of China (Grant No. 32200514, to JY), National Natural Science Foundation of China (Grant No. 32200547, to YW), and Tianjin Natural Science Foundation (22JCQNJC01690). We thank all members in the group for their assistance and constructive suggestions. We gratefully acknowledge the technical support by the High-performance Computing Platform of Tianjin Medical University.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.05.008.

References

[1] Shemesh CS, Hsu JC, Hosseini I, Shen BQ, Rotte A, Twomey P, et al. Personalized cancer vaccines: clinical landscape, challenges, and opportunities. Mol Ther 2021;29:555–70.
[2] Kimura T, Egawa S, Uemura H. Personalized peptide vaccines and their relation to other therapies in urological cancer. Nat Rev Urol 2017;14:501–10.
[3] Parkhurst MR, Robbins PF, Tran E, Prickett TD, Gartner JJ, Jia L, et al. Unique neoantigens arise from somatic mutations in patients with gastrointestinal cancers. Cancer Discov 2019;9:1022–35.
[4] Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol 2017;18:1009–21.
[5] Yang W, Lee KW, Srivastava RM, Kuo F, Krishna C, Chowell D, et al. Immunogenic neoantigens derived from gene fusions stimulate T cell responses. Nat Med 2019;25:767–75.
[6] Cheng R, Xu Z, Luo M, Wang P, Cao H, Jin X, et al. Identification of alternative splicing-derived cancer neoantigens for mRNA vaccine development. Brief Bioinform 2022;23.

[7] Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, et al. Intron retention is a source of neoepitopes in cancer. Nat Biotechnol 2018;36:1056–8.
[8] Srivastava AK, Guadagnin G, Cappello P, Novelli F. Post-translational modifications in tumor-associated antigens as a platform for novel immuno-oncology therapies. Cancers (Basel) 2022;15.
[9] Marasco LE, Kornblihtt AR. The physiology of alternative splicing. Nat Rev Mol Cell Biol 2023;24:242–54.
[10] Mitschka S, Mayr C. Context-specific regulation and function of mRNA alternative polyadenylation. Nat Rev Mol Cell Biol 2022;23:779–96.
[11] Ni TK, Kuperwasser C. Premature polyadenylation of MAGI3 produces a dominantly-acting oncogene in human breast cancer. ELife 2016;5.
[12] Lee S-H, Singh I, Tisdale S, Abdel-Wahab O, Leslie CS, Mayr C. Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. Nature 2018;561:127–31.
[13] Singh I, Lee S-H, Sperling AS, Samur MK, Tai Y-T, Fulciniti M, et al. Widespread intronic polyadenylation diversifies immune cell transcriptomes. Nat Commun 2018;9:1716.
[14] Zhao Z, Xu Q, Wei R, Wang W, Ding D, Yang Y, et al. Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. Genome Res 2021;31:2095–106.
[15] Chai S, Smith CC, Kochar TK, Hunsucker SA, Beck W, Olsen KS, et al. NeoSplice: a bioinformatics method for prediction of splice variant neoantigens. Bioinform Adv 2022;2:vbac032.
[16] Zhang Z, Zhou C, Tang L, Gong Y, Wei Z, Zhang G, et al. ASNEO: Identification of personalized alternative splicing based neoantigens with RNA-seq. Aging (Albany NY) 2020;12:14633–48.
[17] Pan Y, Phillips JW, Zhang BD, Noguchi M, Kutschera E, McLaughlin J, et al. IRIS: discovery of cancer immunotherapy targets arising from pre-mRNA alternative splicing. Proc Natl Acad Sci USA 2023;120:e2221116120.
[18] Liu X, Chen H, Li Z, Yang X, Jin W, Wang Y, et al. InPACT: a computational method for accurate characterization of intronic polyadenylation from RNA sequencing data. Nat Commun 2024;15:2583.
[19] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;37:907–15.
[20] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
[21] Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. F1000Res 2020;9.
[22] Yewdell JW, Reits E, Neefjes J. Making sense of mass destruction: quantitating MHC class I antigen presentation. Nat Rev Immunol 2003;3:952–61.
[23] Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 2014;30:3310–6.
[24] Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide-MHC Class I interaction predictions integrating eluted ligand and peptide binding affinity data. J Immunol 2017;199:3360–8.
[25] Consortium TU: UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 2022, 51:D523-D531.
[26] Rozanov DV, Rozanov ND, Chiotti KE, Reddy A, Wilmarth PA, David LL, et al. MHC class I loaded ligands from breast cancer cell lines: A potential HLA-I-typed antigen collection. J Proteom 2018;176:13–23.
[27] Schuster H, Peper JK, Bösmüller HC, Röhle K, Backert L, Bilich T, et al. The immunopeptidomic landscape of ovarian carcinomas. Proc Natl Acad Sci USA 2017;114. E9942-e9951.
[28] Hahm ER, Mathan SV, Singh RP, Singh SV. Breast cancer selective disruption of actin cytoskeleton by diallyl trisulfide. J Cancer Prev 2022;27:101–11.
[29] Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 2012;30:918–20.
[30] Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics 2013;13:22–4.
[31] Kolbowski L, Combe C, Rappsilber J. xiSPEC: web-based visualization, analysis and sharing of proteomics data. Nucleic Acids Res 2018;46. W473-w478.
[32] Cai Y, Lv D, Li D, Yin J, Ma Y, Luo Y, et al. IEAtlas: an atlas of HLA-presented immune epitopes derived from non-coding regions. Nucleic Acids Res 2023;51. D409-d417.
[33] Reynisson B, Barra C, Kaabinejadian S, Hildebrand WH, Peters B, Nielsen M. Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. J Proteome Res 2020;19:2304–15.
[34] Li G, Iyer B, Prasath VBS, Ni Y, Salomonis N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. Brief Bioinform 2021;22.
[35] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453–7.
[36] Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med 2019;11:34.
[37] Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. Immunity 2018;48(812-830):e814.
[38] Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res 2020;48. W509-w514.
[39] Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep 2017;18:248–62.
[40] Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol 2017;18:18–30.
[41] Aarntzen EH, De Vries IJ, Lesterhuis WJ, Schuurhuis D, Jacobs JF, Bol K, et al. Targeting CD4(+) T-helper cells improves the induction of antitumor responses in dendritic cell-based vaccination. Cancer Res 2013;73:19–29.
[42] Forero A, Li Y, Chen D, Grizzle WE, Updike KL, Merz ND, et al. Expression of the MHC Class II Pathway in triple-negative breast cancer tumor cells is associated with a good prognosis and infiltrating lymphocytes. Cancer Immunol Res 2016;4:390–9.
[43] Johnson DB, Estrada MV, Salgado R, Sanchez V, Doxie DB, Opalenik SR, et al. Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy. Nat Commun 2016;7:10582.
[44] Xie N, Shen G, Gao W, Huang Z, Huang C, Fu L. Neoantigens: promising targets for cancer therapy. Signal Transduct Target Ther 2023;8:9.
[45] Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature 2017;547:217–21.
[46] Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, et al. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature 2019;565:234–9.
[47] Zhu Y, Qian Y, Li Z, Li Y, Li B. Neoantigen-reactive T cell: an emerging role in adoptive cellular immunotherapy. MedComm 2021;2(*2020*):207–20.