INTRON RETENTION INDUCED NEOANTIGEN AS BIOMARKERS IN DISEASES

Chuanpeng Dong

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

August 2022

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____
Jingwen Yan, PhD, Chair

_____
Yunlong Liu, PhD

April 27, 2022

_____
Kun Huang, PhD

_____
Jun Wan, PhD

_____
Xiaowen Liu, PhD

ACKNOWLEDGEMENT

I would like to express my gratitude to all those who helped me during my PhD study. My deepest gratitude goes first to my advisor Professor Yunlong Liu, for his constant encouragement and mentorship. I would like to thank you for shaping my research taste and nurturing me into an independent scientist. Your advice both on my career as well as on research has been priceless. I would also like to thank all of my committee members, Professor Jingwen Yan, Professor Kun Huang, Professor Jun Wan, and Professor Xiaowen Liu for serving as my committee members. I also want to thank you for your brilliant comments and suggestions. I would like to thank Professor Huanmei Wu and all the faculty members of the School of Informatics and Computing. Thanks for their tremendous effect in design and delivering excellent course material.

A special thanks to my family. I cannot even reach here without their unreserved support and love. I would like to express appreciation to my beloved wife for her loving consideration and firm confidence in me for the past decades. You always provide support at the moment I need it most. I would also like to thank all of my friends who have helped me and shared with me my worries, frustrations, and happiness. Their supports encourage me to strive towards my goal through these years.

Chuanpeng Dong

INTRON RETENTION INDUCED NEOANTIGEN AS BIOMARKERS IN DISEASES

Alternative splicing is a regulatory mechanism that generates multiple mRNA transcripts from a single gene, allowing significant expansion in proteome diversity. Disruption of splicing mechanisms has a large impact on the transcriptome and is a significant driver of complex diseases by producing condition-specific transcripts. Recent studies have reported that mis-spliced RNA transcripts can be another major source of neoantigens directly associated with immune responses. Particularly, aberrant peptides derived from unspliced introns can be presented by the major histocompatibility complex (MHC) class I molecules on the cell surface and elicit immunogenicity. In this dissertation, we first developed an integrated computational pipeline for identifying IR-induced neoantigens (IR-neoAg) from RNA sequencing (RNA-Seq) data. Our workflow also included a random forest classifier for prioritizing the neoepitopes with the highest likelihood to induce a T cell response. Second, we analyzed IR neoantigen using RNA-Seq data for multiple myeloma patients from the MMRF study. Our results suggested that the IR-neoAg load could serve as a prognosis biomarker, and immunosuppression in the myeloma microenvironment might offset the increasing neoantigen load effect. Thirdly, we demonstrated that high IR-neoAg predicts better overall survival in TCGA pancreatic cancer patients. Moreover, our results indicated the IR-neoAg load might be useful in identifying pancreatic cancer patients who might benefit from immune checkpoint blockade (ICB) therapy. Finally, we explored the association of IR-induced neo-peptides with neurodegeneration disease pathology and susceptibility. In conclusion, we presented

a state-of-art computational solution for identifying IR-neoAgs, which might aid neoantigen-based vaccine development and the prediction of patient immunotherapy responses. Our studies provide remarkable insights into the roles of alternative splicing in complex diseases by directly mediating immune responses.

Jingwen Yan, PhD, Chair

Yunlong Liu, PhD

Kun Huang, PhD

Jun Wan, PhD

Xiaowen Liu, PhD

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Alzheimer's disease
AS	Alternative splicing
AUC	Area under the curve
AUD	Alcohol use disorder
AUPRC	Area under the precision-recall curve
CCLE	Cancer Cell Line Encyclopedia
CDS	Coding sequence
COGA	Collaborative Studies on Genetics of Alcoholism Study
CTLA-4	Cytotoxic T-lymphocyte-associated protein 4
DAI	Differential agretopicity index
DNA	Deoxyribonucleic Acid
dsRNA	Double-strand RNA
GBM	Glioblastoma
GEO	Gene Expression Omnibus
GTEx	Genotype-Tissue Expression Project
GWAS	Genome-wide association studies
HLA	Histocompatibility leukocyte antigen
hnRNPs	Heterogeneous nuclear ribonucleoproteins
HR	Hazard ratio
HRD	Chromosomal hyperdiploidy
ICGC	International Cancer Genome Consortium
IEDB	Immune Epitope Database

INDEL	Insertion/Deletion
IR	Intron retention
IR-neoAg	Intron retention induced neoantigen
ISS	Myeloma International Staging System
MHC	Major histocompatibility complex
MM	Multiple myeloma
MMRF	Multiple Myeloma Research Foundation CoMMpass Study
MSBB	Mount Sinai Brain Bank study
MsigDB	Molecular Signatures Database
NDMM	Newly diagnosed multiple myeloma
ORF	Open reading frame
OS	Overall survival
PD1	Programmed cell death protein 1
PDAC	pancreatic ductal adenocarcinoma
PD-L1	Programmed death-ligand 1
PSI	Percent splicing
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
ROC	Receiver operating characteristic
ROSMAP	Religious Order Study and the Memory and Aging Project
RPKM	Reads per kilo base per million mapped reads
SNV	Single nucleotide variant
SRSFs	Serine/arginine-rich splicing factors

ssGSEA	Single-sample gene set enrichment analysis
TCGA	The Cancer Genome Atlas
TMB	Tumor mutation burden
TPM	Transcripts per million

Chapter 1 Introduction

1.1 Background

Alternative splicing is a regulatory mechanism that generates multiple mRNA transcripts from a single gene, allowing significant expansion in proteome diversity [1]. Dysregulation of ribonucleic acid (RNA) splicing commonly occurs in tumor transcriptomes, producing various protein isoforms that promote tumor cell survival and proliferation [2, 3]. Disruption of the splicing mechanism can greatly impact the transcriptome and be a significant driver of disease. Most recently, accumulating studies on RNA alternative splicing, especially intron retention, may function a new role in inducing cell immunity in multiple ways [4].

Clinical response to checkpoint blockades has been shown to be highly associated with the presence of tumor-specific antigenic peptides, or neoantigens, which are thought to enhance the efficacy of cancer immunotherapy by providing endogenous tumor-specific immune targets [5]. Neoantigens have the potential to be loaded on major histocompatibility complex (MHC) class I molecules and presented to cytotoxic T-cells as immunogenic targets that can generate an anti-tumor immune response [6]. Cytotoxic T-cells activated against tumor-specific neoantigens can kill tumor cells presenting mutagenic peptides and lead to anti-tumor immunological memory that resists tumor recurrence [7]. Deoxyribonucleic acid (DNA) level alternation as somatic mutations in coding regions of the genome has the potential to generate endogenous MHC class I neoantigens [8]. However, it has been reported that increasing mutation and neoantigen load were found to not correlate with prognosis in multiple cancers [9]. Recently research and computation efforts mainly focused on DNA mutation induced neoantigen, the

potential of other forms of neoantigen was not fully investigated. Intron retention (IR) occurs when the splicing complex fails to remove introns from the primary messenger RNA transcript [10]. This type of aberrant splicing results from cis-acting mutations at or near a splicing site or dysregulation of trans-acting splicing regulators [11]. Most aberrant intron retention events introduce premature termination codons in the resulting transcripts that trigger degradation of the partially translated protein into peptides via nonsense-mediated decay [12]. Such novel tumor cell-specific peptides can be presented on the cell surface by the MHC class I molecules encoded by the histocompatibility leukocyte antigen (HLA) genes. Additionally, some recent studies have shown that unspliced intron tends to form double-strand RNA, which may trigger the anti-virus-like effect in triple-negative breast cancer cells [13]. Identification of IR events and IR-neoAg will provide important aid to selecting patients for immunotherapies and inspiring new treatment strategies towards various diseases.

1.2 Develop In-silico Tools for Identifying Intron Retention Induced Neoantigen from RNA-Seq Data

1.2.1 Significance

In addition to DNA level genomic mutations, dysregulation of RNA splicing commonly occurs in tumor transcriptomes. IR occurs when the splicing complex fails to splice introns from the primary messenger RNA transcript [14], producing tumor cell-specific neo-peptides that can be presented on the cell surface by MHC class I molecules. The IR-induced neo-peptides may provide suitable targets for designing personalized cancer vaccines and predicting patients that can benefit from immunotherapy [4].

1.2.2 Critical Barrier

Most in silico approaches for neoantigen prediction focus on identifying DNA mutation (SNV/INDEL) derived neoantigens, such as pVAC-Seq, Neopepsee pTuneos, and ScanNeo. Attempts in predicting splicing-derived neoantigen have not been reported until most recently. One study based on 8705 patients in the Cancer Genome Atlas (TCGA) cohort revealed that tumors samples carried more alternative splicing events, presenting a large new class of neoantigens that could be exploited in cancer immunotherapy. Smart et al. proposed intron retention as a neoepitope source and developed their pipeline using a pseudo-alignment tools-based approach [4]. Zhang et al. have developed a computing pipeline for identifying novel splicing isoform-derived neoantigen relying on the assembly of novel junctions into existing isoforms [15]. Both methods depended on transcripts' accuracy, which is still a challenge in computational biology. Therefore, unbiased intron-centric tools for identifying high-quality splicing, especially IR-neoAgs, are urgently needed.

1.2.3 Innovation

Herein, we present an unbiased intron-centric computational tool for identifying and prioritizing the IR-neoAg from RNA-seq. The program enables an all-in-one workflow for identifying patient-specific, IR-derived neo-peptides and a refined random forest classifier for prioritizing the neoepitope with the highest probability to induce T cell response. This work may foster the identification of new neoantigen targets for immunotherapy with huge clinical utilizing potential.

1.3 Intron Neoantigen as Biomarkers for Multiple Myeloma

1.3.1 Significance

Neo-peptides generated from somatic mutations have shown potential as personalized cancer vaccines and a positive predictor of response in immune checkpoint therapy [16]. Disruption of the normal splicing patterns of RNA is reported as a potentially important driver mechanism in multiple myeloma (MM) [17]. Recently studies also highlighted tumor-specific splicing presents a large new class of splicing-associated potential neoantigens that may affect the immune response.

1.3.2 Critical Barrier

DNA level alternation as somatic mutations in coding regions of the genome has the potential to generate endogenous MHC class I neoantigens. Previously studies also demonstrate mutation-derived neoantigen can elicit T-cell responses in MM, with a conflicting discovery that high mutation neoantigen burden is correlated with unfavorable prognosis in MM. Dysregulation of RNA splicing via retaining introns, which is common in hematological cancer transcriptomes, represents another potential source of neoepitopes but has not been previously explored in MM. Therefore, we sought to investigate the landscape of IR and IR-neoAg landscape in MM patients and their relationship with clinical outcomes.

1.3.3 Innovation

We demonstrate that IR-neoAgs in MM is associated with unfavorable survival in primary and relapse MM. Our results also provide molecular clues of how MM may escape the immune system by expressing more inhibitory genes and downregulating MHC-II

genes. These discoveries will assist in developing treatment strategies of immune checkpoint therapy and cancer vaccines in patients with multiple myeloma.

1.4 Intron Neoantigen in Pancreatic Cancer

1.4.1 Significance

Clinical response to checkpoint blockades has been shown to be highly associated with the presence of tumor-specific antigenic peptides, or neoantigens, which are thought to enhance the efficacy of cancer immunotherapy by providing endogenous tumor-specific immune targets [5]. Intron retention (IR), resulting from aberrant RNA splicing in cancer cells, is another source of neo-peptides that could potentially trigger an immune response [18]. Systematically investigation of intron neoantigen may provide novel targets for pancreatic cancer immunotherapy.

1.4.2 Critical Barrier

Tumor neoantigen load frequently correlates with the efficacy of immune checkpoint blockade therapy (ICB) in many cancer types. Although mutation-derived neoantigen load is not associated with prognosis in pancreatic ductal adenocarcinoma (PDAC), recent studies have suggested that some pancreatic cancer cells express neoantigens that could elicit intratumoral T-cell responses. Due to the low mutation load in pancreatic cancer, we have to search for a new source of genetic events that can introduce neoepitopes. Most recent studies suggested aberrant transcripts with retained introns are translated and degraded through the nonsense-mediated decay mechanism, which can be a major source of neo-peptides. However, whether IR-neoAgs could elicit an immune response and affect cancer prognosis is still unknown.

1.4.3 Innovation

In the section, we demonstrated that high IR-neoAg load, but not mutation-derived neoantigen load, predicted better overall survival (log-rank p=0.011). Further analysis revealed patients with both high IR-neoAg load and low expression of inhibitory checkpoint genes had the longest overall survival time. Moreover, our results indicated that high IR-neoAg load and high HLA class I gene expression could select PDAC tumors correlated with response to anti-PD1 checkpoint therapy. Our findings demonstrate that high IR-neoAg load is associated with more prolonged overall survival in PDAC patients. IR-neoAg could serve as a biomarker for selecting PDAC patients who might benefit from ICB therapy.

1.5 Objective

The main objective of this dissertation is to explore the novel roles of intron retention in cell immunity. We conducted our research through the following three aspects: 1). develop in silico tools for identifying IR-neoAg from RNA-Seq data; 2). explore the IR-neoAg as a biomarker for cancer prognosis and immunotherapy, and 3). determine the potential of IR as a biomarker in other diseases such as neuron degeneration and diabetes.

Chapter 2 reviews the mechanism and regulation of RNA alternative splicing and its association with a wide range of diseases. It introduces the general knowledge about gene expression and regulation procedure of eukaryotic cells, emphasizing the significant role of alternative splicing in post-transcriptional regulation. It also covers the molecular mechanism of RNA splicing, the cis-and trans regulation of alternative splicing, and its association with various diseases. Moreover, we discussed the novel role of intron retention in cell immunity reported in the most recent literature.

Chapter 3 describes the major findings we developed IntronNeoantigen, a computational pipeline for prioritizing tumor neoantigens from RNA sequencing data. Based on the putative neoantigens obtained by high-peptide-MHC binding affinity, IntronNeoantigen conducts further machine learning classifiers to predict neoantigens with high potentials for T cell recognition. We demonstrated the utility of IntronNeoantigen by applying it on two melanoma cohorts undergoing checkpoint blockade immunotherapy. Further analysis on the Cancer Genome Atlas cohorts revealed that the combination of high-quality IR-neoAgs and CD8+ T cells could infiltrate advanced-stage melanoma and glioblastoma (GBM) with the longest survival. This work provides a user-friendly solution for prioritizing IR-neoAgs from tumor RNA-seq data, which may aid cancer research in immunotherapies and vaccines development.

In Chapter 4, we analyzed the IR and predicted IR-neoAg load in 892 multiple myeloma samples of MMRF ComPass study using RNA sequencing data. We found the increasing IR events in MM patients compared with healthy control, and IR-neoAg load was correlated with significantly decreased overall survival (OS) time, consistent with mutation-derived neoantigen. Further, we sought to investigate the underlying mechanism of why neoantigen load cannot trigger cytotoxic T cell killing and predict better clinical outcomes. We found the T cell inhibitory molecules expression positively associated with the IR load in multiple myeloma. Results from this section demonstrate that the IR-neoAg load could function as a prognosis biomarker and uncover the immunosuppression in the MM microenvironment offset the increasing neoantigen load effect.

In Chapter 5, we quantified both somatic mutation-derived neoantigen and IR-neoAg loads by analyzing RNA sequencing data from the PAAD cohorts in the TCGA and

the International Cancer Genome Consortium (ICGC). We demonstrated that high IR-neoAg load, but not mutation-derived neoantigen load, predicted better OS. Further analysis revealed patients with both high IR-neoAg load and low expression of inhibitory checkpoint genes had the longest OS time. Moreover, our results indicated that the high IR-neoAg load and high HLA class I gene expression might be associated with PDAC tumors responding to anti-PD1 checkpoint therapy. Our findings demonstrate that high IR-neoAg load is associated with longer overall survival in PDAC patients. In addition, the IR-neoAg load may be useful in identifying PDAC patients who might benefit from immune checkpoint blockade (ICB) therapy.

Chapter 6 illustrated the potential usage of IR and IR neo-peptides in neurodegeneration diseases from our preliminary study. We found the IR number was increased in Alzheimer's disease (AD) and alcohol use disorder (AUD) brains. Also, the IR number was associated with malignancy of AD pathology. We further discussed the potential of IR whole-genome neoantigen load for a personalized disease risk prediction.

Chapter 7 concludes that our research is valuable for understanding intron retention as a biomarker in diseases. Our analysis in cancer and neurodegeneration diseases highlights the potential of using IR as a powerful tool for diagnosis, prognosis, and immunotherapy prediction. This chapter also discusses the further direction of our study.

Chapter 2 Literature Review

2.1 Current Progress in Cancer Immunotherapy

Avoiding immune destruction is one of the well-known hallmarks of cancer cells [19]. Under normal circumstances, the human immune system can recognize and clear necrotic or cancerous cells. However, tumor cells have evolved different strategies to escape the surveillance and killing of the immune system and can survive and proliferation through all aspects of the immune response [20]. Cancer immunotherapy has rapidly proved to be a new way to against malignant tumors by boosting immune systems, which have shown great therapeutic efficacy in multiple human cancers. Immunotherapies can also be further categorized into active immunotherapy and passive immunotherapy. The former one includes immune checkpoint blockade (ICB) therapy, adoptive chimeric antigen receptor (CAR) T-cell therapy and oncolytic viruses. Active immunotherapy aims to stimulate the body's immune system to elimination of cancer cells, such as cancer vaccines [21].

In 2010, the field was reinvigorated by a landmark randomized clinical trial that showed that cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) inhibitor ipilimumab could dramatically improve overall outcomes in patients with metastatic melanoma [22]. Early clinical trials of antibodies for checkpoint molecules programmed cell death protein 1 (PD-1) and programmed death-ligand 1 (PD-L1) have demonstrated clinical activity in a variety of tumor types including melanoma, lung, bladder, et al [23]. Clinical results of the individualized tumor vaccine have obtained positive data in melanoma [24] and glioma [25], which is exciting and adds new hope for the success of the human anti-tumor cause. The implicit conclusion from these clinical data is that, in a substantial proportion of

patients, the endogenous T cell compartment is capable of recognizing epitopes/neoantigens displayed on the major histocompatibility complex (MHC) on the surface of malignant cells.

2.2 Concepts of Neoantigen

Tumor neoantigen, as the name suggests, is a new class of antigen compared to "old" antigens as tumor associated antigens. The traditional tumor-associated antigens are a class of antigens that are lowly expressed in normal cells and highly expressed in tumor cells, which is more likely to cause autoimmune side effects and immune tolerance [26]. While the biggest feature of neoantigen is that it is uniquely expressed by tumor cells and not expressed by other normal cells [27].

Neoantigens are non-self-specific proteins produced by non-synonymous mutations in tumor cells [28]. Tumor cells generally have a large number of somatic mutations aligned with the uncontrolled malignant proliferation of cancer cells. On the basis of gene mutation, tumor cells often produce proteins with specific amino acid sequence variations. These mutated proteins are hydrolyzed by intracellular proteases, combined with MHC molecules, and presented to the cell surface, where they are then specifically T cells recognize and activate cancer immune responses. Generally, the greater the difference between the mutated sequence and the original coding sequence, the more obvious the "non-self" feature of the abnormal peptides, and the stronger the immunogenicity [29].

2.3 Challenge of Neoantigens Application in Immunotherapy

Antibody-mediated ICB eliminates the inhibitory effect of tumor cells on immune cells and augments anti-tumor effects. However, only a small fraction of patients with

solids tumors shows a response to ICB [30, 31]. Recent studies have linked mutation and neoantigen loads with clinical benefit in patients receiving ICB treatments in cancers [32]. Tumors with high somatic mutation numbers can produce more neoantigens and are most likely to be recognized by self-immune system, and thus might impact the response to immune therapy (Fig. 1). Snyder et al reported the association between somatic mutation neoantigen with clinical response to anti-CTLA4 in melanoma patients. Similarly, Rizvi et al demonstrated that mutation neoantigen could predicts response to anti-PD1 therapy in non-small cell lung cancer (NSCLC) patients [11].

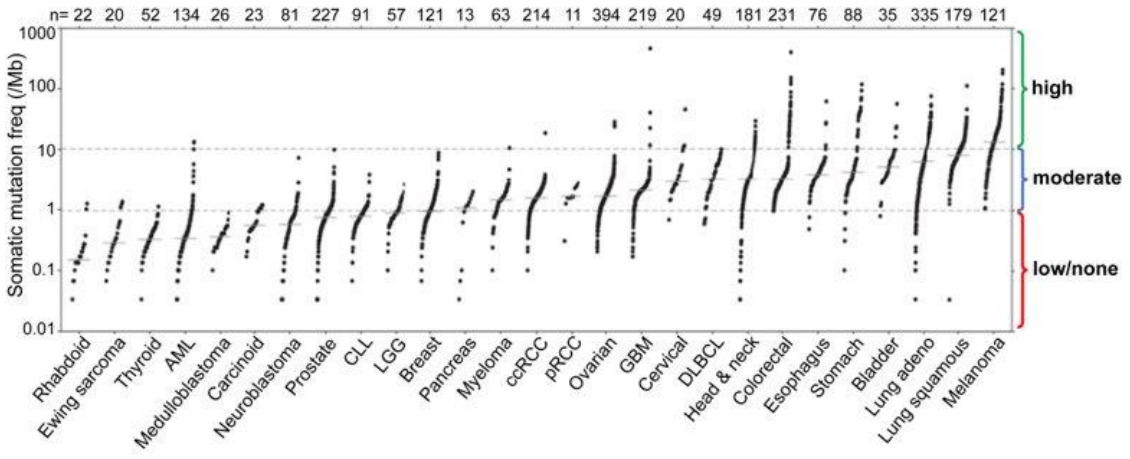

Figure 1 Tumor mutational load across various tumor types. The right green/blue/red box show tumor response level to immune checkpoint blockade therapy. Green boxes show some tumor types with high somatic mutation frequencies are associated with high response rates. Adopted from Lee et al. Trends in Immunology (2018) [32].

But the responses to ICB have not been restricted cancers with high mutation load. For example, renal cell carcinoma is an immunotherapy-responsive with a moderate mutation frequency (< 10% mutations numbers of melanoma) [32, 33]. It is clear that mutational neoantigens alone may not explain response to checkpoint inhibitor immunotherapy. Noteablly, only a small proportion non-synonymous mutations can produce high affinity neoantigens in high mutational tumors [26]. There is an urgent need to discover the underestimated source of tumor neoantigens.

A diversity of tumor-specific alterations may serve as suitable sources for non-pathogen-derived neoantigens, including single-nucleotide variant (SNV), insert and deletion (INDEL) frame shifts, chromosomal translocations, RNA splicing, and posttranslational modification [34]. SNV accounts for 95% of tumor cell mutations, but they only cause a few changes in amino acid sequence and spatial structure. Moreover, only a small fraction of patients had a considerable tumor mutation load in some cancer types such as glioblastoma (GBM) [35] and pancreatic cancer [36].

Low immunogenicity neoantigen is one of the key reasons for tumor immune escape [36]. Therefore, the search for more immunogenic neoantigens has become an urgent need for immunotherapy.

2.4 Aberrant Splicing Serves as Source of Neoantigen

2.4.1 Widespread RNA splicing in eukaryotic cells

Alternative splicing (AS) is one of the most widespread events of post-transcriptional gene regulation [37]. RNA alternative splicing is the process of mRNA maturation through intron removal and exon ligation. In eukaryotes cells, AS acts like a pair of "scissors", "pruning" the pre-mRNA, greatly enriching the diversity of the

transcriptome and proteome, resulting in functional diversity. Also, AS is an important mechanism for regulating gene expression and generating proteome diversity and plays a key role in tissue- and species-specific differentiation patterns [38]. AS is a post-transcriptional process that allows one gene to code multiple transcripts and proteins. For example, the latest GENCODE [39] human genome annotation (v39, release date 12.2021) contains 244939 transcripts assigned to 61533 human genes. There is an average of 3.98 transcripts per gene. When it comes to the protein-coding genes, there are 87151 transcripts assigned to 19982 human protein-coding genes, with an average of 4.36 transcripts per protein-coding gene. Additionally, the gene PCBP1-AS1(ENSG00000179818), an RNA Gene affiliated with the lncRNA class, has the maximum number of transcripts of 283. MAPK10 is the protein-coding gene with the highest transcripts number of 192.

2.4.2 Molecular mechanism of alternative splicing

In general, the pre-mRNA is spliced by the spliceosome to cut the intron, connect the exon, and form a mature mRNA. The spliceosome is a large RNA-protein complex that catalyzes the removal of introns from nuclear pre-mRNA (Fig. 2). In addition to the well-known mRNA, there are actually a variety of RNA types in the nucleus, one of which is a special type of small nuclear RNA, characterized for its high uracil content. snRNAs are usually associated with multiple proteins to form a small nuclear riboprotein. At RNA splice sites, U1, U2, U4, U5, and U6 snRNPs and other non-snRNP proteins assemble step-by-step, eventually forming spliceosomes for splicing introns and joining exons.

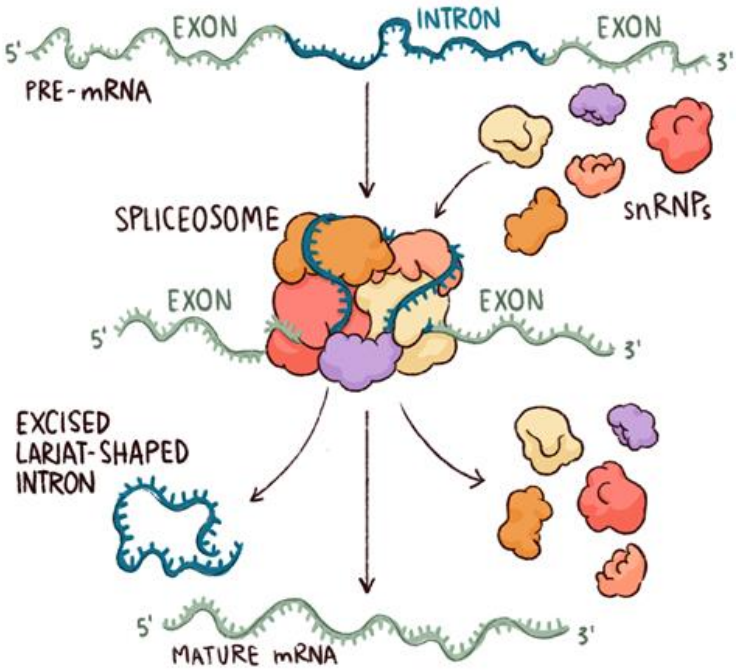

Figure 2 Molecular machinery of spliceosome. The process of spliceosome removes non-coding segments from the strand and then joins the other sections together. Source Claudia Flandoli, Clare Samson 2020 Chemistry World.

2.4.3 Alternative RNA splicing types

In organisms, there are mainly five types (Fig. 3) of AS being depicted [40]:

- Skipping exon (***SE***): Exons are skipped during splicing of precursor mRNA to form mature mRNA and eventually do not appear on some mature mRNAs. SE is the most prevalent AS type.
- Intron retention(***IR***): part of the intron is retained in mature mRNAs;
- Mutually exclusive exon(***MXE***): inclusive exons cannot exist in the same transcript simultaneously but can only exist separately in different transcripts;
- Alternative 5' splice site(***A5SS***): alternative splicing at the 5' end;
- Alternative 3' splice site(***A3SS***): alternative splicing at the 3' end.

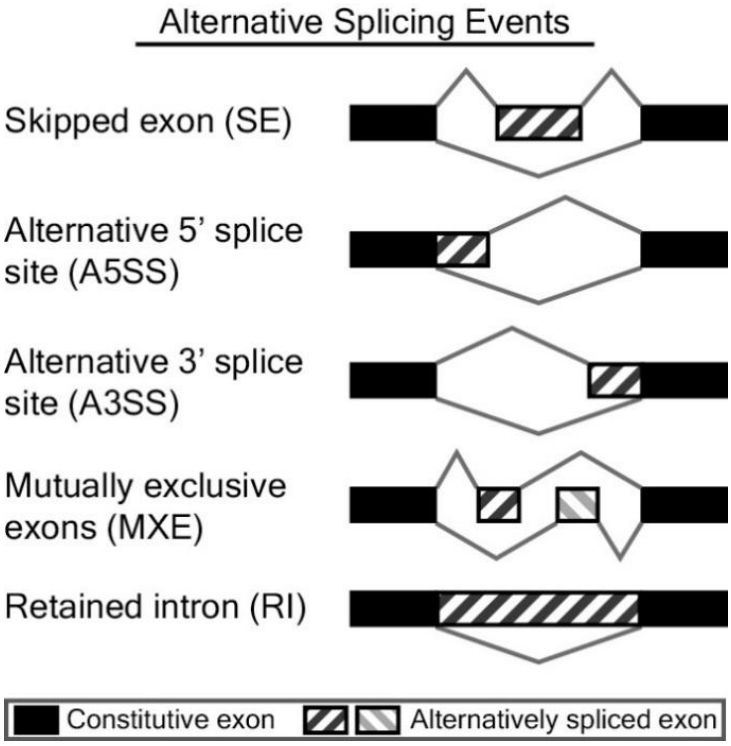

Figure 3 Different types of alternative splicing events. Source Shen S. et al. PNAS, 2014 [40].

2.4.4 Alternative splicing and common disease

Aberrant AS can substantially affect normal cellular functions and cause abnormal cell growth. A recent review reported dysregulated splicing features of cancer hallmarks, suggesting a critical role of AS in cancer biology [41, 42]. For example, the gene integrin subunit α6 (ITGA6) can produce two isoforms via alternative splicing: ITGA6A and ITGA6B. ITGA6A was reported upregulated and attributed to pro-proliferative function for colon cancer cells [43]. Previous studies have shown that splicing factors such as PTB/nPTB and SRSF3 could modulate the PKM1 and PKM2 ratio by alternatively splicing two mutually exclusive exons, suggesting the role of RNA splicing involved in the regulation of tumor metabolism [44]. Also, there is accumulated evidence that reported somatic mutations in components of the human splicing machinery gene were associated with multiple human solid tumors [45], including bladde [46], brain [47], breast [48], colon [49], liver [50], lung [51] as well as hematological malignancies [17].

Alternative splicing has been associated with a wide range of neurological disorder diseases [41]. For example, amyloid precursor protein (APP) plays an important role in AD development, with changes in APP isoform ratios associated with AD cases. AS of APP transcripts at exon 15 can alter the production and ratio of the specific APP isoforms, which will affect amyloid plaque accumulation in AD [52].

Disruptions RNA splicing of genes related to pancreatic β cell apoptosis and endocrine function were associated with type I and type II diabetes [53, 54]. For example, dysregulation of serine/arginine-rich splicing factors (SRSFs), heterogeneous nuclear ribonucleoproteins (hnRNPs) are linked to type I diabetes by mediating splicing apoptosis-

regulating genes such as caspase-2 and caspase-9, giving rise to isoforms with pro-apoptotic and anti-apoptotic activities of pancreatic β cells [55].

2.4.5 Alternative splicing and cancer immunity

AS has been reported to be closely linked with immune cells development and differentiation. A survey from the ImmGen Consortium found that alternative splicing is ubiquitous through high throughput sequencing. About 60% of genes show different alternative splicing isoforms in T or B cells [58], of which about 70% of alternative splicing events are related to lineage differentiation.

In addition to specific isoforms that could play a critical role in the immune cells, the products of aberrant splicing in cancer cells are also the potential new immunogenic targets for the immune system. Most recent studies revealed that a high burden of mis-spliced RNA (especially IR) might be an unexplored feature of some cancers that engage in adaptive immunity response and tumor antiviral signaling [4, 13]. Bowling et al. report that splicing inhibitor treatment in triple-negative breast cancer cell lines could elevate RNA splicing errors and promotes cytosolic accumulation of intron retained transcripts that form double-stranded RNAs. The increased intron-induced double-stranded RNAs can boost antiviral signaling and extrinsic apoptosis in cancer cells, inspiring that anti-splicing could function as a novel approach to augment cancer cell immunogenicity [13].

Importantly, aberrant splicing has been reported involving in adaptive immunity by generating neoantigen as well. One study based on 8,705 patients in the TCGA cohort found that tumor samples carried more AS events than healthy tissue [56]. Tumor-specific RNA splicing presents a large new class of potential neoepitopes that could be exploited in immunotherapy. A recent study showed that retained intron neoepitopes could be present

on MHC I on the surface of cancer cell lines using RNA-Seq and proteomics data. RNA splicing-induced neoepitopes could be used for personalized cancer vaccine development [4]. This thesis mainly focused on investigating aberrant splicing inducing neoantigen in the following sections.

2.5 Identification of Alternative Splicing Events and Neoantigens

2.5.1 Alternative splicing events from RNA-Seq

Due to the popularity of Illumina RNA-seq, many computational tools have been developed for estimating mRNA isoform expression and quantifying alternative splice variants using short-read RNA-seq data.

Transcript-based tools attempt to estimate the abundance and relative proportions of full-length mRNA isoforms using short-read RNA-seq data [57]. A disadvantage of transcript-based approaches is that it is not easy to infer full-length mRNA isoforms from short reads. Event-based tools that attempt to quantify individual AS events using RNA-seq data. This approach estimates alternative splicing events based on reads aligned to specific exons or splice junctions. In the event-based analysis, a widely used metric is percent splicing (PSI or ψ), which expresses the percentage of a gene's mRNA transcripts containing a particular exon or splice site. Popular computational tools for alternatively spliced RNA sequence analysis include MISO [58], rMATS [59], MAJIQ [60], SUPPA [61], LeafCutter [62] et,. al.

2.5.2 In-silico identification of neoantigen

Due to the progress of sequencing technology and bioinformatics algorithms, scientists have developed numerous tools to identify the potential neo-peptides by comparing the DNA or RNA sequences from tumors and normal cells [63]. Factors such

as sequencing depth, tumor tissue quality, source of sequencing materials, and single-nucleotide mutation algorithms will all affect the identification of neoantigens. Currently, most neoantigens identification tools contain the following steps:

1.Whether the mutant gene sequence can be translated into peptides and chopped into small peptide fragments.

2.The affinity of the mutant neo-peptide to the patient's MHC molecule.

3.The affinity of the mutant neo-peptide-MHC complex with T cell receptor.

The prediction of neoantigens requires not only the identification of expressed mutations but also the patient-specific HLA genotypes [64]. There are also a variety of tools that can help to screen and identify neoantigens, such as workflows that combine whole-exome sequencing and RNA sequencing to screen for neoantigens, such as pVAC-Seq [65], MuPeXI [66], Neopepsee [67], pTuneos [68], and ScanNeo [69].

Chapter 3 Development Tools for Identifying Intron Neoantigen from RNA-Seq Data

3.1 Introduction

Neoantigens are normally referred as tumor-specific peptides that arise from the expression of mutated genes and are recognized by cytotoxic T cells. Recently, the roles of neoantigen in other diseases, such as type I diabetes [70] and rheumatoid arthritis [71] , also start to emerge. Cytotoxic T cells can directly and specifically lyse abnormal cells upon binding to antigenic neoantigens presented by the major histocompatibility complex (MHC) on the cell surface. In cancer, these antigenic neoantigens are novel targets for cancer therapy[72] . Recently, neoantigens have been reported to play a critical role in mediating the response of multiple cancers to immune checkpoint blockade (ICB) [30, 31, 36, 73-75]. Moreover, vaccines developed with neoantigens have several advantages, such as multi-targeting capacity, good safety, and broad-spectrum application for treating melanoma and glioma patients [76, 77] .However, basic research and computational studies have mainly focused on DNA mutation-induced neoantigens. The potential effect of other forms of neoantigens has yet to be fully investigated.

RNA modification, especially AS, contributes to proteomic diversity and can be a major source of neoantigens [78, 79]. RNA splicing dysregulation commonly occurs in the transcriptome of disease tissues and produces protein isoforms that lead to aberrant phenotypes [2, 3] . IR is a form of AS that occurs when the splicing complex fails to remove introns from the primary messenger RNA transcript. This type of aberrant RNA splicing results from the dysregulation of *trans*-acting splicing regulators or *cis*-acting mutations at or near a splicing site [11] . Many aberrant IR events introduce premature termination

codons in the product transcripts that trigger the degradation of the partially translated protein into peptides via nonsense-mediated decay [12]. The resulting novel disease-specific peptides can then be presented on the cell surface by the MHC class I molecules encoded by the HLA genes. The identification of IR-neoAgs will provide important information for understanding the etiology of complex diseases, and for the selection and development of therapeutical strategies, such as immunotherapies and vaccines for cancers when DNA mutations are rare.

Here, we present IntronNeoantigen, an integrated computational pipeline for identifying and prioritizing IR-neoAgs from RNA-seq data. IntronNeoantigen utilizes an all-in-one workflow for identifying IR-derived neo-peptides, patient-specific MHC-I alleles, and a refined random forest classifier for prioritizing the neoepitopes with the highest likelihood to induce a T cell response.

3.2 Materials and Methods

3.2.1 Reparsing an intron-centric gene annotation

The gene annotation files with detailed coordinate information were obtained from GENCODE (GRCh38, v32 from https://www.gencodegenes.org/). The exon and intron coordinates of human protein-coding genes were re-annotated with the following steps: the GTF files were parsed by first merging the exons from every transcript of the same gene to generate a union set of exons. Next, the set of introns was obtained by extracting the intervening sequences from the union exons located in the coding sequence (CDS) regions (Fig. 4-A).

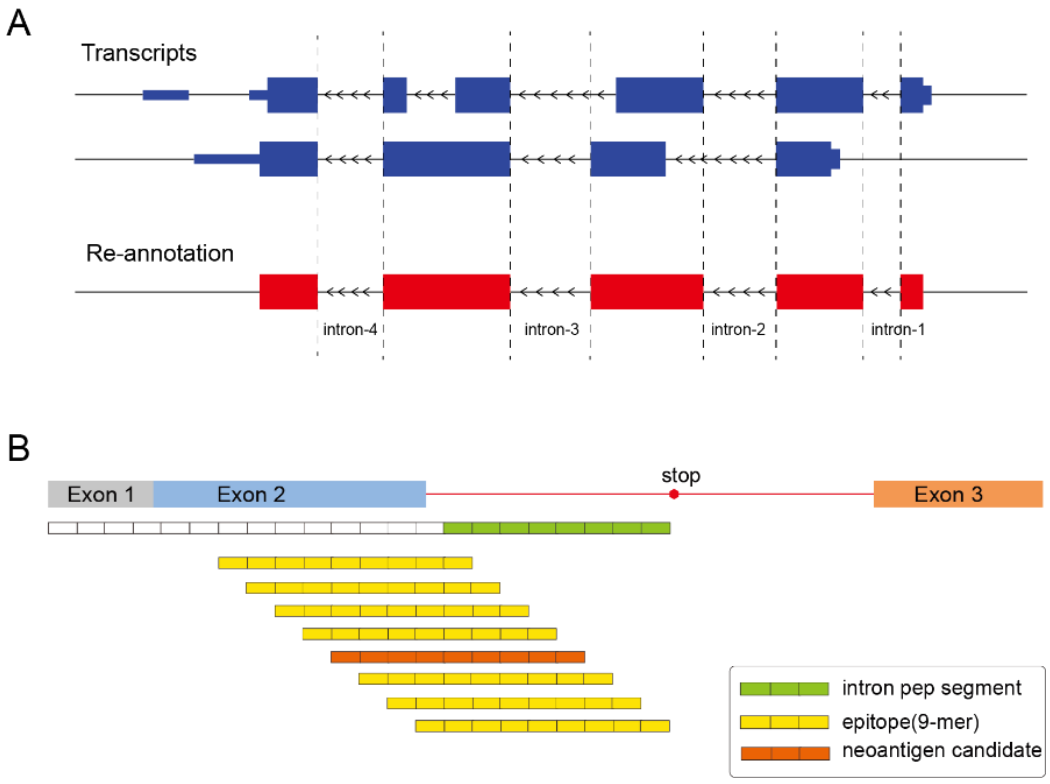

Figure 4 Extraction of intron information and generation of intron-derived peptide sequences for neoantigen prediction. **A** For protein-coding genes with multiple transcripts, exons (shown as boxes) were re-annotated as union exon sets by considering all alternatively spliced protein-coding transcripts; the set of introns (shown as thin lines) was then obtained by extracting the intervening sequences from the union exons in coding regions. **B** Retained introns were translated into peptides that were in-frame with the upstream exon and were terminated at the first in-frame stop codon (2nd line). Intron-derived peptides between 8 to 11 amino acids were generated from the exon-intron ORF, containing at least one amino acid from the intron region (green boxes).

3.2.2 Identifying intron retention events using RNA-seq data

The fastq files are first aligned to the reference genome using a splice-aware RNA-seq mapper, such as STAR [80] or TopHat [81] . We use only the uniquely and perfectly aligned reads with a number of hits (NH) tag value equal to 1 or a mapping quality value (MAPQ) larger than 30. Using the re-annotated intron-centric gene annotation file, we quantify the reads covering individual exon and intron regions using [82]. Only reads with at least ten bases in the intron region were counted for the following analysis. The IR events were selected with the following criteria (1) read counts on flanking exonic regions larger than 10; (2) read count of the intronic region larger than 10; and (3) a ratio of the TPM (transcript per million) values of the intron and flanking exons between 0.05 and 0.5. We hope these criteria will help capture the high confident IR events with considerable abundance.

3.2.3 IR-derived neo-peptide

Each intronic sequence is translated into peptides using the open reading frame (ORF) of the upstream exon (Fig. 4-B). The ORF for each exon is determined as the one shared with the most protein-coding transcripts (the ORF of the longer transcripts is selected if more than one ORF existed with the same usage frequency). The translated peptides are segmented into 8-11 amino acid peptides that contained at least one intron-encoded amino acid. We further discard intron-derived peptides that could be generated from human reference proteins. The protein-coding transcript translation sequences from Gencode (GRCh38, v32) are used as reference proteins.

3.2.4 Evaluating binding affinity between patient-specific HLA-I allele and IR-derived neo-peptide

HLA genotype for each sample is inferred from RNA-seq data using arcasHLA [64] with default parameters. The IntronNeoantigen presentation function is used to estimate the binding affinity of the intron-derived peptides against HLA A/B/C alleles of the same patient using netMHCpan (v4.1) [83] . Peptides with a percentile rank affinity score less than 0.5 are reported as strong peptide candidates. The rank of the predicted binding score is calculated by comparing to a set of random natural peptides and was suggested by the original study.

3.2.5 Collect of peptides immunogenicity features

We collected potential features that had been previously reported or hypothesized to have an impact on T cell recognition. A total of 14 features were included, which were categorized into the following 3 groups: (A) Features characterizing the physical properties of the neo-peptides, including the hydrophobicity score (sum of pre-defined hydrophobic score of each amino acid) [84], the peptide molecular size (sum of individual residues weights after removing H_2O weight) [85] , the peptide entropy (sum of pre-defined entropy value of each amino acid) [86], the polarity [87] and the charge value of the peptides (sum the pre-defined amino acid polarity and charge scores at residues 2, 3, 5, and 6) [88] [89] ; (B) Five features associated with neo-peptide processing and MHC-I binding, including NetMHCpan predicted IC50 value in nM' and 'percentile rank'-based predictive values for MHC-I binding affinity, and another three NetCTLpan [90] output scores for protein cleavage, TAP transport efficiency, and the combined score; and (C) Four previously reported neoantigen immunogenicity scores, including Immune Epitope Database (IEDB)

Class I Immunogenicity for T-cell recognition [91] , differential agretopicity index (DAI) of the neo-peptide [29], probability of the neoantigen to be recognized by the TCR repertoire (fitness score) [36] , and neoantigen recognition potential calculated as a DAI time fitness score.

3.2.6 Training and testing dataset for predicting the quality of the neoantigen

To construct the positive set of immunogenic epitopes, we collected immunogenic epitopes from infectious virus-derived epitopes from the IEDB [91] and mutation-derived peptides with verified immunogenicity in human tumor studies [92] . We searched peptides that matched the infectious virus epitopes with less than five mismatching from human reference proteins fragments pool using the pepmatch_db_x86_64 program from MuPeXI [66] . In summary, 964 infectious epitopes and 231 mutation-derived peptides with verified immunogenicity in human tumor studies were included in the positive immunogenic set (accompanied with matched wild-type peptides). The corresponding HLA alleles were also retrieved from original reports.

The negative set was composed by collecting non-immunogenic epitopes using mutated peptides from common nonsynonymous SNVs from the 1000 Genomes Project [93] with minor allele frequency $\geq$ 0.05 in all populations (ANNOVAR [94], hg38_ALL.sites.2015_08 version, refGene database annotated), with the assumption that common peptide variants would not lead to an immunogenic response. The 9-mer peptides that harboring SNV in the 5^{th} amino acid were generated. The wild-type peptides of the reference allele were collected as well. The HLA alleles for non-immunogenic peptides were randomly assigned using the top HLA alleles in the 1000 Genomes Project with a frequency over 5% across all populations.

3.2.7 RNA-seq dataset

Immunotherapy studies. RNA-seq datasets of pre-treated tissues from the melanoma patients with well-noted immunotherapy response metadata were included in this study. The Riaz cohort (GSE91061) includes the RNA-seq data of 51 samples from advanced melanoma patients before they were treated with PD-1 inhibitor Ipilimumab [95] , and the Hugo cohort (GSE78220) includes RNA-seq data from 27 melanoma patients who received PD-1 inhibitor Pembrolizumab [96] . Raw data were downloaded and converted to fastq format using sra-tools (v 2.10.1) [97] .

TCGA and GTEx.We downloaded the raw data of the TCGA-SKCM and TCGA-GBM cohorts from the GDC data portal (https://portal.gdc.cancer.gov/) [98] . The RNA-seq data of the normal skin (n=184) and brain (n=420) tissues were downloaded from the GTEx consortium [99]. All of the TCGA and GTEx bam files were converted to fastq files with samtools (v 1.10) [100], facilitating the uniform IR-neoAg quantification.

The fastq were aligned to the reference human genome (GENCODE annotation.v32) using STAR (v.2.7.2b) [80] . The aligned bam files were used to identify the IR events and the HLA alleles. The transcript expression abundance of all RNA-seq data was quantified with Salmon (v1.2.1) [101] and further summed to reflect the gene-level expression using the 'tximport' package in R.

3.2.8 RNA-seq data processing

The quality of raw RNA-seq data was first evaluated using FastQC (version 0.11.5). FastQC results showed that all raw reads were qualified for downstream analysis. RNA-seq reads for two melanoma datasets (GSE78820 and GSE91061) were aligned to the human reference genome (GRCh38) using STAR software (v2.7.2b) with the following

parameters: *STAR --runThreadN 16 --genomeDir $INDEX --readFilesIn $FASTQ1 $FASTQ2 --readFilesCommand zcat --outSAMmapqUnique 60 --outSAMtype BAM SortedByCoordinate*. For the mono-allelic RNA-seq dataset GSE131267, with reads length < 50 bp (~38 bp), bowtie2 was applied to align the short reads to the human reference genome (GRCh38).

3.2.9 MS data analysis

Raw MS files were analyzed using SearchGUI (V.3.3.20) [102] and PeptideShaker (V.1.16.45) [103]. MS/MS spectra were searched using X!Tandem, MS-GF+, and OMSSA search engines against the target-decoy neoantigen database [104] containing forward and reverse sequences of predicted neoantigens. The following analysis settings were used: (i) Trypsin (Specific), (ii) precursor mass tolerance 10 ppm, (iii) fragment mass tolerance 0.5 Da, (iv) one fixed modification (carbamidomethylation of cysteine), and two variable modifications (oxidation of methionine, acetylation of protein N-term) (v) peptide length allowed (8–30 amino acids) (vi) Protein FDR 1%, Peptide FDR 1%, PSM FDR 1%, and (vii) a maximum of two miscleavages allowed. Peptides identified to be shared between two proteins were combined and reported as one protein group.

3.2.10 Estimation of immune cell infiltration with RNA-seq data

CIBERSORT analysis was conducted with gene expression profiles from RNA-seq, using the default LM22 dataset as the signature gene file (1000 permutations) and other default parameters [105].

3.2.11 Data availability

There are no new data associated with this article. The IntronNeoantigen program is available at https://github.com/cpdong/IntronNeoantigen.

3.3 Results

3.3.1 Re-annotation of intron coordinate

We first re-annotated the exon and intron sets for protein-coding genes using a gene annotation file from GENCODE (GRCh38, v32). Unique exon sets for each protein-coding gene were determined by merging intersected exons. The introns were defined as the interval regions of exon sets (Fig. 4A). We identified 189,458 unique from 17,689 unique human protein-coding genes. These introns were further classified into the following four types according to their potential to be translated into peptides (more details shown in Fig. 4B and Methods): (1) large introns > 10 kilobases (n=20,072; 10.6%); (2) no ORF due to a stop codon at the first position (n=35,282; 18.6%); (3) continuous ORF to the downstream exon (n=7,850; 4.1%); and (4) partial ORF with a stop codon anywhere in the intron (n=126,254; 66.7%) (Fig. 5A). Long introns and no-ORF introns were excluded from the subsequent analysis because of their relatively low abundance or inability to generate novel peptides. Of the introns with partial or complete ORFs, 55.7% (n=70,307) produced peptides < 20 amino acids in length, 87% (n=109,951) produced peptides < 50 amino acids in length, and 97.6% (n=123,226) produced peptides < 100 amino acids in length (Fig. 5B).

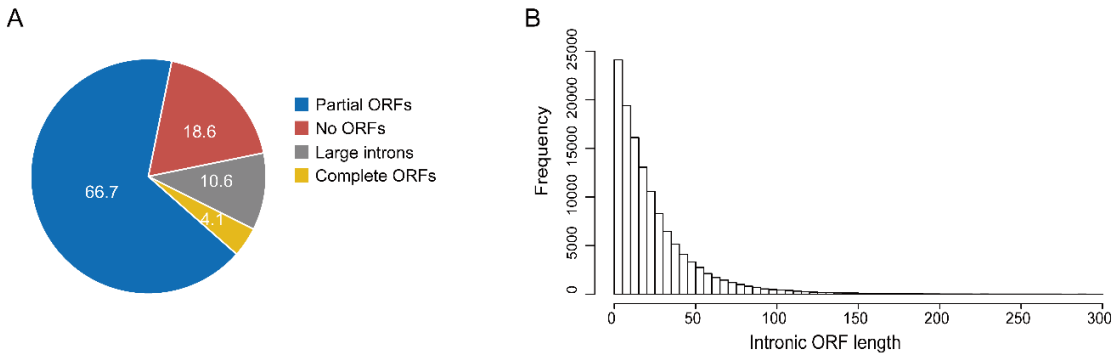

Figure 5 Distribution of re-annotated introns. **A** Pie chart showing the frequency of four types of retained introns. **B** Length distribution of peptides translated from retained intron sequences with complete or partial ORFs.

3.3.2 Workflow of IntronNeoantigen pipeline

We developed an all-in-one workflow for quantifying the IR-neoAg, as shown in Fig. 6. The only required input is RNA-seq data in a BAM file aligned using a splice-aware aligner, such as STAR or TopHat. We first assessed the HLA class I genotypes from the RNA-seq data using arcasHLA [64] . Users can also input their own HLA alleles following a pre-defined format requirement. Overall, the workflow consists of the following steps: (1) Identification of a disease-specific IR event from RNA-seq data, (2) Quantification of the binding affinity of IR-derived neo-peptides with corresponding HLA alleles, and (3) Evaluation of neoantigen quality by assessing their ability to be recognized by T cells using a random forest classifier (positive or negative immunogenicity). IntronNeoantigen will report all results, including predicted binding affinity, neoantigen recognition potential, and other conventional predictions, peptide characteristic metrics, and information listed in the methods.

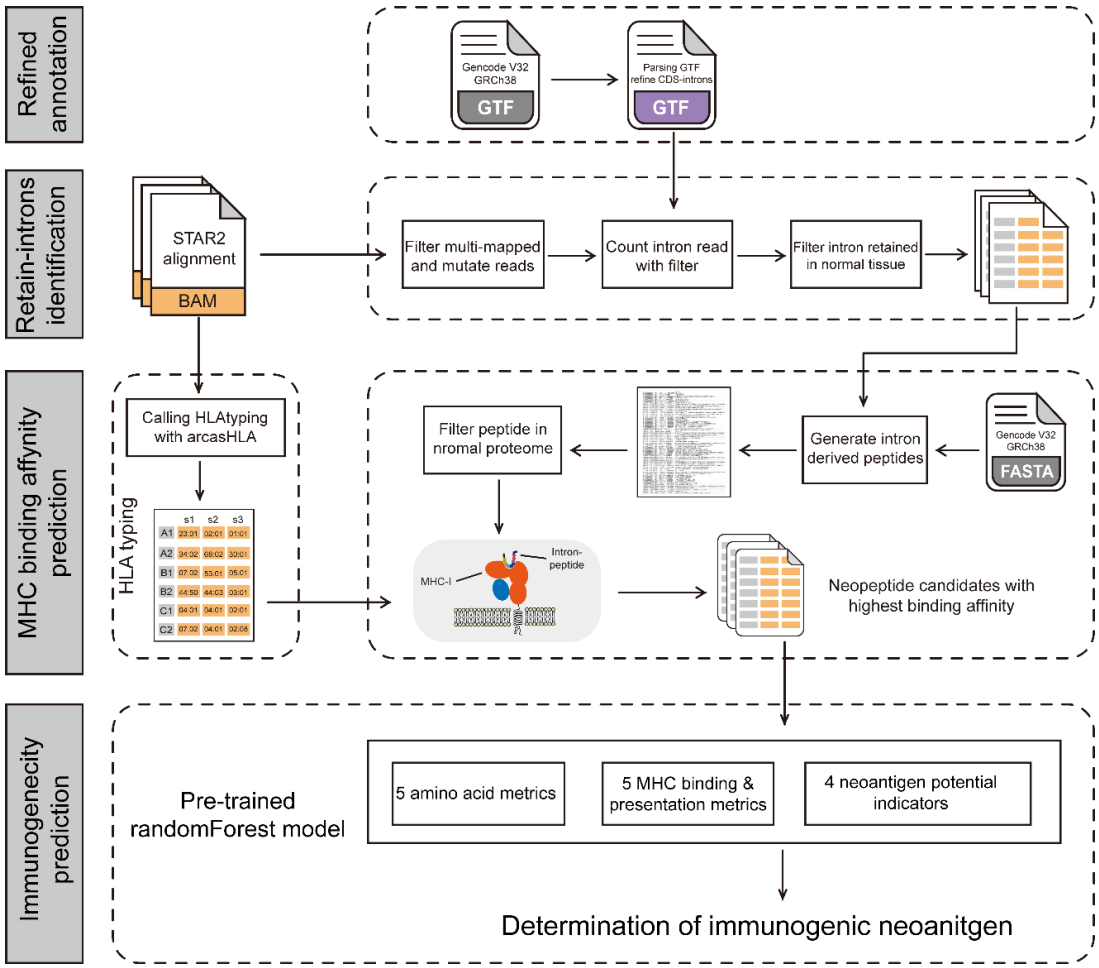

Figure 6 Workflow of IntronNeoantigen. Starting with an aligned RNA-seq BAM file, IntronNeoantigen consists of the following three steps: **(i)** count intron reads and filter qualified events, **(ii)** predict intron neoantigens with netMHCpan tools, and **(iii)** predict immunogenicity with a random forest model. The HLA typing was measured with arcasHLA.

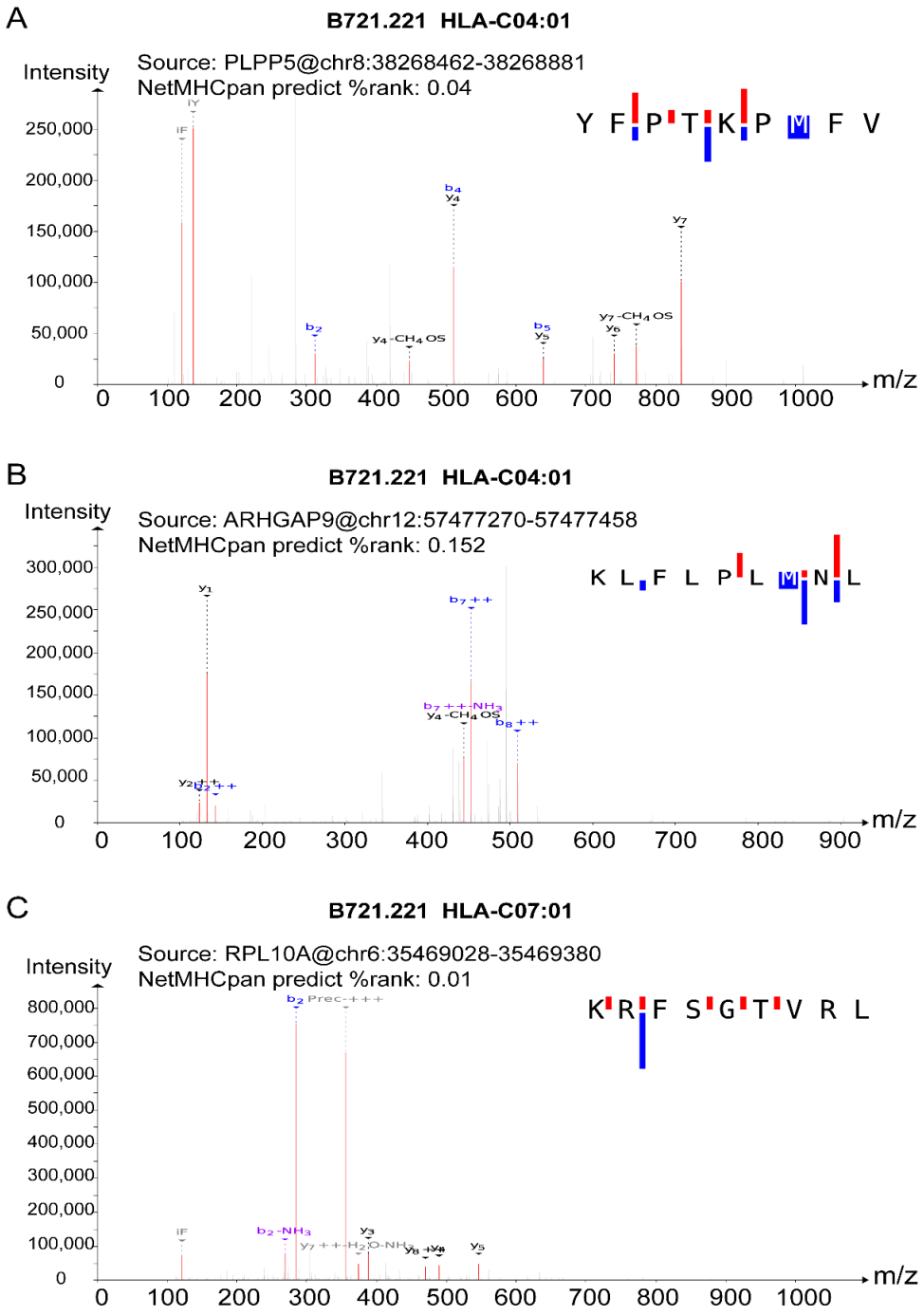

Figure 7 Potential intron-retention-derived epitopes confirmed in HLA-Associated Peptidomes data. Peptides verified in HLA-peptidomes in HLA-C04:01 mono-allele B721.221 cells (**A** and **B**), and HLA-C07:01 mono-allele B721.221 cells (**C**).

We validated our algorithm-predicted IR neoepitope identified from HLA-proteomic data. We re-analyzed the RNA-seq and HLA-proteomics data from mono-allelic B721.221 cells transfected with individual C0401 or C0701 mono-alleles in Sarkizova et al.'s study [106]. Interestingly, two out of 107 IR-neoAg candidates were found in the matched mono-allele LS-MS proteomics data (Fig. 7).

3.3.3 Evaluation of the neo-peptide immunogenicity predictive model

We collected verified immunogenic peptides and non-immunogenic peptides to train the random forest classifier and perform internal validation. The immunogenic peptides consisted of 964 infectious virus-derived epitopes and 231 high confidence mutation-derived peptides from human tumors (data not shown). The corresponding HLA alleles to the epitopes were obtained from original studies. A total of 21,258 nonsynonymous SNPs, with minor allele frequencies larger than 0.05 from the 1000 Genomes Project, were used to generate the non-immunogenic peptides set (Detailed described in the Methods, supplementary table not shown). We trained the random forest classifier using 75% of the above dataset with optimal parameters. The importance of the input features for determining immunogenicity is shown in Fig. 8A.

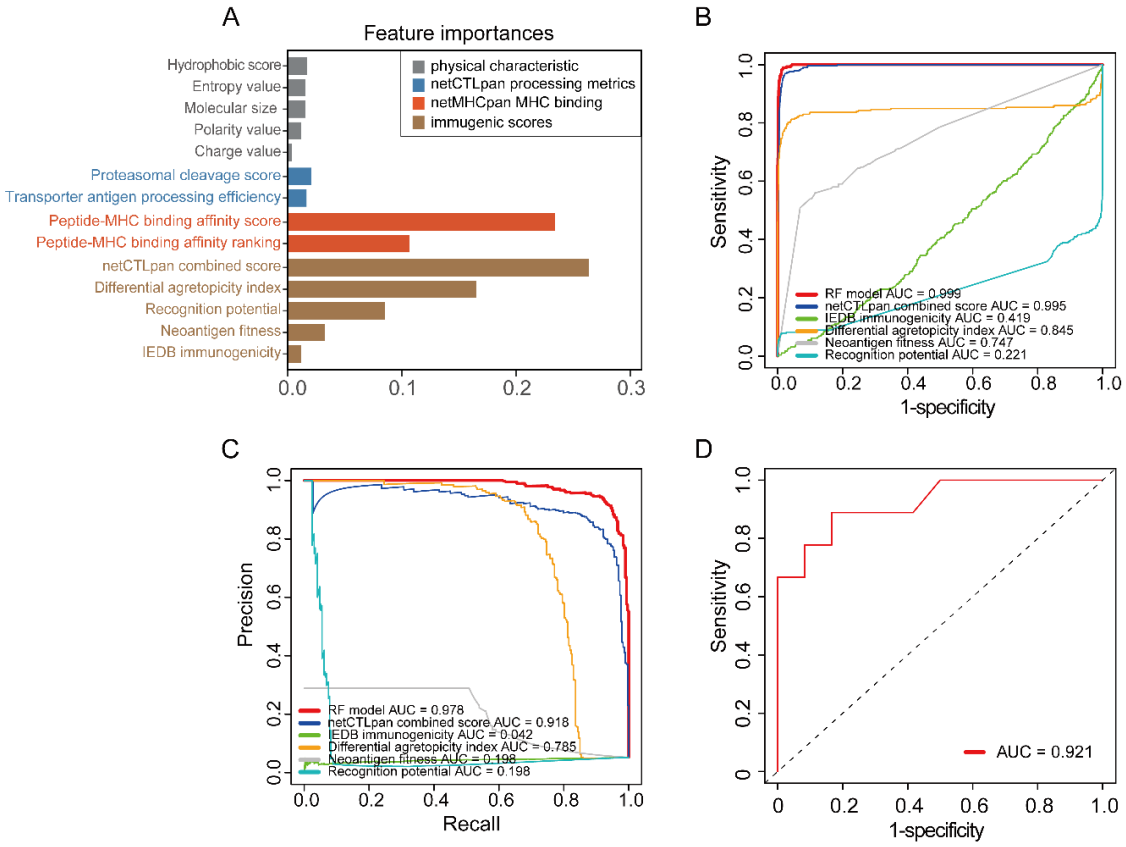

Figure 8 Performance evaluation of the random forest classifier. **A** The importance of features used in the random forest training models used in this study. **B** ROC curves of the random forest and the top four indicators in the internal validation dataset. The random forest models outperformed the conventional single indicators. **C** Precision-recall curves of the random forest model and the top four indicators. The random forest models showed better performance than conventional features. **D** ROC curve of the random forest model in the external validation dataset.

We performed internal validation with the remaining 25%. Receiver operating characteristic (ROC) curve was used to evaluate the sensitivity and specificity of the predictive performance of the model. The random forest classifier showed the highest area under the curve (AUC) value of 0.9986, outperforming netMHCpan-predicted metrics and other complicated immunogenic indicators, such as the differential agretopicity index (DAI) [29] and neoantigen fitness scores [36] (Fig. 8B). A precision-recall curve was used to avoid false positives resulting from training-set imbalance (immunogenic/non-immunogenic peptides ratio 1:17.8). The area under the precision-recall curve (AUPRC) of the random forest classifier (Fig. 8C) outperformed other indicators, reaching 0.978. Furthermore, we applied the machine-learning classifier to an external testing neo-peptides dataset from Carreno et al. study [PMID:25837513]. Nine out of 21 neo-peptides were confirmed to elicit an immune response. The AUC of the external validation set reached 0.921 (Fig. 8D). Overall, the results demonstrate that the machine-learning-based classifiers have a higher capacity to predict neoantigen immunogenicity.

3.3.4 Application in an immunotherapy-treated datasets

As previous studies have demonstrated that neoantigens are closely linked with patient response to immunotherapy, we sought to investigate the potential of using IR-neoAgs to predict patient responses to ICB. We analyzed the RNA-seq data from two immunotherapy studies of pre-treatment melanoma patients with well-noted clinical responses. We quantified the melanoma-specific IR events by filtering out normal IR events occurring in over 25% of the normal skin samples in the Genotype-Tissue Expression Project (GTEx) (supplementary table not shown). We identified 412 IR-neoAgs per sample (Fig. 9A) in 27 patients before administration of Pembrolizumab in the

Hugo cohort [96]. In the Riaz cohort [95], 49 melanoma patients were analyzed prior to administration of Nivolumab, and we identified an average of 525.3 IR-neoAgs per sample (Fig. 9D). The IR-neoAg load showed no significant correlation with immunotherapy response in the two cohorts (Fig. 9B and 10E). However, we found that our predicted immunogenic peptides' average neoantigen fitness score was significantly associated with patient response to anti-PD1 treatment. The Wilcoxon rank-sum test p-value was 0.018 in the Hugo cohort (Fig. 9C) and 0.046 in the Riaz cohort (Fig. 9F). These data suggest that IR-neoAgs can potentially be used for predicting immunotherapy response.

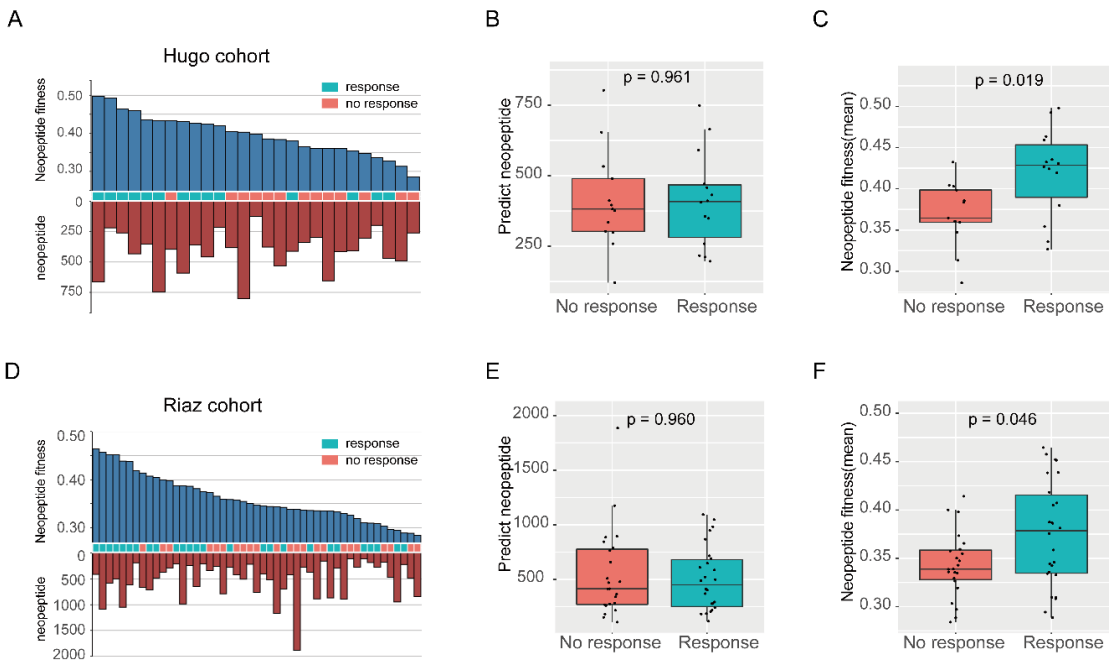

Figure 9 IntronNeoantigen in immunotherapy patient cohorts. **A** and **D**, IntronNeoantigen burden and neoantigen immunogenicity score distribution in two immunotherapy-treated patient cohorts; **B** and **E**, the intron neoantigen burden was not associated with response to ICB; **C** and **F**, the overall neoantigen immunogenicity score showed more differences than the neoantigen burden between the long-benefit group and the no-benefit group.

3.3.5 Application to The Cancer Genome Atlas dataset

Several studies have revealed that mutation-derived neoantigen load is directly associated with the survival rate of many cancers. Thus, we investigated whether IR-neoAgs may also function as favorable biomarkers for cancer survival rates. We identified IR-neoAgs in pathological stage III/IV melanoma patients included in the TCGA-SKCM cohort using the IntronNeoanitgen pipeline. In a total of 193 patients, 73,660 IR-neoAgs were identified, with a median of 304 neoantigens per patient. We found that the IR-neoAg load (Fig. 10A) was not significantly associated with overall survival rates (p=0.47). As infiltrating CD8+ T cells play an indispensable role in triggering a neoantigen-mediated immune response in tumor tissue [107], we estimated the tumor-infiltrating immune cell proportions using Cibersort [105]. We found that higher CD8+ T cell infiltration was associated with a favorable survival outcome in patients with stage III/IV melanoma (p=0.02) (Fig. 10B). Notably, when we combined IR-neoAg load and CD8+ T infiltration data, patients with a high IR-neoAg and CD8+ T cell infiltration level exhibited better prognoses than those with a low IR-neoAg and a low CD8 T cell infiltration level (p=0.02) (Fig. 10C). The 5-year survival rate was 0.83 (95% Cl 0.36 - 0.63) in the group with a high IR-neoAg and CD8+ T cell infiltration level. This rate dramatically decreased to 0.43 (95% Cl 0.67 - 1.00) in the group with a low IR-neoAg and CD8+ T cell infiltration level.

Isocitrate dehydrogenase wild-type GBM (IDH wild-type GBM) is the most aggressive brain tumor subtype, with a median survival time below 15 months [108]. Similarly, we quantified IR-neoAgs in 132 IDH wild-type GBM samples from the TCGA-GBM cohort. Five thousand four hundred forty-eight normally occurring IR events in more than 25% of GTEx normal brain samples were discarded for the analysis. We identified a

total of 88,683 IR-neoAgs, with a median value of 432 per sample. The IR-neoAg load was significantly associated with the OS of IDH wild-type GBM patients (p=0.038) (Fig. 10D). While the CD8+ T cell infiltration level was not associated with survival outcomes (p=0.664) (Fig. 10E), the combination of IR-neoAg load and CD8+ T cell infiltration level could help identify IDH wild-type GBM patients with a favorable prognosis (p=0.019) (Fig. 10F). Patients with high IR-neoAg load and CD8+ T cell had a significant higher 2-year OS rate compared to patients with low IR-neoAg and CD8+ T cell (0.40 vs. 0.08, respectively).

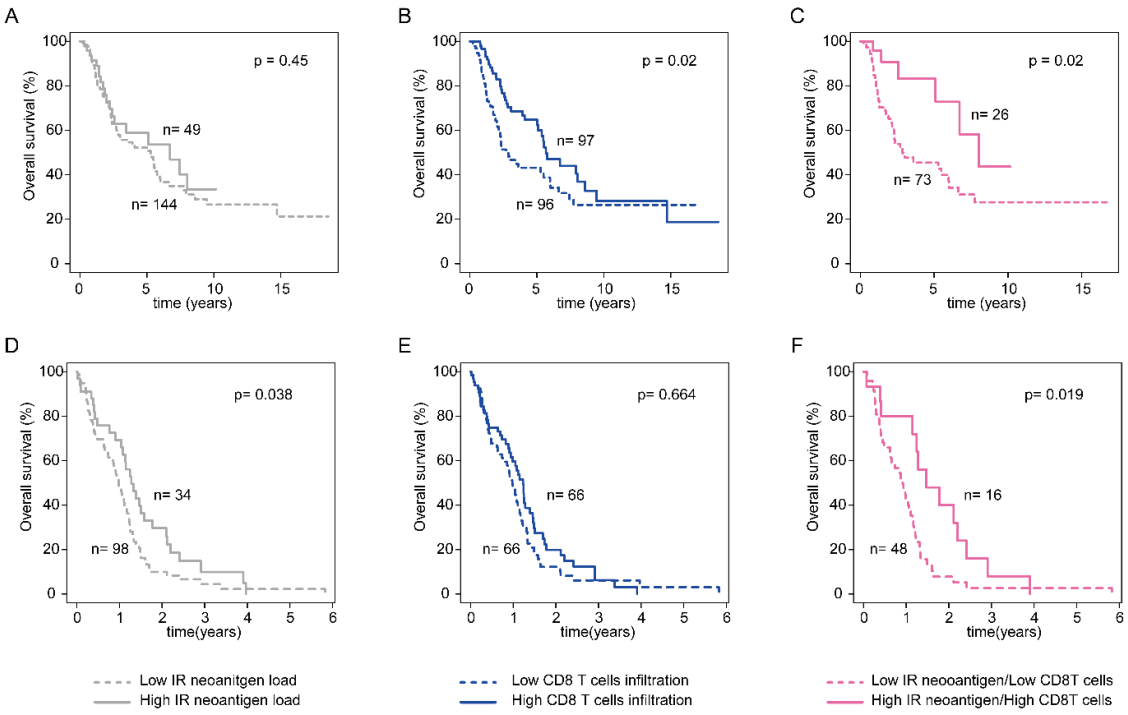

Figure 10 The combination of IR-neoAg load and CD8+ T lymphocyte infiltration level identified advanced cancer patients with longer survival times. Kaplan–Meier survival curves were used to compare advanced melanoma (stage III/IV) and IDH wild type GBM patients: (**A** and **D**) Patients with high (defined as top quantile) IR-neoAg load vs. those with low (below the top quantile) IR-neoAg load; (**B** and **E**) patients with high CD8+ T cell infiltration levels vs. those with low CD8+ T cell infiltration levels (using the median as the cutoff) and (**C** and **F**) patients with IR-neoAg^{high}/ CD8T^{high} versus patients with IR-neoAg^{low}/ CD8T^{low}.

3.4 Discussion

Recently, it has been reported that neoantigens are closely linked with favorable prognoses and responses to ICB therapy in multiple cancers [36, 74, 75]. Transcripts carrying un-spliced introns have a higher potential to generate neo-peptides, serving as a source of novel tumor neoantigens [4]. To date, the majority of *in silico* approaches to neoantigen prediction have been focused on identifying neoantigens resulting from DNA-level alterations, such as pVAC-Seq [65], MuPeXI [66], Neopepsee [67], pTuneos [68], and ScanNeo [69]. Most attempts at predicting splicing-derived neoantigen have been directed towards identifying neoantigens arising from aberrant transcripts and have been limited due to the lack of transcript assemble accuracy.

Here, we present IntronNeoantigen, an integrated computational pipeline for identifying personalized IR-neoAg information from RNA-seq data. We further established a random forest classifier to evaluate the immunogenicity of putative neo-peptides. Moreover, we verified the function of IntronNeoantigen using publicly available cancer datasets and demonstrated its utility for predicting clinical outcomes.

Interestingly, we found that the average IR-neoAg fitness score of our predicted candidate neoantigens was significantly associated with patient anti-PD1 treatment responses in both cohorts. However, IR-neoAg load was not directly associated with immunotherapy outcomes in two ICB datasets. We propose a more comprehensive method based on mutation load, IR-neoAg load, checkpoint gene expression level, and other factors to improve the accuracy of predicting patient immunotherapy responses.

Traditional treatments, such as surgery and chemotherapy, are unsuitable for many advanced melanoma and glioma patients. Immunotherapy offers new hope and a potential

therapeutic option for such patients [109]. Several tumors mutation-derived neoantigens have been linked to improved patient survival rates and responses to ICB in multiple cancers [36, 110]. We applied our pipeline to advanced melanoma and GBM cohorts. Our results showed the combination of CD8+ T lymphocyte infiltration level, and high quantity IR-neoAg load was the best predictive model for predicting the survival outcomes of advanced-stage melanoma and IDH wild-type GBM patients. Our findings suggest that the combination of high IR-neoAg load and elevated CD8+ T cell infiltration level may be useful for identifying patients with a high likelihood of achieving a clinical response to immunotherapy. It is worth mentioning that, for some cancer types that harbor few mutations, IR-neoAg analysis could allow for the prediction of immunotherapy response, and the data could be used to design precision cancer vaccines against those cancers. The successful identification of IR-neoAgs may also reveal further mechanisms underlying splicing-mediated immune responses in other diseases, such as type I diabetes and neurodegenerative disorders.

Our study presents the first comprehensive pipeline for identifying immunogenic IR-neoAgs from RNA-seq data, providing a useful tool for screening splicing-induced neoantigens for research and clinical applications. Currently, the evaluations of our model are limited to publicly available datasets with both RNA-seq and mass spectroscopy data of eluted epitopes available. Further experimental validations are required to verify the role of IR-neoAgs in promoting immune responses.

3.5 Conclusion

We presented IntronNeoantigen, a computational pipeline for identifying IR-neoAgs from RNA-seq data. IntronNeoantigen enables the efficient identification and

prioritization of personalized IR-neoAgs, which might aid neoantigen-based vaccine development and the prediction of patient immunotherapy responses. We believe this tool will help researchers explore the potential roles of IR-neoAgs in other immune-dysregulated diseases.

Chapter 4 Intron Retention-induced Neoantigen Load Correlates with Unfavorable Prognosis in Multiple Myeloma

4.1 Introduction

MM is characterized by the clonal expansion of malignant plasma cells in the bone marrow [111]. Recent therapeutic advances have extended overall survival, but most MM patients ultimately relapse [112]. ICB therapy has revolutionized the treatment of many solid tumors by harnessing the immune system for effective anti-cancer treatment [113]. In these diseases, clinical response to ICB therapy is associated with the presence of tumor-specific antigenic peptides, or neoantigens [114], a source of potential neoepitopes that can be loaded onto MHC class I molecules to generate an antitumor immune response [115]. Cytotoxic T-cells recognize tumor neoantigens as foreign and kill the presenting tumor cells, which initiates an antitumor immunological memory that hinders tumor recurrence. An important source of cancer neoantigens is somatic DNA mutations in the genome's coding regions [116] and the mutation-derived neoantigen load in several types of solid tumors corresponds with better prognosis [96, 117-119]. However, MM has a relatively low mutation frequency. In contrast to solid tumors, mutation-derived neoantigen load in MM has been associated with unfavorable outcome [120, 121].

Another potential source of tumor neoantigens is aberrant RNA splicing [18, 79, 122, 123]. AS is a regulatory mechanism that generates multiple mRNA transcripts from a single gene and significantly expands proteome diversity [124]. Consequently, disruption of splicing mechanisms has a large impact on the transcriptome and is a significant driver of disease [125]. IR occurs when the spliceosome fails to remove specific introns from pre-mRNA molecules, and they remain in the mature polyadenylated mRNA. In normal cells,

IR functions to further reduce the levels of relatively low abundance transcripts that are not needed in specific cell types, such as the expression of developmentally regulated genes [12, 126]. This type of regulation has been termed transcriptome-tuning and is brought about through both nuclear RNA degradation and nonsense-mediated mRNA decay [127].

IR occurs more frequently in nearly all cancer types compared with normal control tissues, even in the absence of DNA mutations in genes encoding proteins involved in splicing. Additionally, in cancer cells, transcripts with IR are present at relatively high levels in cytoplasmic mRNA [128]. These transcripts are translated and degraded by the nonsense-mediated decay (NMD) pathway, a translation-coupled mechanism that eliminates mRNAs containing premature translation-termination codons [129]. Although most IR transcripts are subject to NMD-induced degradation, this process does not occur until after the pioneer round of translation, which can result in the production of neo-peptides that bind to MHC molecules [4, 130]. Therefore, we hypothesized that IR-neoAgs in MM might impact immune response.

Herein, we used RNA-seq data from the Multiple Myeloma Research Foundation's CoMMpass Study (MMRF) to identify IR events and predict IR-neoAgs. We found cells in bone marrow aspirates from MM patients exhibited high levels of IR events. However, consistent with the findings of that high mutation-neoantigen load predicts unfavorable prognosis, high IR-neoAg load was correlated with shorter OS in MM. To investigate why high IR-neoAg load was not correlated with better MM patient survival, we performed gene set enrichment analysis on MM samples with high versus low IR-neoAg load. This analysis revealed that high IR-neoAg load was positively associated with the expression of T-cell inhibitory molecules, such as those involved in interferon (IFN) and tumor necrosis

factor (TNF) alpha signaling activity. In addition, flow-cytometric analyses of four MM cell lines showed an inverse correlation between IR levels and MHC-II abundance, while treatment with a splicing inhibitor increased MHC-I protein abundance, especially in MM cells bearing high IR levels.

4.2 Materials and Methods

4.2.1 RNA-seq data sets

The raw data from MMRF was obtained through an authorized data access request for dbGaP study accession: phs000748.v7. p4. RNA-seq data from 893 samples, including both newly diagnosed and relapsed subjects, were downloaded and converted to fastq format using SRA-tools (v2.10.0). Curated survival and clinical data were downloaded from the UCSC Xena cancer browser (http://xena.ucsc.edu). The revised International Staging System (R-ISS) was calculated as defined by the International Myeloma Working Group [131], by considering the presence of del(17p), t(4;14), and t(14;16) and information on serum β2-microglobulin, albumin, and lactate dehydrogenase levels. B2M mutations and the status of TP53 in baseline samples were obtained from Dr. Brian Walker, as described previously [132].

Two other RNA-seq studies with normal plasma cells were retrieved from the Gene Expression Omnibus (GEO). Data of bone marrow-derived plasma cells from five healthy individuals and five newly diagnosed MM patients were obtained from GSE110486 [133]. Data of plasma cells from bone marrow or tonsil of another eight normal subjects were acquired from GSE114816 [134].

For all RNA-seq data, an initial sequence-level quality assessment was performed using FastQC (v0.11.5). The alignment-free quantification tool Salmon (v1.2.1) [101] was

used to quantify the expression of gene transcripts from RNA-seq data using the reference transcriptome built from Gencode (GRCh38, v32) gtf annotation as the index. The gene-level transcript abundance was calculated using the tximport package in R.

The normalized gene expression data of 887 cancer cell lines (dated 2018.09.29) and their annotations (dated 2018.12.26) were downloaded from the Cancer Cell Line Encyclopedia (CCLE) data portal [135].

4.2.2 Identification of intron retention events

To quantify the IR events for MM samples, RNA-seq reads were aligned to the GRCh38 reference genome using STAR (v2.7.2) [80]. Uniquely mapped RNA-seq reads were used to quantify the expression levels of retained introns using HTseq [82] package. Additional criteria were applied to filter the identified IR events: (1) read counts for both the intron region and its flanking exon regions were > 10; (2) read coverage of the intron was comparable to its flanking exons, such that the transcripts per million (TPM) ratios of introns to flanking exons was > 0.05 and < 0.5. MM-specific IR events were further selected by removing the events that were observed in normal plasma cells using the same filtering criteria.

4.2.3 IR-neoAg prediction

We used arcasHLA (v1.1) to infer HLA class-I genotypes from the RNA-seq data [64]. Sequences from the retained introns were translated into peptides by extending the open reading frame of the upstream exon using the standard codon table. The translated peptides were segmented into 8 to 11 amino acid lengths that contained at least one intron-encoded amino acid. For each patient, NetMHCpan4.1 was used to estimate the binding affinity of the IR-derived neo-peptides with the patient's HLA alleles [83]. A binding

affinity rank score less than 2 (default parameter of NetMHCpan4.1) was regarded as a neoantigen candidate. Expression levels for each neoantigen were determined by the abundance of IR events that generated the specific neoantigen.

4.2.4 Differential expression and pathway enrichment analysis

Differentially expressed genes in plasma cells between MM and healthy bone marrow samples were identified using the limma [136] package in R. A total of 1790 gene sets from Molecular Signatures Database (MsigDB, v7.2), including KEGG, REACTOME, and HALLMARK gene sets, were used for enrichment analysis. Fisher's exact test was used to test for pathway enrichment significance, and the p-value was adjusted for multiple hypothesis correction using the Bonferroni method [137]. ClusterProfiler was used to visualize the pathway enrichment results [138]. Single-sample gene set enrichment analysis (ssGSEA) was used to assess the pathway activity in each individual using GSVA [139] package in R, using default parameters.

4.2.5 Cell culture of MM cells and spliceosome inhibition

KMS11, U266, JJN3, AMO1 MM cell lines were kindly provided by Dr. David Roodman and cultured at 37°C in a humidified atmosphere containing 5% CO2 and maintained in RPMI media supplemented with 10% fetal bovine serum (FBS). Cells were tested for mycoplasma infection monthly as a regular lab routine. Pladienolide-B (Cayman Chemical Company, Ann Arbor, MI; cat# 16538) was dissolved in dimethyl sulfoxide and used at the following concentrations: 0, 0.1, 1, 5, 10, and 100 nM. Flow-cytometric analyses were performed at 96 hours post-treatment, two independent biological replicates were analyzed for each treatment.

4.2.6 Antibodies and flow cytometry analysis

The following flow-cytometry antibodies were used: HLA-DR, DQ, DP-APC (Biolegend, San Diego, CA cat# 361714), Isotype control-APC (Biolegend cat# 400222), MHC class I-PE (LSBio, Seattle, WA cat# LS-C751033-0.1). Isotype control-PE (Abcam, Waltham, MA cat# ab91357). Flow cytometric data were acquired using the LSR II flow cytometer (BD Biosciences, San Jose, CA) and analyzed with FlowJo software.

4.2.7 Statistical considerations

Survival analysis and Cox-proportional hazard comparison were performed using the R package Survival with log-rank test and hazard ratio (HR) stat istical tests [140]. Significant differences in the value of the two given groups were assessed using the Mann-Whitney-Wilcoxon test [141]. Statistical analyses were performed in R (v4.0.2).

4.2.8 Code Availability

The computational algorithm and source code allowing for reproduction of the IR-neoAgs quantification in this manuscript are available at https://github.com/cpdong/IntronNeoantigen.

4.3 Results

4.3.1 Genes involved in spliceosome activities are differentially expressed between MM and normal plasma cells

To investigate whether the expression of genes involved in RNA splicing was altered in MM compared to normal plasma cells, we analyzed differential expressed genes using RNA-seq data of plasma cells from 5 newly diagnosed MM patients (NDMM) and 5 healthy controls (GSE110486). These results showed that the spliceosome pathway was among the top upregulated pathways in MM (Fig. 12A), where 67 out of 126 genes in the

spliceosome pathway were upregulated significantly in MM samples. Gene set enrichment analysis also demonstrated that the spliceosome pathways were enriched in MM samples with a normalized enrichment score of 1.46 (p-value < 0.001, FDR = 0.016, Fig. 12B). We also found that the 230 upregulated differentially expressed genes identified in GSE110486 were also highly enriched in the NDMM samples from the MMRF cohort, as compared to the normal plasma cells (Fig. 11A). In addition, the expression levels of 11 out of 12 SRSF protein genes, a conserved family of proteins involved in RNA splicing, were upregulated in MM (Fig. 11B). Additional analysis of the MMRF data suggested that the increased expression of each of these 12 SRSF family genes was associated with decreased overall survival time (Fig. 12C).

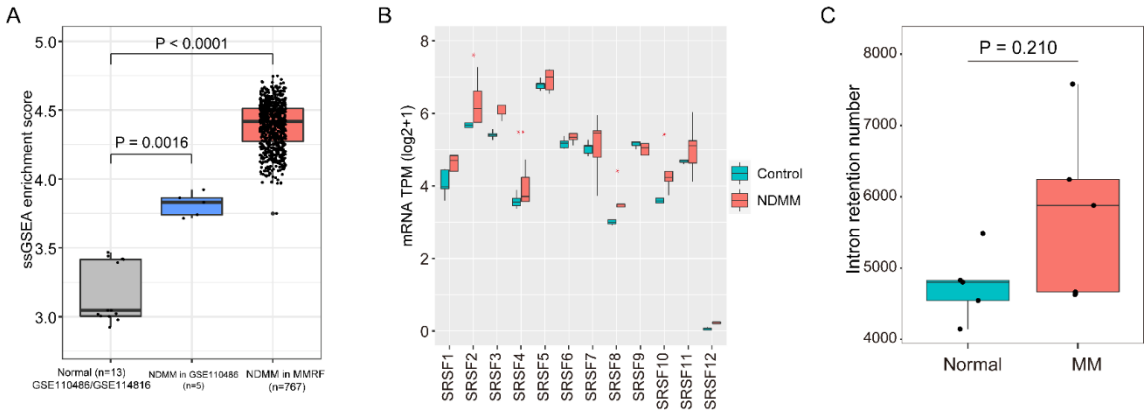

Figure 11 Plasma cells from NDMM patients show higher levels of differential gene expression, upregulation of splicing factor genes and more IR events compared to normal plasma cells. A ssGSEA of the 230 upregulated genes in the MMRF NDMM patients. The enrichment scores for the NDMM samples in the MMRF data were significantly higher than the 13 normal plasma samples (5 in GSE110486 and 8 in GSE114816) and were also higher than the 5 NDMM samples in GSE110486. B Comparison of TPM of SRSF protein genes from NDMM (GSE110486) and normal plasma (control) RNAseq data. Asterisks denote significant differences in expression (*, p-value < 0.05 and **, p-value <= 0.01). C Comparison of the number of IR events in normal plasma cells and plasma cells from NDMM patients from GSE110486 (n = 5 per group). P values were determined using the Mann-Whitney test. Box plots show the median and 25th and 75th percentiles (box) and the 95% confidence interval (whiskers).

4.3.2 IR events are more common in MM compared to control plasma cells

Accumulating studies provide strong evidence that IR is an important source of tumor neoantigens [18]. We sought to characterize IR in MM and its association with MM progression. The number of IR events and their expression levels were assessed for both the MM and control samples in GSE110486. We observed an average of 5799 IR events in 5 MM samples and 4761 IR events in healthy controls (Fig. 11C); however, due to the limited sample size, this difference did not reach statistical significance. Next, we compared the number of IR events in 767 NDMM samples from the MMRF cohort with 13 control plasma cell samples from bone marrow and tonsil of healthy subjects (5 from GSE110486 and 8 from GSE114816). The NDMM samples showed more IR events with an average of 5391 per sample compared to 4065 IR events in the normal plasma cells. The result of this comparison was statistically significant (Wilcoxon test, p-value = 0.001) (Fig. 12D).

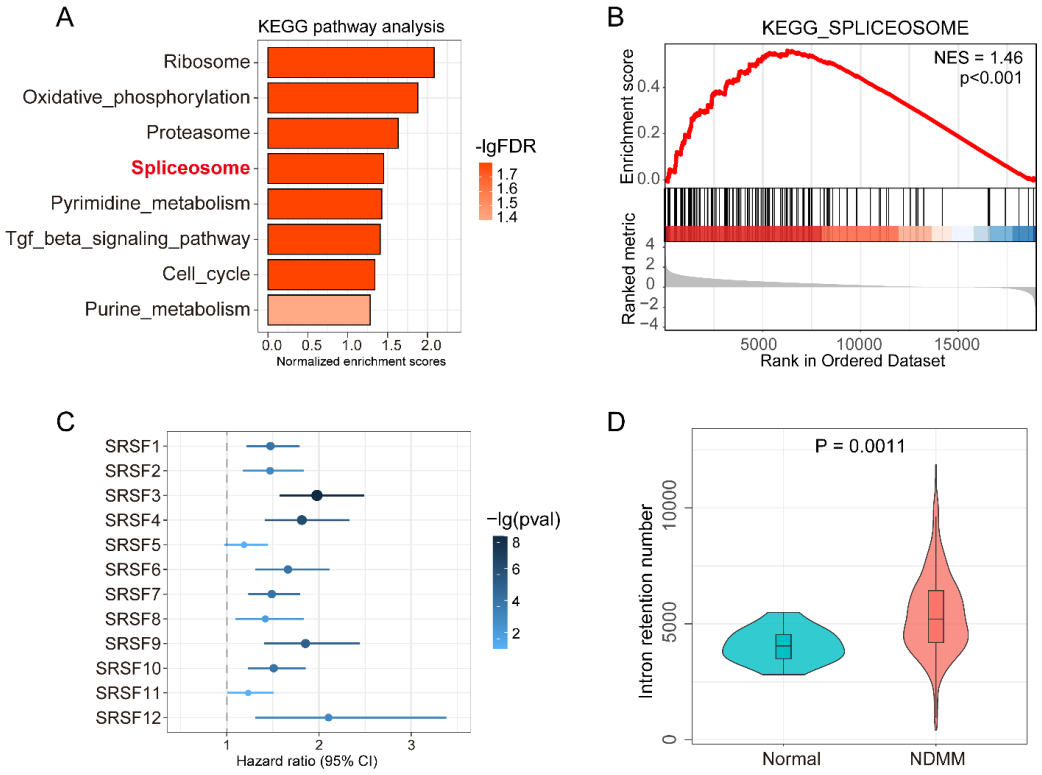

Figure 12 IR events in plasma cells from MM patients are associated with altered RNA splicing. **A** Spliceosome is among the top significant pathways involving upregulated genes in newly diagnosed MM compared with healthy controls from GSE110486. **B** Gene enrichment plot for spliceosome pathway genes in MM samples compared with healthy controls from GSE110486. NES, normalized enrichment score. **C** SRSFs gene expression was associated with shorter OS time in MM; results were obtained from 767 NDMM patients in the MMRF cohort. **D** Comparison of the number of IR events in primary MM samples from MMRF (n = 767) compared with normal plasma cell samples from GEO (n = 13). Violin plots show the median and 25th and 75th percentiles (box) and the 95% confidence interval (whiskers). P value was determined using the Mann-Whitney test.

4.3.3 IR-neoAgs are abundant in multiple myeloma

Emerging evidence suggested that IR events in the cancer genome can be a source for immunogenic peptides [4]. Therefore, we investigated the potential for IR events to produce neoantigens in MM. To begin to address this question, we filtered the IR events that also occurred in normal plasma cells from the events identified in MMRF RNA-seq data. IR events occurring in normal plasma cells were removed because they were not expected to produce immunogenic peptides due to host immune tolerance. To identify the IR events in the healthy plasma cells, we analyzed RNA-seq data from the 13 plasma cell samples in GSE110486 and GSE114816. We detected a total of 9715 IR events that appeared in at least one healthy control sample; these IR events were eliminated from the list of events identified in the MM samples. After filtering the normal IR events, the average number of MM-specific IR events per sample was 1009 and ranged from 21 to 4138. Interestingly, gene ontology analysis of 450 genes harboring MM-specific IR events that occurred in more than half of the NDMM samples showed that these genes were enriched in pathways involving RNA processing and RNA transport (Fig. 13A).

To computationally predict IR-neoAgs, we first determined the HLA-I genotype of each MMRF patient using the RNA-seq data. A total of 178 unique HLA-A/B/C alleles were identified from 767 individual patients of the MMRF cohort (HLA alleles and their frequencies in supplementary table not shown). Next, the retained intron sequences were translated into protein sequences, which were then segmented into 8-11 amino acid peptides, where at least one amino acid was translated from the intronic region. Any peptide that could also be generated from normal proteins was further removed. The remaining IR-derived neo-peptides were then evaluated for their predicted binding affinity

with the set of patient-specific HLA alleles using NetMHCpan (v4.1). Peptides with a NetMHCpan predicted rank score less than 2 (the default cutoff from NetMHCpan) were selected as IR-neoAgs. IR-neoAgs were called for 893 RNA-seq samples from the MMRF cohort (including both newly diagnosed and relapsed samples). Not surprisingly, the number of IR events and the IR-neoAg load were highly correlated (Spearman correlation rho = 0.97, p-value < 0.0001, Fig. 13B). We further evaluated whether any HLA allele presented more IR-neoAgs than other alleles at the population level (Fig. 13C). Our results revealed that HLA-B07:02 presented the highest number of neoantigen peptides (N = 59 588); this allele was detected in 10.8% of samples. The most common allele, HLA-A02:01 which was detected in 19.4% of the samples, presenting 31 161 IR-derived peptides.

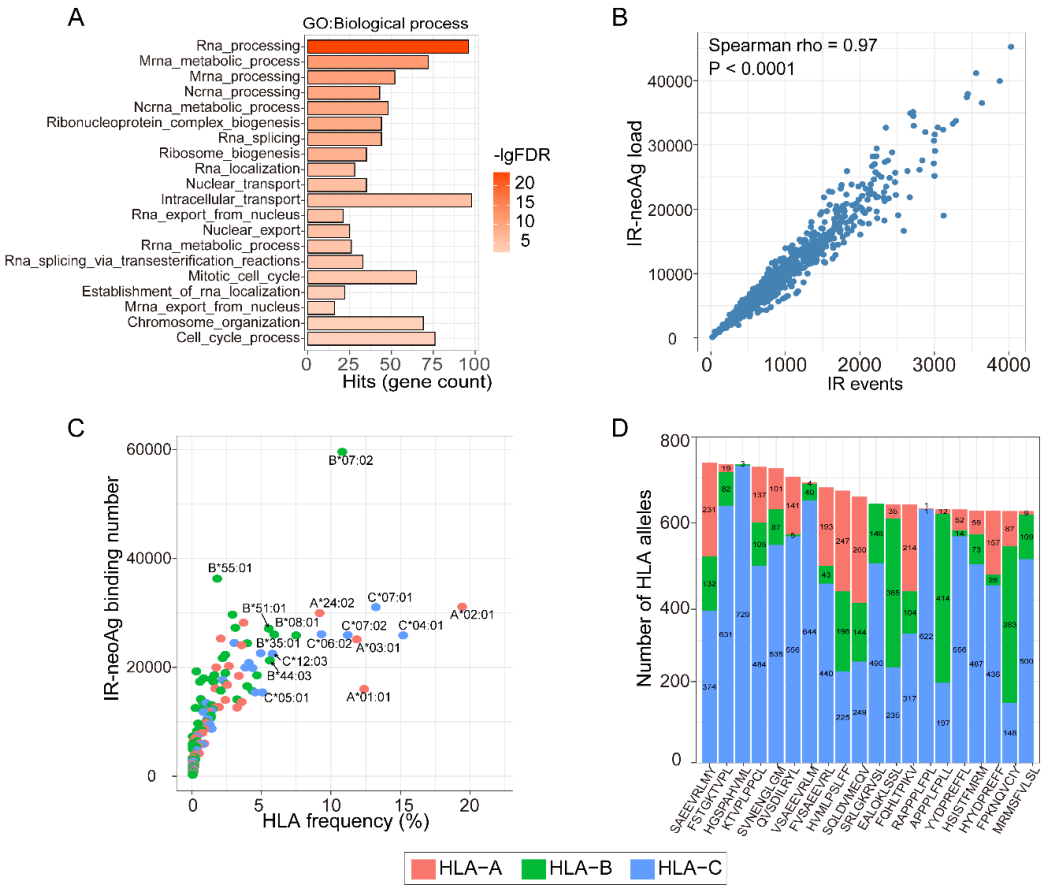

Figure 13 MM-specific IR-neoAgs. **A** Gene ontology enrichment analysis of genes harboring MM-specific IR events in specialized biological processes. **B** Scatter plot showing the number of IR events vs. IR-neoAg load in NDMM patients in the MMRF cohort. Each dot represents an individual patient (N = 767). Spearman correlation rho = 0.97, p-value < 0.0001. **C** Scatterplot showing *HLA* allele frequency vs. the number of predicted IR-neoAgs bound in MMRF NDMM samples. **D** Top 20 most abundant IR-neoAg peptides and the number of *HLA* alleles in the MMRF NDMM samples predicted to bind each peptide; numbers in each bar represent the quantity of *HLA* allele subtypes.

Notably, 24 680 out of 479 685 of the IR-neoAgs were shared across more than 5% of the multiple MM samples. This observation would suggest that there might be potential for developing cancer vaccines in the future based on IR-neoAgs. We also found that 20 neoantigens occurring in more than 80% of NDMM samples were preferentially presented by HLA-C alleles (Fig. 13D), suggesting neoantigens presented by HLA-C alleles could be prioritized for cancer vaccine development.

4.3.4 IR-neoAg load correlates with unfavorable clinical outcome

We next asked whether IR-neoAg load was associated with overall survival (OS) in the MMRF cohort. Kaplan-Meier survival analysis revealed that NDMM patients with higher than the median IR-neoAg load had significantly shorter OS (log-rank test, P = 0.027, Fig. 14A). When considering the expression levels of IR-neoAgs, we observed an even more substantial prognostic effect, with a p-value reaching 0.006 (Fig. 14B). Similarly, higher than the median IR-neoAg load predicted shorter OS for MM patients at the time of first relapse (log-rank test, P = 0.002, Fig. 14C, n=60). Notably, relapsed MM samples with lower IR-neoAg load had a higher 2-year OS rate compared to patients with higher IR-neoAg load (OS 0.85 vs 0.57).

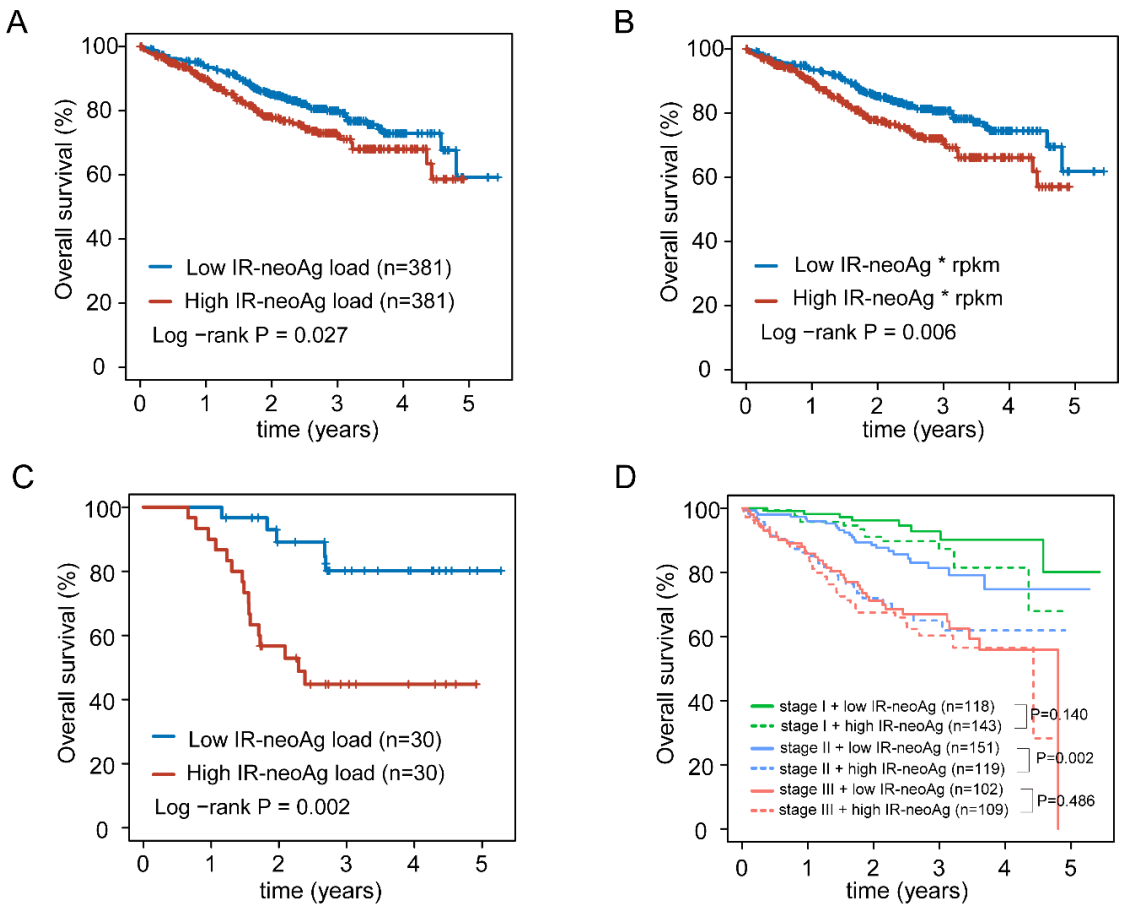

Figure 14 Association of IR-neoAg load with overall survival in the MMRF cohort. Kaplan–Meier survival curves comparing: **A** NDMM patients with high (defined as above the median) IR-neoAg load to those with low (below the median) IR-neoAg load; **B** NDMM patients with high or low expression levels of IR-neoAgs that were quantified using Reads per kilo base per million mapped reads (RPKM) values of the source IR events; **C** MMRF patients with relapsed disease and either high or low IR-neoAg load; and **D** high and low IR-neoAg (IR-neoAg) load subdivided by ISS disease stage.

To determine whether IR-neoAg load was associated with clinical features of MM, we asked whether IR-neoAg load correlated with the International Staging System (ISS) [142], which is a reproducible predictor of MM outcome. We did not find that IR-neoAg load was associated with the ISS disease stage in NDMM from the MMRF cohort (one-way ANOVA P = 0.724, Fig. 15A). To determine whether the addition of IR-neoAg load to ISS stage improved prediction of OS, we performed survival analysis on patients stratified by disease stage and IR-neoAg load. This analysis showed that stage II MM patients with higher than the median IR-neoAg load had significantly shorter OS than stage II patients with low IR-neoAg load (log-rank test, P = 0.002, Fig. 14D). A similar trend was observed with stage I patients, although the association did not reach statistical significance (P = 0.14). IR-neoAg load had no apparent prognostic value for OS in stage III MM patients (P = 0.486). In addition to ISS stages, chromosomal hyperdiploidy (HRD) is widely used in defining genetic subtypes of MM patients, and HRD-myeloma is associated with better survival compared to nonhyperdiploid (nHRD) MM [143]. Although we observed that higher than the median IR-neoAg load was apparently associated with shorter OS in both HRD and nHRD MM patients, these associations did not reach statistical significance (Fig. 15B).

We further examined whether the prognostic performance of IR-neoAg load was independent of other clinical factors. We used multivariate Cox analysis to test the performance of IR-neoAg load after adjusting for other clinical factors, including age, sex, P53 status, ISS stage, as well as the revised ISS stage after adjusting for lactate dehydrogenase (LDH) level, chromosomal aberrations and other factors. In the multivariate analysis, the HR of high versus low IR-neoAg load for OS in NDMM was

1.491 (p-value = 0.027; 95% CI 1.056 to 2.492) (Table 1), indicating that the IR-neoAg load offers prognostic power that is independent of other clinical factors.

Table 1 Univariate and multivariate Cox regression analysis of OS in NDMM.

	Univariate			Multivariate		
Variable	HR	95% CI	P-val	HR	95% CI	P-val
IR-neoAg (high/low)	1.431	1.040-1.968	0.027	1.622	1.056-2.492	0.027
Age (years)	1.038	1.021-1.055	<0.001	1.046	1.023-1.069	<0.001
Gender (male/female)	1.536	1.089-2.165	0.0140	1.537	0.954-2.476	0.077
Stage						
ISS (I/II/III)	2.038	1.640-2.532	<0.001	1.442	0.963-2.159	0.076
Revised ISS	2.398	1.760-3.266	<0.001	1.496	0.865-2.588	0.150
TP53 status						
TP53_Loss	1.088	0.823-1.438	0.555	1.343	0.806-2.238	0.257
BI_TP53	0.622	0.450-0.859	0.004	0.652	0.206-2.062	0.467
NS_TP53	2.755	1.603-4.732	<0.001	1.153	0.145-9.146	0.893

IR-neoAg, Intron retention-induced neoantigen; ISS stage, Myeloma International Staging System; HR, Hazard ratio; CI, confidence interval; Revised Stage (R-ISS) was calculated as defined by the International Myeloma Working Group, by considering LDH, β2-microglobulin, albumin, deletion of chromosome17p, and translocations; TP53_Loss: TP53 copy number variation; BI_TP53: bi-allelic p53 status; NS_TP53: presence of non-synonymous mutation on TP53.

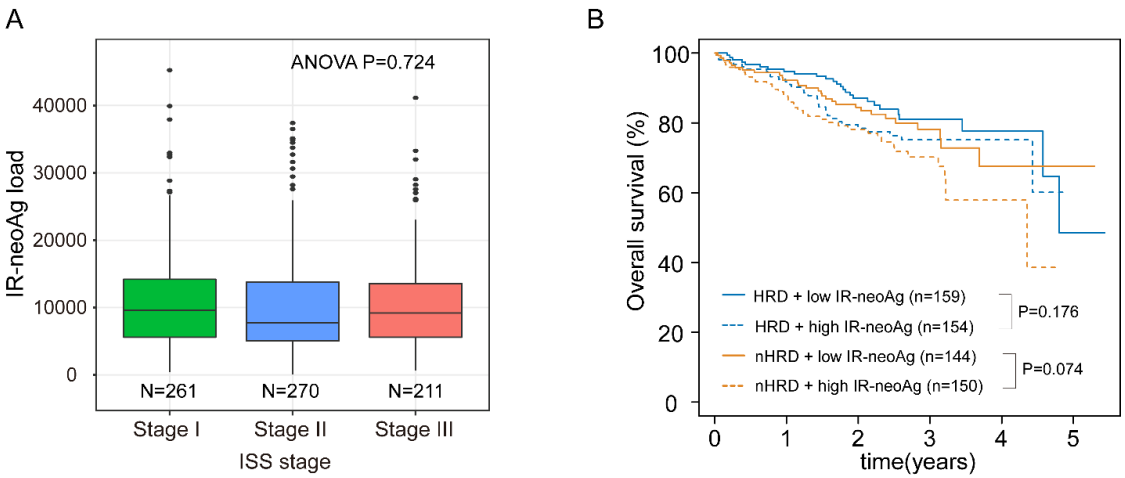

Figure 15 IR-neoAg load correlated with unfavorable clinical outcome of newly diagnosed MM. **A** Distribution of the IR-neoantigen load in MM patients with different ISS stages. **B** Kaplan–Meier survival curves show overall survival in newly diagnosed MM patients with hyperdiploidy (HRD) or nonhyperdiploid (nHRD) and low or high IR-neoantigen load.

4.3.5 Higher T cell inhibitory signals associate with IR-neoAg and poor prognosis in MM

Our observation that higher IR-neoAg load was associated with shorter OS is consistent with previous reports of mutation-derived neoantigen load in MM [120, 121]. However, this finding is the reverse of previously reported observations that high mutation-derived and IR-neoAg loads are associated with longer OS in patients with solid tumors, including melanoma [96], lung cancer [117], breast cancer [119], and pancreatic cancer [116]. In addition, there is increasing evidence that T cells present in the MM microenvironment show an exhausted and suppressed phenotype [144]. This would suggest that additional changes in MM plasma cells may affect the anti-MM immune response.

To test this hypothesis, we conducted differential expression and gene set enrichment analysis on RNA-seq data from the MMRF cohort, comparing samples from NDMM patients with either higher or lower than the median IR-neoAg load. Notably, we observed a significant enrichment of the pathways related to T cell suppression. We found that IFN gamma signaling and TNFα signaling via NF-κB pathways were upregulated in patients with high IR-neoAg load (Fig. 16A). These pathways are involved in the recruitment of T-regulatory (Treg) cells that control cytotoxic T-cell killing. In addition, the B-cell receptor (BCR) signaling pathway was significantly enriched in patients with high IR-neoAg loads. Previous studies demonstrated that sustained activation of BCR signaling plays critical roles in B-cell malignancies [145]. This result suggests that molecular features in cells with higher IR-neoAg load might contribute to T-cell and B-cell dysfunction in MM.

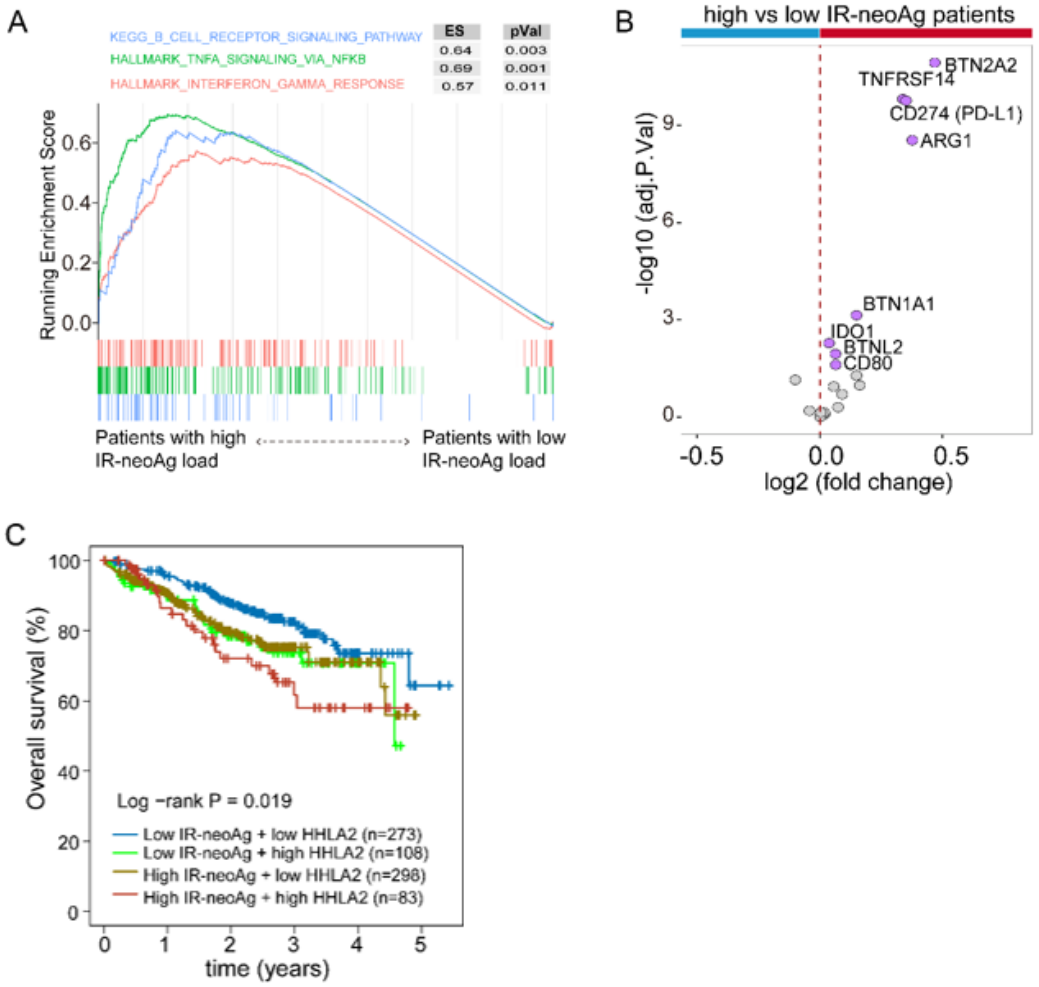

Figure 16 High T-cell inhibitory signature in MMRF patient cohort. **A** Gene set enrichment analysis comparing NDMM patients with high and low IR-neoAg loads. Pathways involved in T-cell suppression and B-cell receptor signaling were enriched in patients with high IR-neoAg load. **B** T-cell signaling co-inhibitory genes were upregulated in patients with high IR-neoAg load. The genes with adjusted p-value < 0.05 are labeled in purple; the blue bar indicates downregulated genes in NDMM patient samples with high IR-neoAg load and the red bar indicates upregulated genes. **C** Kaplan–Meier survival curves showing overall survival in MMRF cohort patients with high (>= 75 percentage) and low (< 75 percentage) expression of *HHLA2* and high (above the median) or low (below the median) IR-neoAg load.

Based on this finding, we postulated that increased expression of T-cell co-inhibitory molecules in MM cells exhibiting high IR-neoAg load might be a partial explanation for the reduced antitumor immunity and thereby facilitate cancer immune evasion [146]. These co-inhibitory molecules function as brakes to inhibit T-cell activation. Higher expression levels of co-inhibitory ligands on the cancer cell surface can negatively impact T-cell function. To begin to address this question, we first analyzed the expression levels of 20 co-inhibitory genes identified by Dufva and colleagues [147], which include genes for 8 B7 ligands, 6 enzymes impacting T-cell activity, and 6 other genes from the butyrophilins and CD226 family. We found that these co-inhibitory genes tend to have higher expression levels in NDMM samples with higher IR-neoAg load (Fig. 16B). We found that CD274 (PD-L1) expression was 1.3-fold higher in patients with high IR-neoAg load (adjusted p-value < 0.0001), suggesting there could be a stronger immune suppression in patients with higher IR-neoAg load. Next, we analyzed co-inhibitory gene expression from 29 MM cell lines compared to other cancer cell lines in CCLE. Surprisingly, we found that the average expression level of PD-L1 in the MM cell lines was lower than most other types of cancer cell lines (Fig. 17), which might partially explain why anti-PD1 therapy has had a limited response rate in MM. Other B7 co-inhibitory ligands, such as CD86, CD80 and HHLA2, showed high expression levels in myeloma cell lines relative to the other cancer cell lines, implying that these B7 ligands might serve as potential targets for immune checkpoint therapy. Kaplan-Meier survival analysis revealed that the patients with higher HHLA2 and IR-neoAg load had the worst outcome (Fig. 16C), which provides further support that HHLA2 may be a druggable target for treating MM in the future [148].

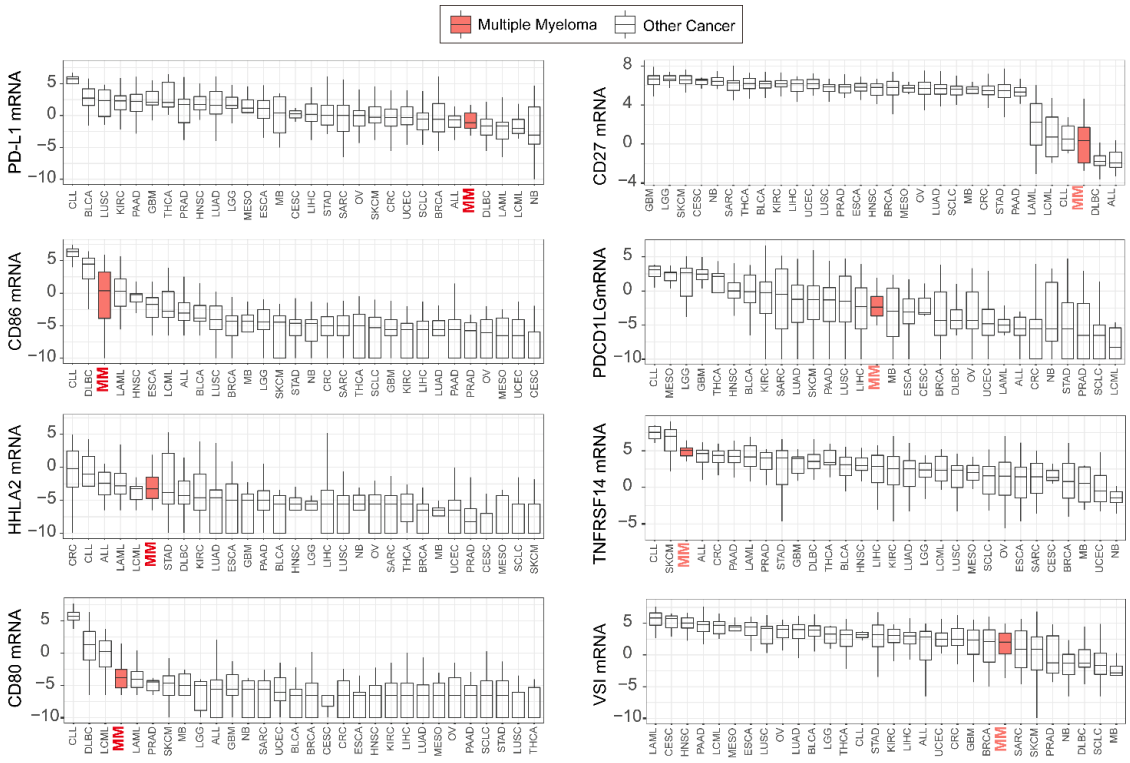

Figure 17 The B7 ligand genes in MM and other cancer cell lines in CCLE. The red box represents the 29 MM cell lines in CCLE. mRNA expression values were log2(TPM + 1) normalized. CCLE, the Cancer Cell Line Encyclopedia (CCLE) database

4.3.6 RNA splicing inhibition impacts MHC-I protein expression in MM cells

MHC molecules encoded by the HLA-I and HLA-II genes are essential components in IR-neoAg presentation on the cell surface. Therefore, we investigated the relationship between IR events (IR levels) and MHC protein abundance in four MM cell lines, namely JJN3, U266, KMS11, and AMO1 cells. These cell lines were selected because KMS11 and U266 had the highest levels of IR, while JJN3 and AMO1 had the lowest levels of IR based on RNA-seq data from the CCLE consortium (Fig. 18A). We measured MHC-I and MHC-II cell surface abundance in these MM cells by flow cytometry before and after treatment with the splicing inhibitor pladienolide-B for 96 hours.

As demonstrated in Fig. 18B, the basal cell surface level of MHC-II was lower in the KMS11 and U266 cell lines bearing higher IR levels, compared to the JJN3 and AMO1 cell lines bearing lower IR levels. Low MHC-II abundance in MM cells with high IR levels is consistent with our observation in MM patients where higher IR levels and low HLA-II gene mRNA expression was associated with worse clinical outcomes (data not shown).

Next, we investigated whether splicing activities, as measured by the mRNA expression levels of genes encoding key splicing factors and regulators, correlated with the mRNA expression levels of the HLA genes encoding MHC-I and MHC-II complexes in the MMRF RNA-seq data. We observed a negative correlation between the expression levels of MHC-I genes and spliceosome pathway activities, as measured by ssGSEA enrichment scores (Fig. 18C, Spearman correlation rho = -0.245, $p < 0.001$). No correlation was observed between the expression of genes encoding MHC-II molecules and spliceosome pathway activities (Fig. 19A).

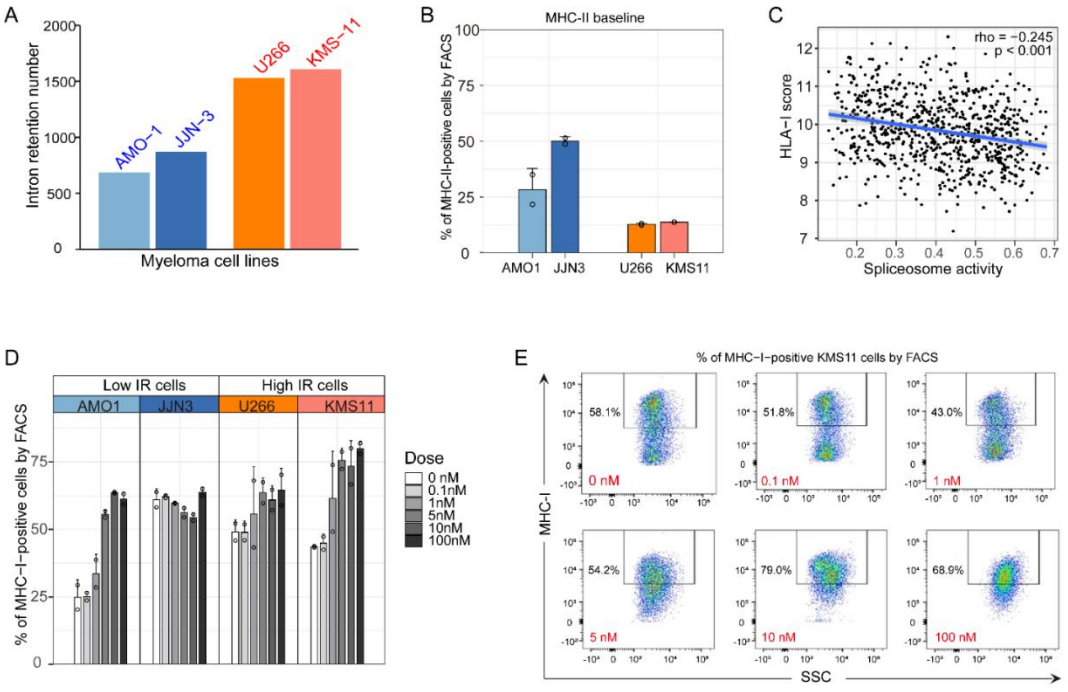

Figure 18 Intron retention, spliceosome activity, and MHC abundance in MM cells. **A** Number of intron retention events determined from CCLE RNA-seq data in a panel of four MM cell lines. **B** MM cell lines with high IR levels had lower baseline MHC-II cell surface expression determined by flow cytometry. **C** *HLA class I* gene score (average expression of *HLA-A/B/C* alleles) from MMRF RNA-seq data was negatively associated with spliceosome pathway activity. **D** MHC-I genes were upregulated in a dose-dependent manner in 3 of 4 MM cell lines following treatment with the splicing inhibitor pladienolide-B (0 to 100 nM) for 96 h. **E** Representative flow cytometry charts illustrating the percentage of KMS11 cells with increased MHC-I (*HLA-A/B/C*) gene expression following treatment with pladienolide-B (0 to 100 nM) for 96 h. Gates were set based on isotype controls and unstained controls.

To investigate whether low MHC-I expression might be a result of increased splicing activity, we treated MM cell lines with the splicing inhibitor pladienolide-B, which targets SF3B1, a gene encoding subunit 1 of the splicing factor 3b protein complex, and measured MHC-I cell surface expression by flow cytometry. We found that MHC-I expression levels were significantly increased in 3 of the 4 MM cell lines, including both cell lines with higher IR levels (Fig. 18D). As shown in Fig. 18E, MHC-I cell surface abundance in KMS11 MM cells exhibiting high IR increased following pladienolide-B treatment. This finding strongly suggests that modulation of splicing activity may regulate the abundance of MHC-I class proteins along with the antigen presentation potential in MM cells. Consistent with the lack of correlation between HLA-II gene expression and spliceosome pathway activity, no significant changes in MHC-II protein abundance were observed in the MM cell lines after splicing inhibition (Fig. 19B).

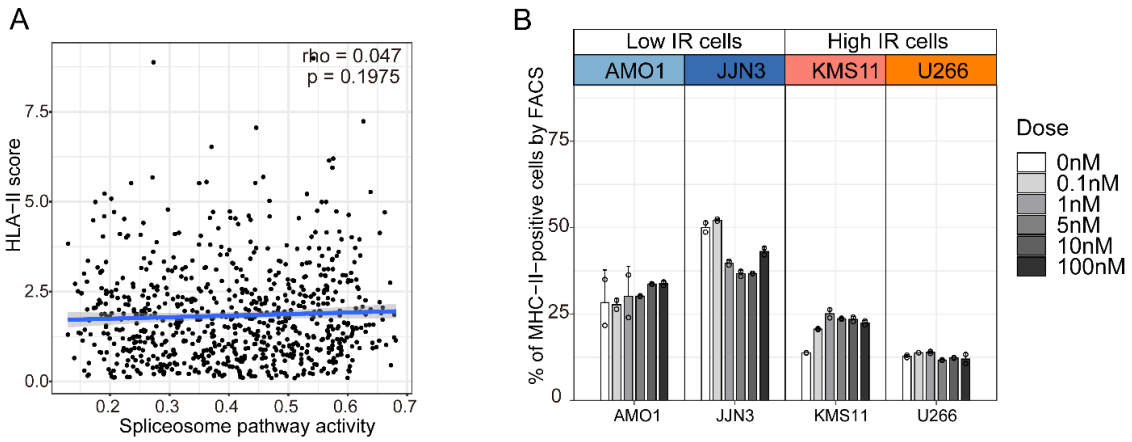

Figure 19 MHC-II expression levels were defined in MM cell lines by flow cytometry. While significant differences were observed between cells bearing high IR (KMS11 and U266) and cells bearing low IR (JJN3 and AMO1), no significant differences in MHC-II abundance were observed following spliceosome inhibitor treatment for 96 hours. **A** HLA class II gene scores (average expression of HLA-DPA1/DPB1/DQA1/DQB1/DRB1 alleles) from MMRF RNA-seq data were not associated with spliceosome pathway activities. **B** MHC-II cell surface levels were not changed in either high or low IR MM cell lines following pladienolide-B treatment.

4.4 Discussion

In this study, we demonstrate that intron retention is an important source of neoantigens in multiple myeloma, which impacts patient clinical outcome. We showed that newly diagnosed MM samples exhibited more intron retention events than normal plasma cells and that higher IR-neoAg load was significantly associated with unfavorable survival in both newly diagnosed and relapsed MM. Our findings indicate that bioinformatic predictions of immune recognition of neoantigens arising from genomic or transcriptomic alterations in MM might not be useful in selecting patients for immune checkpoint therapy. Further, our analyses revealed that poor outcome in MM patients with high IR-neoAg load is associated with higher expression levels of checkpoint genes and elevated IFN signaling activity, which implies strong T-cell suppression. Therefore, our results suggest a potential mechanism for MM cell immune evasion despite having an increased neoantigen load compared to normal plasma cells.

Whereas high neoantigen load generally predicts favorable survival and higher likelihood of response to checkpoint blockade in many solid tumors such as breast cancer [119], lung cancer [117], glioblastomas [118] and melanoma [96], we found that a high neoantigen load in MM patients was associated with poor prognosis [120]. In addition, the immune context of the bone marrow microenvironment is more complex compared with solid tumors, where cytokines and immune cell components in the bone marrow provide a unique seedbed for myeloma cell growth [149]. Therefore, the underlying mechanisms that allow for MM cell immune escape are apparently different from other tumors.

In addition to somatic DNA mutations, RNA alternative splicing, including intron retention, was reported to be a novel source of neoantigens [4]. Numerous studies have

reported that the splicing machinery is dysregulated in multiple cancer types, including bladder cancer [46], breast cancer [48], melanoma [150], prostate cancer [151] and hematological cancers [152, 153]. In addition, IR events have been observed frequently in prostate cancer [154] and pancreatic cancer [155]. Yang et al. reported that blood cells have a high level of splicing diversity comparing to other tissues, next to testis, brain, and muscle-skeletal tissue, in the GTEx transcriptional data [156]. IR events represent a large proportion of alternative splicing events in blood tissue. These findings prompted us to investigate whether the IR-neoAg could contribute to immune responses in MM, in particular to antigen presentation and T-cell mediate responses. We found that higher IR-neoAg load was significantly associated with shorter survival time, both in newly diagnosed and relapsed MM. This finding was further strengthened when the expression levels of IR-neoAg were considered (p-value reached 0.006).

Over the past decade, ICB therapy has revolutionized cancer therapy in several tumor types [157]. However, response to the immune checkpoint inhibitor pembrolizumab (anti-PD1) has been limited in MM [120]. Clinical response to ICB has been closely linked with the abundance of tumor-specific neoantigens, the presence of cytotoxic T-cell infiltration, and distinct tumor microenvironment profiles. Previous reports have demonstrated an increase in the mutation-derived neoantigen load in MM and have also confirmed a neoantigen T-cell response in relapsed patients with MM [121]. These results implied that a T-cell mediated immune response might be suppressed or impaired in MM. Zelle-Rieser et al. reported that CD8+ T-cells expressed several molecules associated with T-cell exhaustion (PD-1, CTLA-4, CD160) as well as the T-cell senescence marker CD57 at the MM tumor site [144, 158]. Our results showed that higher IR-neoAg load was

positively correlated with higher expression levels of T-cell inhibitory molecules and genes belonging to the Tregs activating pathway. Dufva et al. reported that decreased HLA-II gene expression might be a potential immune evasion mechanism in hematological cancers [147]. We also found that gene expression levels of HLA-II genes were significantly lower in newly diagnosed MM compared with healthy control cells.

Despite these RNA-seq-based observations, direct evidence of IR-neoAg presentation on MM cells using immune-peptidomics technology could strengthen our conclusion. We hypothesize that standard MM treatment options do not generate an effective immune response that leverages the neoantigen immunotherapeutic potential. Indeed, we observed lower levels of MHC-II activity in MM cell lines with higher intron retention. In addition, we observed that MHC-I activity appeared to be inhibited in cells with elevated expression levels of splicing factors, a hallmark of MM, and that inhibition of spliceosome activity resulted in increased MHC-I activity. Collectively, these two mechanisms may partially explain why higher IR-induced neoantigen load in MM samples was not associated with better prognostic outcome. This result also suggests that splicing inhibitors could possibly boost the efficacy of immune checkpoint blockade therapy in MM by activating MHC-I presentation [159, 160]. Further analysis with integrated multi-omics data from different aspects of the immune landscape is needed to further understand the potential determinants of responsiveness to cancer immunotherapies in MM.

In conclusion, while neoantigen load has been associated with favorable survival in many solid cancers, our study strongly suggests that IR-neoAg load may serve as a clinically relevant risk factor that negatively impacts myeloma patient survival. Our analysis provides evidence that MM cells bearing high levels of IR-neoAgs also present T-

cell inhibitory gene signatures, which may offset the neoantigen load in eliciting a cytotoxic T cell response. Moreover, we found that aberrant RNA splicing may also regulate MHC abundance and thus, contribute to MM immune escape. Our findings highlight the need to integrate multi-omics data to uncover the immune context and understand the factors that determine responsiveness of MM to immunotherapies. Also, this works suggests that targeting splicing may represent an additional therapeutic strategy to promote anti-MM immune response.

Chapter 5 Intron Retention Neoantigen Load Predicts Favorable Prognosis in Pancreatic Cancer

5.1 Introduction

Advances in ICB therapy using antibodies that block PD-1 and CTLA-4 have resulted in remarkable clinical responses in a wide variety of cancer patients [30, 161, 162]. Response to ICB therapy correlates with high tumor mutation burden (TMB) and the cell surface display of tumor neoantigens by MHC molecules, which is critical for T cell recognition and immune-mediated killing of tumor cells [163-166].

Pancreatic cancer consists of two categories, namely cancers developing from the exocrine cells that make up the exocrine glands and ducts of the pancreas and cancers that develop from cells in the endocrine glands. More than 95% of pancreatic cancers arise from exocrine cells, with pancreatic ductal adenocarcinoma (PDAC) being the most common histologic type, as well as one of the deadliest with a 5-year survival rate less than 8% [167]. PDAC is considered an immune-privileged tumor and to date, ICB therapy has not shown efficacy in PDAC patients, nor affected OS [168, 169]. One explanation for why PDAC responds poorly to ICB therapy is a low TMB [170]. However, recent studies have shown that some PDAC tumors do express neoantigens and exhibit T cell infiltration [171, 172]. These findings imply that some patients are able to generate an antitumor immune response and therefore, may benefit from ICB therapy.

In addition to somatic DNA mutations, aberrant RNA transcripts can be a major source of neo-peptides when retained introns are translated and degraded through the nonsense-mediated decay mechanism [4]. IR commonly occurs in a wide variety of cancer transcriptomes, including PDAC, compared to normal tissues [14, 173]. Additionally, high

IR levels in PDAC appears to be an independent predictor of tumor progression [155]. Therefore, we hypothesized that in the presence of low TMB, aberrant IR induced neoantigens (IR-neoAg) contribute to immune-mediated clearance of pancreatic cancer cells.

To begin to test this hypothesis, we herein employed in silico prediction on RNA-sequencing data from two large independent cohorts of pancreatic cancer patients from TCGA and ICGC pancreatic cancer cohorts to determine whether IR-neoAg load was associated with longer overall survival for these patients. We also estimated tumor infiltrating immune cell proportions and determined the association of IR-neoAgs with various tumor lymphocyte populations. In addition, we investigated the association of IR-neoAg load with expression of immune checkpoint genes and HLA genes encoding the MHC class-I molecules in pancreatic cancer. Finally, we compared gene expression profiles in pancreatic cancer with those of melanoma tumors that responded to anti-PD1 checkpoint therapy. The results from this study could be useful in selecting pancreatic cancer patients who might benefit from ICB therapy.

5.2 Methods

5.2.1 Pancreatic cancer and normal pancreas datasets

RNA-seq data from TCGA-PAAD (n = 178) with clinical and pathologic characteristics were downloaded from the Genomic Data Commons (GDC) [98]. RNA-seq data from the ICGC Pancreatic Cancer cohorts (ICGC-PDAC-AU, n = 81) were downloaded from the ICGC data-portal [174]. One sample each from TCGA-PAAD and ICGC-PDAC-AU cohorts was missing survival event time and was excluded from survival analyses. RNA-seq data from normal pancreas samples (n = 68) were retrieved from GTEx

projects [175]. Four PDAC microarray datasets, GSE15471, GSE16515, GSE28735 and GSE62452 [176-179], were downloaded from the GEO database and the Bioconductor affy package was used for raw data processing and normalization.

5.2.2 Identification of IR events

Raw fastq files were aligned to GRCh38 reference genome using STAR (v.2.7.2) [80]. The exon sets for each protein-coding gene were re-annotated using GTF files (Gencode.v32). The union of each exon set was used to define introns as the interval between exon sets. Exons and introns were quantified using uniquely mapped reads based on the re-annotated GTF file using the HTseq [82]. IR events were further filtered using the following criteria: (1) both the intron region and its flanking exon regions had read counts >10 and (2) the TPM ratios of the intron to flanking exons was greater than 0.05 and less than 0.5. These filters allowed identification of IR events that had expression levels that were comparable with the flanking exons in the mRNA transcripts that were composed of a mixture of normal and aberrant splicing products. IR events that were also observed in at least 25% of normal pancreas RNA-seq datasets from GTEx [175] were filtered to obtain the final set of pancreatic tumor-specific IR events.

5.2.3 IR-neoAg prediction

To obtain the set of neo-peptides derived from retained introns, the open reading frames of the upstream exons were extended into introns until the first stop codon. The translated peptides were segmented into fragments of 8 to 11 amino acids that contained at least one intron-encoded amino acid. We estimated the binding affinity of each IR-derived neo-peptide with specific MHC class-I molecules for each sample using NetMHCpan (v.4.1)[83]. Each patient's HLA genotype was deduced from RNA-seq data using

arcasHLA (v1.1) [64]. NetMHCpan compares raw prediction scores to a set of random natural peptides to calculate the % rank, which provides robust binding metrics[83] . Neo-peptides with % rank < 0.5 were defined as strong binders, as recommended by NetMHCpan, and were considered as IR-neoAgs in this study.

5.2.4 Tumor immune cell proportions and prediction of immunotherapy response

The relative proportion of 22 types of tumor-infiltrating immune cells were inferred from bulk RNA-seq expression data using the online CIBERSORT application [180], with parameters set as LM22 signatures and 100 permutations. We used SubMap[181] to evaluate the similarity between global gene expression patterns of TCGA-PAAD tumors with either high or low IR-neoAgs and samples from immunotherapy-treated melanoma patients[182]. The mapping information generated by SubMap identifies subclasses common to two independent datasets and calculates the probability that they share similar biological properties. Default parameters were used [183].

5.2.5 Differential expression and pathways enrichment analysis

Salmon (v1.2.1) was used to quantify the gene expression levels from RNA-seq data[101], using the reference annotation GRCh38 (gencode.v32). Differentially expressed genes among patient groups were identified with the *limma* package in R[136]. The *clusterProfiler* package was used to test the pathway enrichment significance. P values were adjusted using the Benjamini-Hochberg method [138].

5.2.6 Statistical analysis

All analyses and visualization were performed in R (v.4.0.2). Kaplan-Meier survival estimates and Cox-proportional hazard analysis were performed with log-rank test and hazard ratio (HR) to compare patient groups, using the *survival* package [184].

Mann-Whitney-Wilcoxon Test was used to compare differences in the value distribution between groups [185].

5.3 Results

5.3.1 IR is a potential source of neoantigens in PDAC

We first surveyed PDAC primary tumors in the TCGA SpliceSeq database to examine the frequency of IR compared to other alternative splicing events. We found that IR accounted for 16.5% of all alternative splicing events, which was second behind skipped exon (56%) and more common than alternative 5'- and 3'-splice site and mutually exclusive exons (13.7%, 12.1%, and 0.7%, respectively). Using KEGG gene set enrichment analysis, we also found that the spliceosome pathway was significantly upregulated in PDAC compared to adjacent normal tissue (Fig. 20A-D). These findings are consistent with those of Wang et al., who first reported the frequency of different alternative splicing events and dysregulation of the spliceosome machinery using Affymetrix exon array data from PDAC tissues [173].

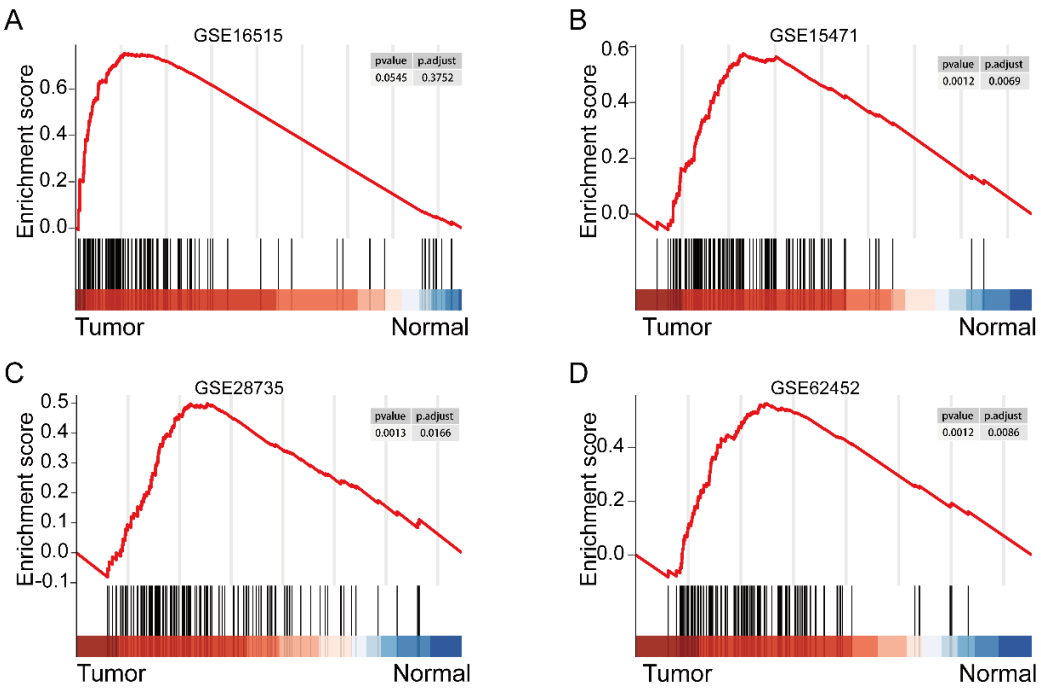

Figure 20 Spliceosome pathway is unregulated in PAAD. A-D. The spliceosome pathway is altered in PAAD tumors. Gene enrichment plots revealed that spliceosome pathway genes were enriched in tumor samples compared with adjacent normal tissue. Differential gene expression data were from GSE16515 (**A**), GSE15471 (**B**), GSE28735 (**C**), and GSE62452 (**D**).

Next, we investigated the immunogenic potential of IR-derived neo-peptides using RNA-seq data from 178 pancreatic cancer patient samples in the TCGA-PAAD cohort. An overview of the data processing steps in our computational pipeline to identify IR events and IR-neoAgs is diagrammed in Fig. 21A. IR events were first identified in 68 normal pancreas GTEx RNA-seq datasets. We found a total of 5,927 IR events that occurred in at least 25% of the GTEx samples (supplementary data not shown). After filtering these normal IR events, there were an average of 600 tumor-specific IR events per sample. Each patient's *HLA* genotype was further deduced from RNA-seq data and used to predict the number of IR-neoAgs in their tumor sample. A total of 171,526 IR-neoAgs were predicted with an average IR-neoAg load of 963 (range 43 - 4,284) per tumor sample. The number of IR-neoAgs per tumor was strongly correlated with the number of IR events (Spearman correlation rho = 0.93, $p < 0.001$, Fig. 21B). Interestingly, the predicted number of IR-neoAgs presented by each *HLA* type was not related to *HLA* allele frequency (Fig. 21C). For example, *HLA-A02:01* showed the highest allele frequency (24.2%), but this allele was only predicted to present 11,014 (6.4%) of the IR-neoAg peptides. In contrast, *HLA-B07:02* had a lower allele frequency (8.7%) but was predicted to present 23,481 (13.7%) IR-neoAgs. These findings suggest that a patient's *HLA* genotype is an important factor in predicting immunogenicity of potential IR-neoAgs.

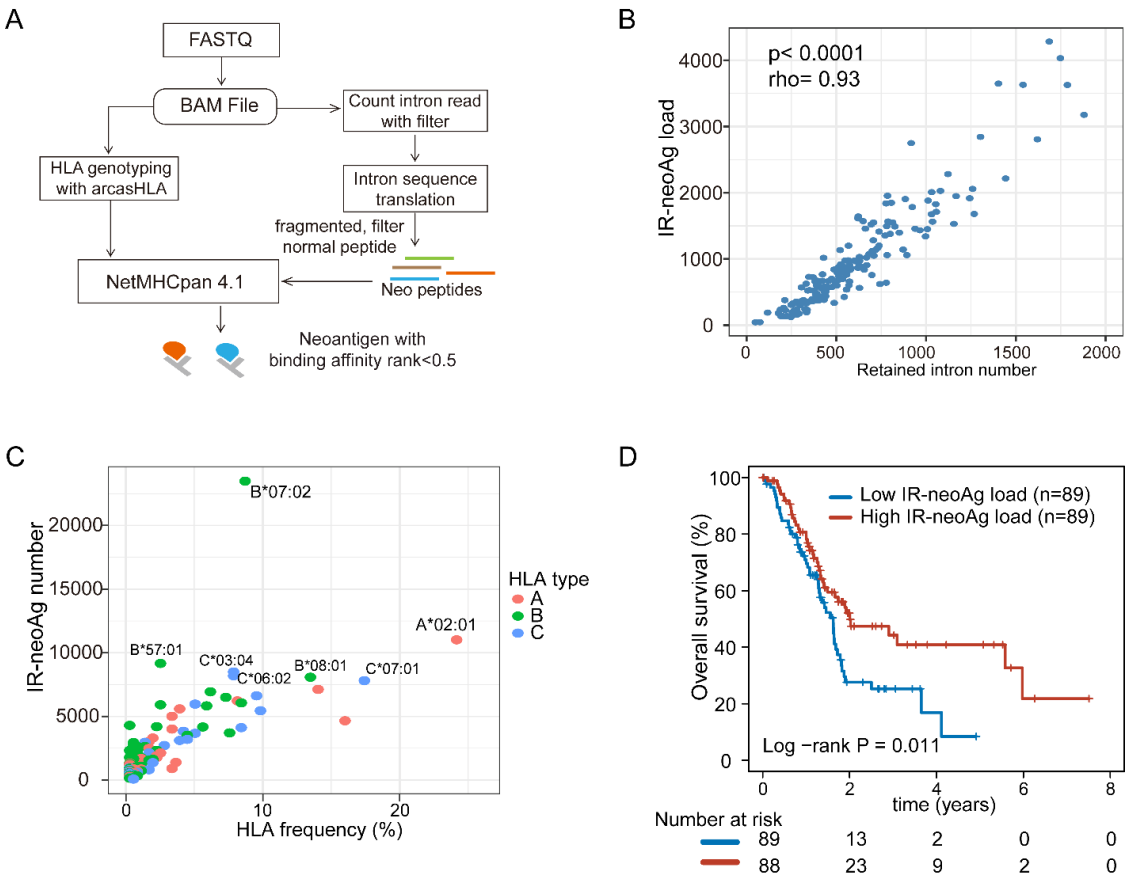

Figure 21 IR-neoAgs predicts favorable survival. **A** Workflow for identifying IR-neoAgs from RNA-seq data. **B** Scatter plot showing the correlation between the number of IR events and the IR-neoAg load in the TCGA-PAAD RNA-seq data. Each dot represents an individual patient (N = 178). **C** Scatter plot showing HLA allele frequency in the TCGA-PAAD dataset vs. the number of IR-neoAgs each HLA allele can potentially present. Each dot represents an individual allele type. **D** Kaplan–Meier curves for overall survival of groups with high and low tumor IR-neoAg load in the TCGA-PAAD cohort. A risk table is displayed below the plot.

5.3.2 IR-neoAg load is an independent prognostic factor for pancreatic cancer

TCGA-PAAD tumors (n = 178) were divided into high and low IR-neoAg groups based on the median IR-neoAg number. The high and low IR-neoAg groups were not significantly different with respect to age, sex, tobacco smoking, or for clinical features such as tumor stage, grade, or microsatellite instability (Table 2). However, the median OS time for patients with high IR-neoAg load was 24 months compared to 19 months for those with low IR-neoAg load (Kaplan Meier log-rank test, P = 0.011; Fig. 21D). The association between high IR-neoAg load and survival was independent of clinicopathological factors (p = 0.008, multivariate Cox regression analysis; Table 3). The HR was 0.55, indicating that high IR-neoAg load reduced the risk of poor outcome by 45%. High IR-neoAg load was also associated with longer survival in the ICGC PDAC cohort (n = 81), although these findings did not reach statistical significance likely due to the smaller sample size (Fig. 22).

Table 2 Clinical and pathologic characteristics of TCGA-PAAD dataset.

Characteristics	No.	Low IR-neoAg load (%)	High IR-neoAg load (%)	P-val
Age (years)				
Median	65	65	65	
IQR	57-73	57-74	56-72	
≤60	59	29 (32.6)	30 (33.7)	0.999
>60	119	60 (67.4)	59 (66.3)	
Gender				
Female	80	41 (46.1)	39 (43.8)	0.880
Male	98	48 (53.9)	50 (56.2)	
Tumor Stage				
I–II	168	83 (93.3)	85 (95.5)	0.859
III-IV	8	5 (5.6)	3 (3.3)	
Unknown	2	1 (1.1)	1 (1.1)	
Tumor Grade				
G1+G2	126	61 (68.5)	65 (73.0)	0.807
G3+G4	50	27 (30.3)	23 (25.8)	
Unknown	2	1 (1.1)	1 (1.1)	
MSI/MSS Status				
MSS	141	68 (76.4)	73 (82.0)	0.169
Indeterminate	28	18 (20.2)	10 (11.2)	
MSI-L	9	3 (3.4)	6 (6.7)	
Smoking Status*				
NO	121	63 (70.8)	58 (65.2)	0.521
YES	57	26 (29.2)	31 (34.8)	

Smoking status was characterized as smoking exposure by pack-years > 1.

Table 3 Univariate and multivariate Cox regression analysis of OS in TCGA-PAAD patients.

	Univariate			Multivariate		
Variable	HR	95% CI	P-val	HR	95% CI	P-val
IR-neoAg (high/low)	0.586	0.386-0.890	0.012	0.549	0.353-0.854	0.008
Age	1.028	1.007-1.049	0.009	1.023	1.003-1.045	0.028
Gender (male/female)	0.824	0.548-1.238	0.350	0.959	0.617-1.492	0.853
Stage (I/II/III/IV)	1.301	0.889-1.904	0.175	1.211	0.801-1.830	0.364
Grade (1/2/3/4)	1.448	1.089-1.925	0.011	1.270	0.933-1.731	0.129
MSI status	1.111	0.659-1.872	0.694	1.477	0.859-2.541	0.158
TMB	1.000	0.999-1.000	0.660	1.000	1.000-1.000	0.918

TCGA, The Cancer Genome Atlas; HR, hazard ratio; CI, confidence interval;

MSI status includes three categories, Indeterminate/MSS/MSI-L;

MSS, microsatellite stable; MSI, microsatellite instability;

TMB, tumor mutation burden is the count of non-synonymous mutations.

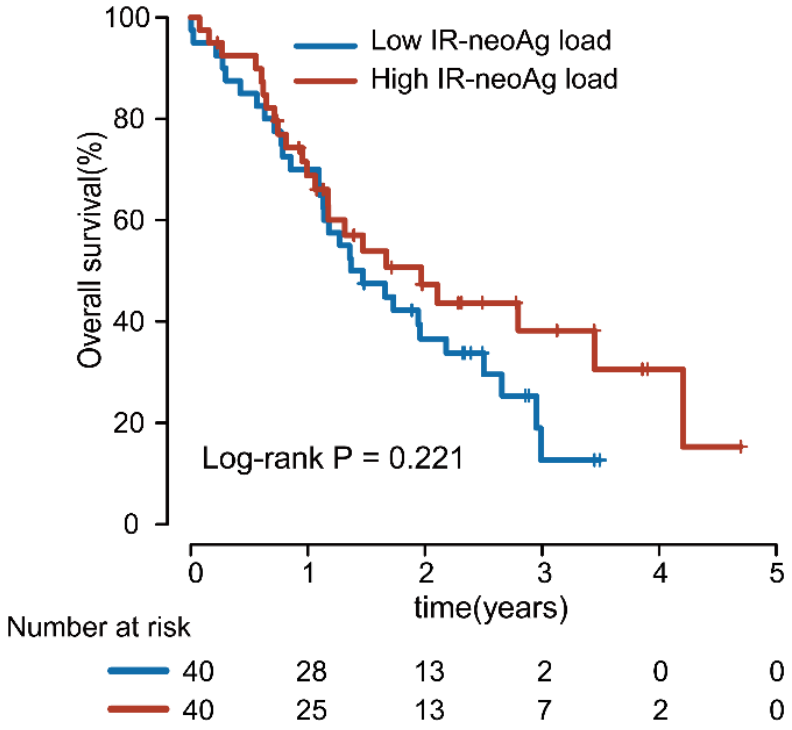

Figure 22 Kaplan-Meier survival curves of ICGC-PDAC patients with high (≥ median) and low (<median) IR-neoAg load.

Our findings agree with those of Tan et al., who showed that IR was associated with PDAC patient outcome in a curated subset of the TCGA-PAAD cohort (n = 150), which excluded non-PDAC tumors such as acinar cell carcinoma, pancreatic neuroendocrine tumors (PanNET), benign neoplasms and tumors with < 1% neoplastic cellularity [186]. Since our initial survival analysis was performed using all 178 TCGA-PAAD samples with RNA-seq data, we repeated the multivariate Cox regression analysis on the curated set of 150 PDAC samples and considered tumor purity as a covariate. Our results showed that high IR-neoAg load remained significantly associated with longer OS (HR 0.605; 95% CI, 0.377 to 0.971; p = 0.037). These results, taken together, support IR and IR-neoAg load as independent predictors of pancreatic cancer progression and patient survival.

5.3.3 IR-neoAg load is associated with features of tumor immune response

To test whether high IR-neoAg load was accompanied by immune cell infiltration in the TCGA cohort, we used a deconvolution algorithm to estimate the relative percentages and activation states of 22 types of immune cells from bulk RNA-seq data [180]. Seven cell types showed statistically significant differences between high and low IR-neoAg load tumors (Fig. 24A). There were more memory B cells (p=0.014), follicular helper T (Tfh) cells (p=0.011), and naive CD4+ T cells (p=0.044) in tumors with high IR-neoAg load. In contrast, these tumors had fewer resting mast cells (p=0.002), resting dendritic cells (p=0.011), M1-macrophages (p=0.013), and neutrophils (p=0.003). However, no significant correlation was observed between IR-neoAg load and the relative percentage of CD8+ T cells or other cytotoxic immune cell types (Fig. 23). While the overall percentage of infiltrating immune cells was low, higher levels of memory B and Tfh cells are generally associated with better prognosis and cancer immunotherapy

response [187-189]. These findings imply that IR-neoAg load may impact the immune cell composition of the tumor microenvironment.

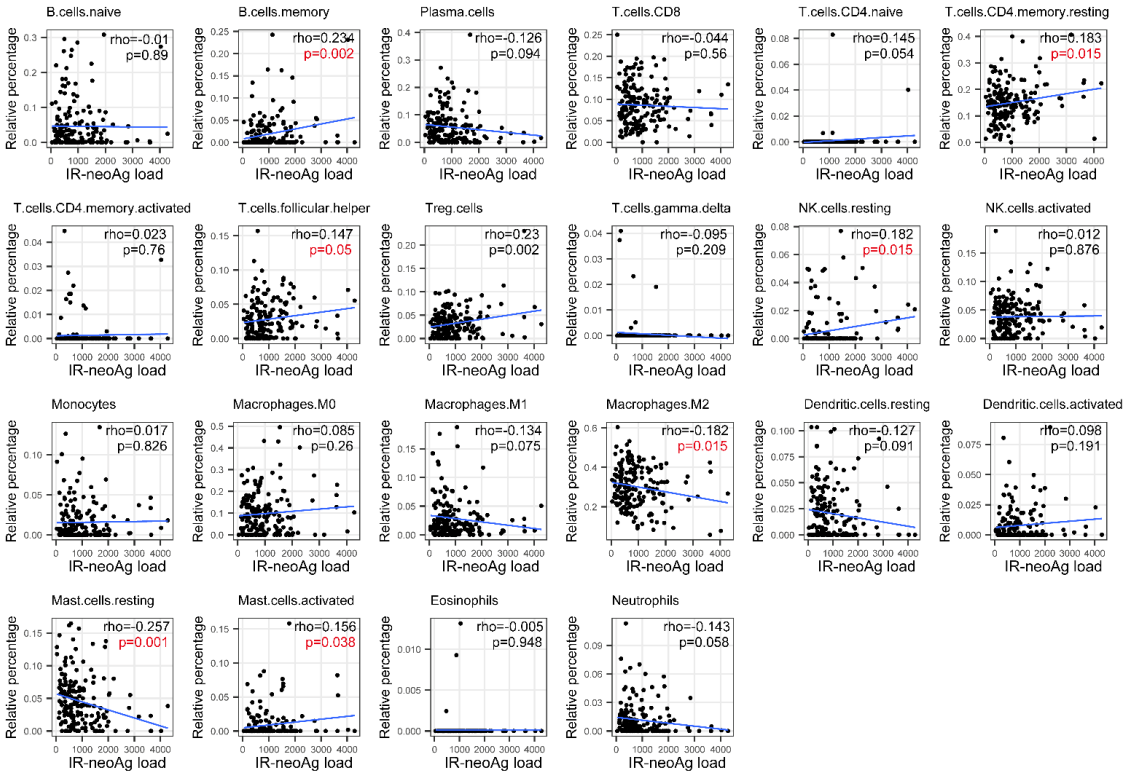

Figure 23 Correlation between IR-neoAg and infiltrated immune cell proportions. The Scatter plot showing the correlation between the IR-neoAg load, and 22 infiltrated immune cell proportion inferred by the Cibersort. Red text shows immune cells significantly associated with IR-neoAgs, $p<0.05$.

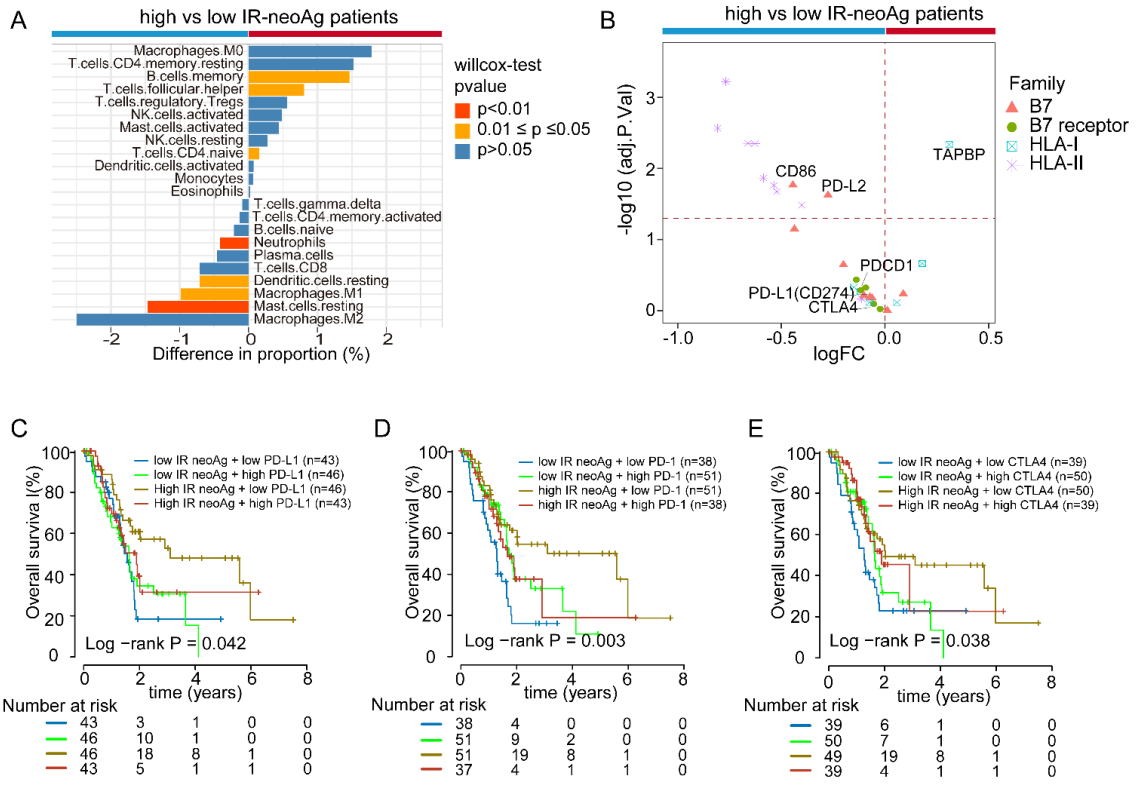

Figure 24 IR-neoAg load is associated with immune features in the TCGA-PAAD cohort. **A** Comparison of the percentage of tumor infiltrated immune cells between tumors with high and low IR-neoAg load. A positive value along the x-axis represents a higher and a negative value represents a lower proportion of cells in the high IR-neoAg samples. P values were calculated using the Mann-Whitney-Wilcoxon Test. **B** Volcano plot showing the expression differences of immune-related genes between the high and low IR-neoAg load tumors. Symbols to the right of the vertical dashed line represent genes with higher expression and symbols on the left represent genes with lower expression in high IR-neoAg tumors. Symbols above the red dashed line represent genes with adjusted p-values < 0.05. **C-E** Kaplan-Meier overall survival curves for four patient groups stratified by IR-neoAg load and the gene expression levels of PD-L1 (**C**), PD-1 (**D**), and CTLA-4 (**E**). A risk table is displayed below the plot.

To better understand the molecular differences that might contribute to OS between pancreatic cancers with high and low IR-neoAg load, we compared the expression levels of a collection of genes related to immune cell response, consisting of nine B7 ligand genes, five B7 receptor genes, six MHC-I genes, and nine MHC-II genes. We found that 12 of 14 B7 ligand and receptor genes had lower expression in samples with high IR neoAg load, with *CD86* and *PD-L2* reaching statistical significance (Fig. 24B). In addition, all nine MHC-II genes were expressed at lower levels in the high IR-neoAg group, with eight of nine reaching statistical significance. Only one MHC-I pathway gene, *TAPBP*, showed significant differential expression between tumors with high and low IR-neoAg load. Thus, longer OS in patients with high IR-neoAg load tumors could be partially explained by low expression levels of immune co-inhibitory genes that dampen effector T-cell responses.

5.3.4 IR-neoAg load together with immune checkpoint gene expression levels are associated with OS

The expression level of immune checkpoint genes, such as *PD-L1*, is associated with ICB response and survival outcomes in multiple cancers [190-192]. Therefore, we asked whether the correlation of IR-neoAg load and patient survival was associated with the expression levels of immune checkpoint genes. To address this question, we stratified TGCA-PAAD samples into four groups by IR-neoAg load and immune checkpoint gene expression levels (median as the cutoffs, data not shown). Kaplan-Meier survival analysis revealed that patients with high IR-neoAg load tumors and low *PD-L1* gene expression (Fig. 24C) had the longest survival time compared to the other groups. Similar findings were observed for the combinations of high IR-neoAg load and low *PD-1* (Fig. 24D) or low *CTLA-4* (Fig. 24E) gene expression. Additional Kaplan-Meier survival curves

comparing these four patient groups stratified by IR-neoAg load and expression of other inhibitory checkpoint genes are shown in Fig. 25. Taken together, our results suggest that although most pancreatic cancers have an immune-privileged phenotype, a subset of patients with high tumor IR-neoAg and low expression of co-inhibitory genes may be able to generate a spontaneous antitumor immune response that could potentially restrain tumor progression and increase patient survival. Further analysis showed that all eight PanNET patients in the TCGA-PAAD cohort fell into this category and were amongst the patients with the longest OS.

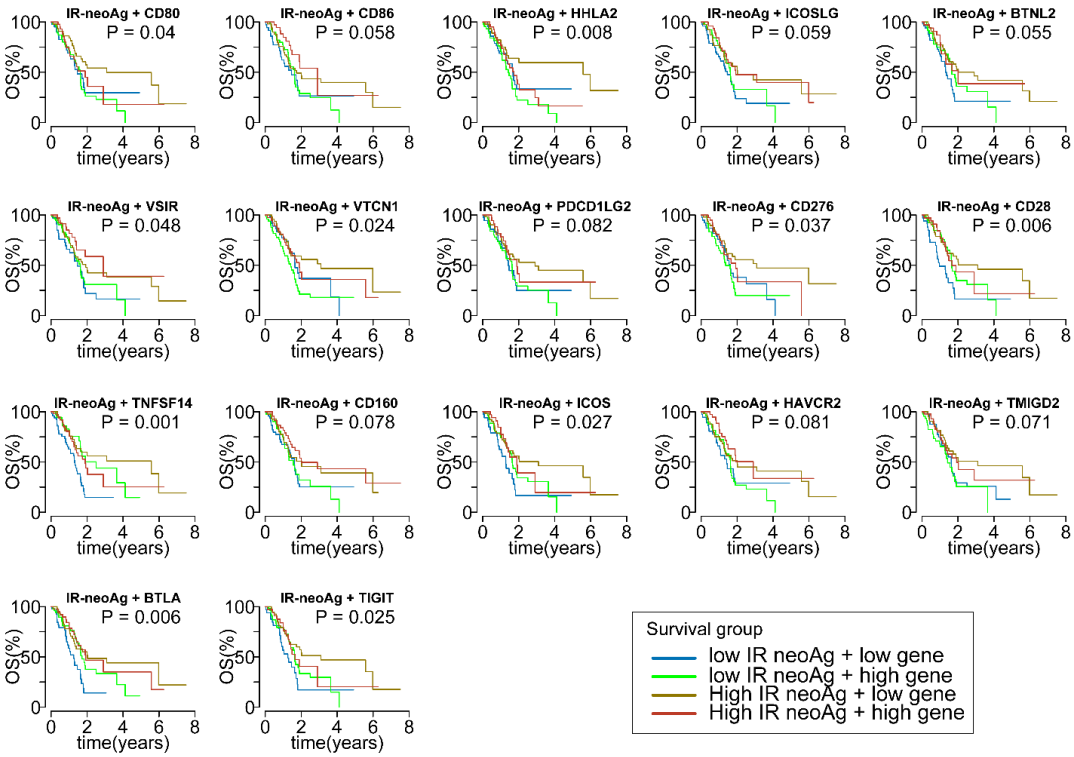

Figure 25 Kaplan-Meier survival curves of overall survival among patient groups stratified by the IR-neoAg and co-inhibitory checkpoint genes.

5.3.5 IR-neoAg load and HLA-I expression identify a subgroup of tumors that have similar gene expression patterns as tumors that respond to ICB therapy

Tumor neoantigens are critical mediators of host immune response and immunotherapy treatment efficacy. In addition, MHC class I expression is essential for neoantigen presentation. Because ICB therapy has not been used routinely for PDAC patients, we used SubMap, a subclass mapping method, to determine whether TCGA-PAAD tumors with high IR-neoAg load and high *HLA-I* gene expression levels shared similar transcriptomic profiles with ICB therapy-responsive melanoma tumors. We calculated an HLA-I score for each TCGA-PAAD tumor sample by averaging the expression value of the three major *HLA* genes, *HLA-A*/*B*/C. We then divided the samples into two groups based on the number of IR-neoAgs and the HLA-I score. One group contained samples with higher than the median IR-neoAg and HLA-I score, denoted as IR-neoAg^{hi}/HLA-I^{hi} (n = 43) and the other group consisted of all other tumors (n = 135). Notably, SubMap analysis indicated that the gene expression profiles of the IR-neoAg^{hi}/HLA-I^{hi} group showed significant similarity with the subset of melanoma tumors that were responsive to anti-PD1 immunotherapy (Fig. 26A, Bonferroni corrected P = 0.028). However, no significant similarities were observed between the TCGA-PAAD samples with IR-neoAg^{hi} alone or combinations of IRneoAg^{hi} with *PD-1*, *PD-L1*, or *CTLA-4* expression and the anti-PD1 responsive melanoma samples (data not shown). Notably, 93% of the IR-neoAg^{hi}/HLA-I^{hi} tumors (40/43) were classified as PDAC; three tumors were excluded from the curated 150 TCGA-PAAD set, which included two samples with low neoplastic cellularity and one tumor that did not arise from the pancreas. We also found that CD8+ T cells, as well as the total number of tumor-infiltrating immune cells, were

both significantly higher in IR-neoAg^{hi}/HLA-I^{hi} samples compared with the other tumors (Fig. 26B, Wilcoxon test, $p < 0.001$). Furthermore, this set of IR-neoAg^{hi}/HLA-I^{hi} tumors had significantly higher expression of *PD-1* and *PD-L1* genes (Fig. 26C), which likely reflects the higher proportion of CD8+ T cells and other cytotoxic immune cells. Together, these results indicate that pancreatic cancer patients with IR-neoAg^{hi}/HLA-I^{hi} tumors may represent a group that is more likely to respond to anti-PD1 treatment.

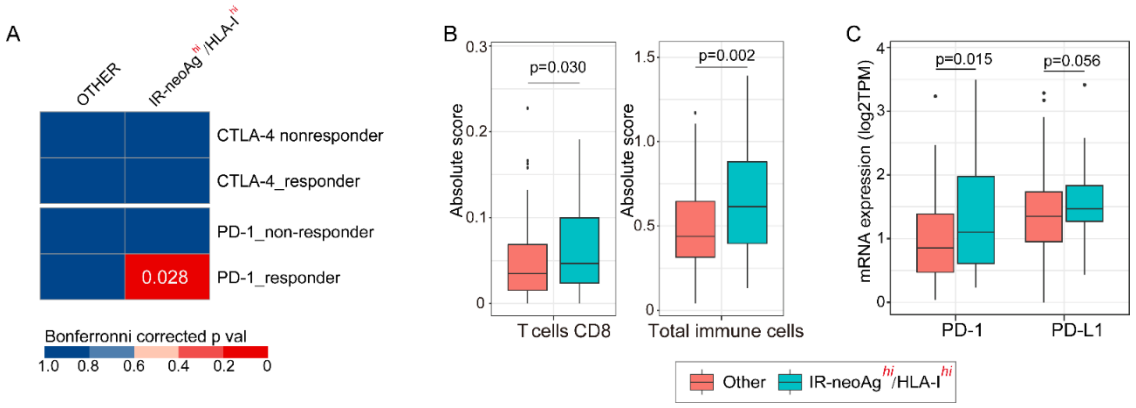

Figure 26 High IR-neoAg and high HLA class-I expression identify pancreatic cancers with similarities to tumors responsive to immune checkpoint blockade therapy. **A** Submap analysis comparing two groups from the TCGA-PAAD cohort [IR-neoAg^{hi}/HLA-I^{hi} (n = 43), and all Others (n = 135)] and four groups from an ICB-treated melanoma patient dataset [anti–PD-1 responders (n = 12), anti-PD1 non-responders (n = 14), anti-CTLA-4 responders (n = 8) and anti-CTLA-4 non-responders (n = 28)]. Bonferroni corrected p values were calculated in the SubMap program. **B** Box plot representation of the proportion of CD8+ T cells and total tumor infiltrated immune cells between tumors with IR-neoAg^{hi}/HLA-I^{hi} and all other tumors in the TCGA-PAAD dataset. Absolute score is estimated as the median expression level of all genes in the signature matrix divided by the median expression level of all genes in the mixture and is used to scale the relative cell fractions to absolute abundances. **C** Box plot representation of the tumor expression levels of immune checkpoint genes (PD-1 and PD-L1) in TCGA-PAAD tumors with IR-neoAg^{hi}/HLA-I^{hi} and all others. Box plots represent median and interquartile range (IQR). Upper whisker extends to the third quartile plus 1.5x IQR. Lower whisker extends to the first quartile minus 1.5x IQR. P values were determined by Wilcoxon test.

5.4 Discussion

We have shown that tumor-specific IR-neoAg load is an independent predictor of OS in pancreatic cancer patients. In addition, we found that the subset of tumors with the combination of high IR-neoAg load and high *HLA-I* gene expression had transcriptome profiles with significant similarities to melanoma tumors that were responsive to anti-PD1 therapy. This subset of IR-neoAg^{hi}/HLA^{hi} tumors showed higher numbers of tumor infiltrating immune cells, including CD8+ T cells. Collectively, our findings suggest that IR-neoAg load identifies patients with better prognosis, and together with *HLA-I* expression levels, could be a useful biomarker for selecting patients who may benefit from ICB therapy.

TMB has emerged as a biomarker of response to ICB therapy because it is a source of tumor-specific neoantigens that are targets of activated immune cells. TMB and neoantigen load have been shown to correlate with patient response to ICB therapy in several cancer types [193-195]. However, pancreatic cancer is characterized by a low TMB that frequently does not meet the threshold defined by clinical trials for ICB benefit [196-199]. Like TMB, intron retention is also a source of tumor neoantigens that can be presented by MHC-I [4, 130]. Thus, in cancers with low TMB, IR-neoAgs could be a potentially important biomarker for selecting patients who might benefit from ICB therapy. Importantly, additional factors other than neoantigens influence the ability of T cells to recognize and kill tumor cells. For example, the tumor microenvironment (TME) is an important factor in the poor responsiveness of PDAC to ICB therapy. PDAC TME is highly desmoplastic due to the presence of cancer-associated fibroblasts and a dense extracellular matrix [200], which impedes drug delivery. Because most PDAC patients present with

advanced disease, combination therapies will undoubtedly be necessary for overcoming resistance and improving immunotherapy strategies. We provide evidence that IR-neoAgs may aid in advancing these efforts by providing a new tool for selecting patients for participation in future clinical trials.

Chapter 6 Preliminary Research on Intron Retention in Cell Immunity of Neurodegeneration Disease

6.1 Introduction

Neurodegenerative diseases are characterized by progressive loss of neurons in the central nervous system, which is a relevant risk factor for the burden of mortality and morbidity [201]. Alzheimer's disease (AD) is an age-dependent disease characterized by the presence of amyloid plaques, neurofibrillary tangles, neuronal death, synaptic loss, and neuroinflammation in the brain [202]. Alcohol use disorder(AUD) is another common neuronal disease globally [203]. Alcoholic liver disease (ALD) and AUD are primary disease outcomes caused by alcohol use [204].

Neuroinflammation is one of the hallmarks of neurodegenerative diseases [205]. Genome-wide association studies (GWAS) recently highlighted a critical role for neuroinflammation in AD development. There are also emerging clues indicating that innate and adaptive immune responses play fundamental roles in neurodegeneration pathology. T cells populations were reported to increase in individuals with mild cognitive impairment and AD [206]. However, the nature of crosstalk between brain disease and the immune system remains unclear.

Abnormal RNA splicing has been shown to play a crucial role in neurodegenerative disease [207], especially AD. Also, according to a survey by the GTEx Consortium [156], brain tissues hold the highest number of splicing events (Fig. 27).

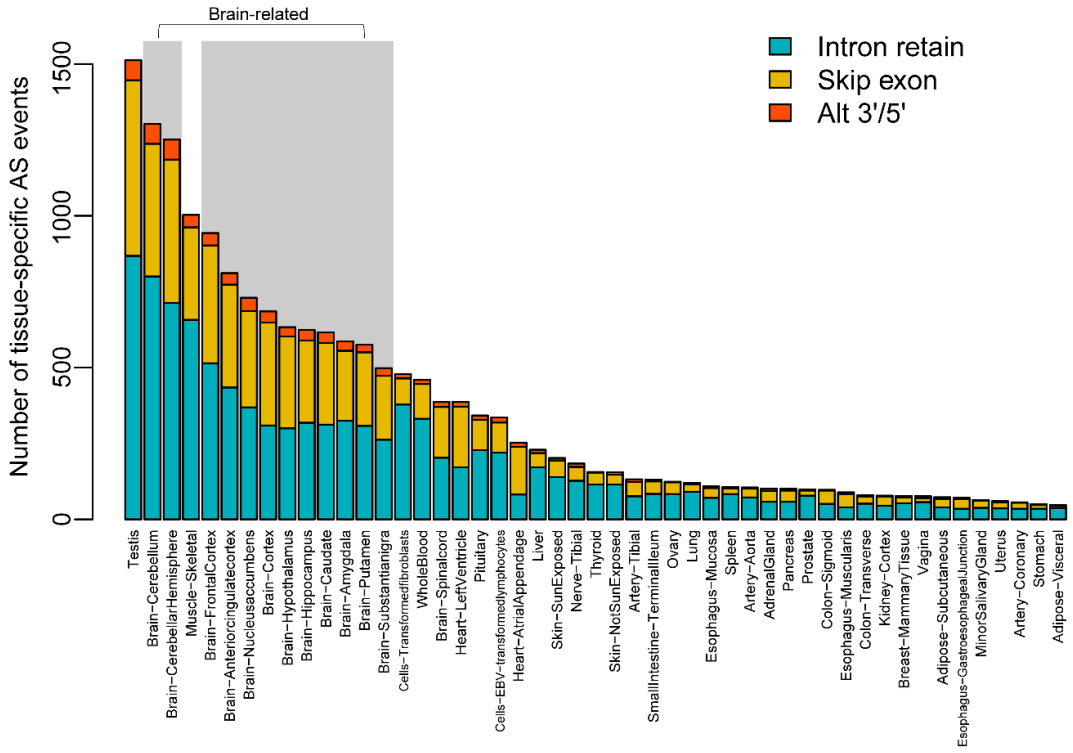

Figure 27 The splicing event abundance across tissues in GTEx. Data source: https://doi.org/10.1101/311563 [156]. GTEx, the Genotype-Tissue Expression project.

Intron retention (IR), one important type of alternative splicing, occurs when the splicing complex fails to remove introns from the primary messenger RNA transcript. Previous studies have demonstrated that IR is associated with AD pathology [208] and other neuron disorder diseases [209]; however, the mechanism of IR in mediating AD and AUD is still unclear.

In this section, we hypothesize that increased IR in neuronal cells might introduce immunogenetic peptide antigen or trigger anti-virus-like effect via forming double-strand RNA (dsRNA), which can be potential mechanisms of neurodegenerative diseases. We will perform a comprehensive bioinformatics analysis using existing human and mouse tissue-based RNA-seq data from the Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD) consortium. Our results indicate that IR was significantly increased in several brain regions in late-onset AD, and was positively correlated with amyloid plaque density. Furthermore, we also observed increasing IR levels are associated with AUD status in brain and liver tissues. The finding of this research suggests targeting aberrant splicing inducing immune response may serve as a potential therapeutic strategy to prevent neurodegenerative diseases.

6.2 Materials and Methods

6.2.1 RNA-Seq Datasets

AD brain tissues. Gene expression RNA-Seq data were downloaded from AMP-AD Consortium. We will focus on our work with two studies: the Mount Sinai Brain Bank (MSBB) study [210] and the Religious Order Study and the Memory and Aging Project (ROSMAP) study [211]. The respective clinical information, including APOE4 status, was also retrieved for these cohorts.

AUD brain tissues. Human frontal cortex brain samples were obtained from the Collaborative Studies on Genetics of Alcoholism Study (COGA) consortium with granted permission. Briefly, there were 47 samples from healthy control and 36 patients with detailed evidence of alcohol use disorder history (age ranged from 22 to 82, median age 54; gender, 16 Female and 67 Male).

Alcoholic liver tissues. RNA-Seq data for alcoholic hepatitis patients were retrieved from GEO via access GSE143318 [212], and gene expression was quantified following a standard RNA-Seq pipeline.

AUD rat brain tissues. RNA-Seq data of alcohol dependence rat models were retrieved from GEO under accession GSE159136 [213] and were processed using a standard RNA-Seq pipeline.

6.2.2 Identify aberrant IR events and IR neo-peptides

Raw data in bam format was first downloaded and converted to fastq files. STAR 2.7.2a was used to realign the fastq to the human reference genome. The GTF file was downloaded from Gencode (GRCh38.p13, version 32) and was parsed by merging intersectional exons to get the union exon sets within a protein-coding gene. The intron regions were obtained from the complement of exonic CDS regions. We count unique mapped the reads of RNAseq that falls on individual intron region level using an adjusted HTseq method against the customed GTF annotation file. The IR events were further filtered with parameters such as minimum read count of the intron region, RPKM value, and ratio versus flanking exon regions, which enable us to identify IR with enough abundance and high confidence.

The intron retention derived peptides were generated using methods described in Chapter 3. The binding affinity between the HLA-I allele and intron-induced neo-peptide was quantified using NetMHCpan 4.1.

6.2.3 Gene sets enrichment analysis

Single-sample gene sets enrichment analysis (ssGSEA) was conducted to infer the pathways activity level per sample using 'gsva' package in R [139]. The pathways genes sets were downloaded from KEGG and GO databases.

6.3 Results

6.3.1 Intron retention increased in AD brains

Previous studies have reported that increasing IR is a post-transcriptional signature associated with progressive aging and AD. Herein, we quantified the IR number for each sample in the MSBB study (see in methods) using RNA-Seq data. Our analysis shows that IR levels were significantly increased in AD samples (BM44, inferior frontal gyrus region), with a Wilcoxon test p-value reaching 0.001 (Fig. 28A). Furthermore, we found the IR levels positively correlated with plaque mean density (Fig. 28B, spearman correlation test, rho=0.34, $p < 0.0001$). APOE-ε4 is one of the strongest genetic risk factors for AD and is associated with an increase in the levels of amyloid deposition. We observed an increasing IR level in ROSMAP APOE ε4+ AD patients compared with APOE ε4- ones, although the difference has not reached statistical significance (Fig 29C). Increasing IR events might generate neo-peptide presenting MHC-I molecules to the neural cell surface and thus aggravate the AD pathological progression through inducing immune response and neuroinflammation. The preliminary result strongly suggested that IR could have an important role in AD pathology.

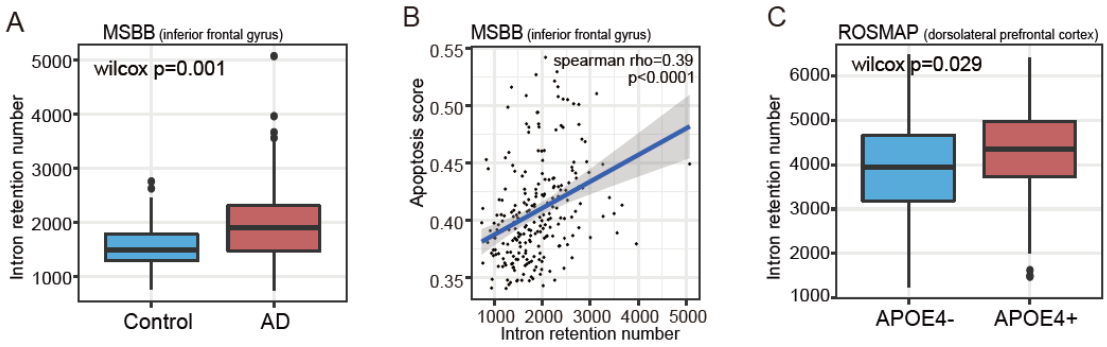

Figure 28 IR increased in AD brains. **A** IR events increased in AD brains (inferior frontal gyrus, denote as BM44) compared with healthy control in MSBB cohort; **B** The IR number was positively associated with apoptosis level in MSBB BM44 samples; **C** The IR level was increased in APOE4 positive samples compared with negative in AD brains in ROSMAP cohort.

6.3.2 Increase intron retention increased in alcoholic brain human and animal model

Alcoholic liver disease (ALD) and alcohol use disorders (AUDs) are important disease outcomes caused by alcohol use. Disruption of splicing on a somewhat broader scale was observed in the brain cortex of human fetuses exposed to alcohol [214]. Aberrant RNA splicing in the brain may lead to the modulation of protein functions, ultimately influencing alcohol dependence and neurotoxicity behaviors [215]. However, few studies have been conducted to elucidate the molecular role of splicing that contributes to AUD pathology.

We quantified the IR number for each human and rat brain sample from the COGA cohort and alcoholic-treated rat model dataset. Our analysis shows that IR levels were statistically significantly increased in AUD brain samples compared with normal brain tissues (Fig. 29 A, Wilcox p=0.021, adjust with age and gender p=0.046). Further analysis revealed that the IR levels were significantly correlated with apoptosis (Fig. 29B) and dsRNA binding protein enrichment scores (Fig. 29C). Similar observations were confirmed that higher IR number in the alcoholic feed rat compared with the control group (Fig. 29D). Moreover, we found the IR number was higher in RNA-Seq data of livers of patients with alcoholic hepatitis than healthy donors (Fig. 29-E) and was positively associated with cell apoptosis levels using data from Hyun et. al study (Fig. 29-F).

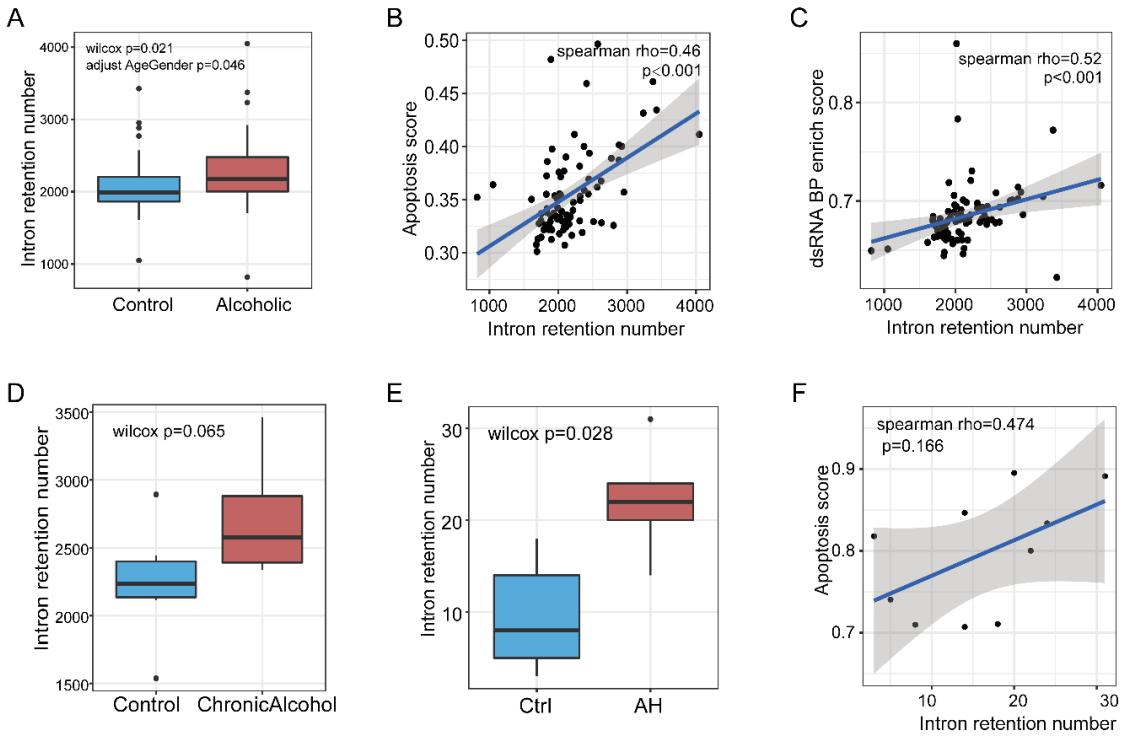

Figure 29 IR increased in alcohol use disorders. **A** IR numbers increased in COGA human brain tissues compared with controls; **B** The increased IR number was associated with apoptosis in COGA brain tissues; **C** IR number was associated with double-stand RNA binding protein levels in COGA brain tissues; **D** IR numbers increased in chronic alcoholic dependent rat brain tissues compared with control rat brains; **E** IR numbers increased in livers of patients with alcoholic hepatitis compared with health control in GSE143318; **F** IR number was associated with cell apoptosis in human liver tissues in GSE143318.

6.3.3 Maximum hypothetical IR neo-peptides across the human genome as an indicator for malignance susceptibility

The HLA genes play an important role in presenting internal and external antigens for the immune system [106]. Previous large genetic association studies have shown strong links between HLA genes in this region and risk for AD [216]. However, the underlying mechanism of how HLA alleles are associated with disease malignance remains unknown. As somatic mutation is rarely involved in neurodegeneration diseases, we sought to investigate the explore the antigenic peptides derived from intron retention induced peptides across the human genome. We tried to quantitively evaluate the IR induced neo-peptide potential of individual HLAs by considering an extreme case: if introns retention happens globally, how many hypothetical IR neo-peptides will be presented. As HLA alleles show different binding affinities for specific peptides, we choose the top HLA alleles in US European population. Totally, 6 HLA alleles with frequency large than 5% were selected (Fig. 30): 2 HLA-A alleles (HLA-A*01:01, HLA-A*02:01), 2 HLA-B alleles (HLA-B*07:02, HLA-B*08:01), and 2 HLA-C alleles (HLA-C*07:01, HLA-C*07:02). The 9-mer length neo-peptides were then sent to NetMHCpan 4.1 software for estimating the binding affinity with top HLA alleles. Peptides with a binding affinity rank score less than 0.5 were used as putative IR neo-peptides. The hypothetically maximum IR neo-peptide numbers for each allele are shown in Table 4. Interestingly, we found a trend of enrichment of IR neo-peptides on some chromosomal regions, such as chr19, chr17, chr16, and chr1. Notable, we found the HLA-C07:01 with the highest IR neo-peptide number, while HLA-A*01:01 with higher allele frequency (16.5%) but the corresponding binding IR neo-peptide number was lower to 15994.

Notably, the HLA alleles A*02:01, B*07:02 associated with AD risk in UCSF cohort and Italian population analysis [217-219], C*0702 is reported associated with AD in the Oxford population [220]. Those HLA prevalent alleles with high IR neo-peptide numbers were more prevalent in the AD group.

Table 4 Statistics of HLA genotype and AD susceptibility.

Locus	Allele	Frequency	IR neo-peptide	ROSMAP	MAYO	MSBB
A	A*01:01	0.165	15994	81	25	68
A	A*02:01	0.275	48124	96	39	52
B	B*07:02	0.131	63529	56	23	20
B	B*08:01	0.114	52512	57	19	62
C	C*07:01	0.16	67246	74	25	104
C	C*07:02	0.114	67105	59	22	24

ROSMAP, the Religious Orders Study and Memory and Aging Project study;

MAYO, the Mayo RNA-Seq study; MSBB, the Mount Sinai Brain Bank study.

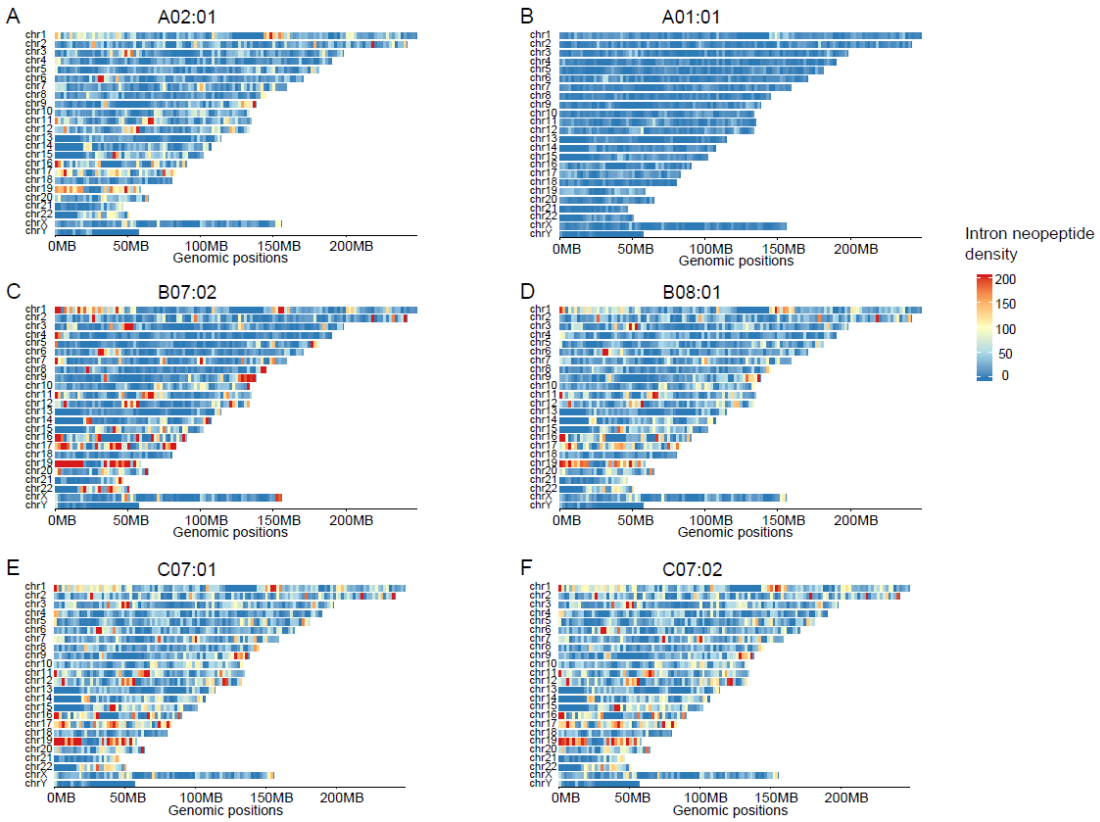

Fig 30 Intron neo-peptide potential and AD susceptibility. Atlas of intron induced neo-peptide across human genome that can be present by specific HLA alleles: **A** HLA-A02:01; **B** HLA-A01:01; **C** HLA-B07:02; **D** HLA-B08:01; **E** HLA-C07:01 and **F** HLA-C07:02. The red color represents a high density of IR neo-peptides, while blue represent low density of IR neo-peptides.

6.4 Discussion

In this section, we observed an increase in intron retention in AD brains compared with healthy control. The increase of IR events was positively associated with apoptosis in MSBB inferior frontal gyrus brain regions. In the alcohol use disorder study, we also found high IR levels were associated with alcohol dependence significantly in human brain and liver tissues. We validate our finding in alcohol drinking rat nucleus accumbent shell tissues. A further survey at whole-genome level revealed the IR neo-peptides was various across different type I HLA alleles. Our findings suggest that increasing IR levels are associated with neurodegenerative diseases, which indicates how personalized HLA contributes to disease burden.

Previous studies have reported no significant difference in somatic mutation numbers between noncancerous AD brains and healthy control. Dysregulation of RNA splicing through IR is frequent in brain tissue transcriptomes [208, 209]. The brain exhibits the highest alternative splicing rate of all tissues [156]. Increased IR has been reported to be a post-transcriptional signature associated with progressive aging and AD [209], which were recently reported as a potential source of neoantigen peptides [4]. The retained intron sequence can be translated into abnormal peptides and generate antigen-peptides presented by MHC-I molecules. Another study also reported the accumulated dsRNA formed when unspliced intron harbored in transcripts [13]. The dsRNA can further induce an innate immune anti-virus response in cancer cells. It indicates that IR, rather than genetic mutations, may play a crucial role in adaptive immune responses involved in inducing neuroinflammation in AD and other neurodegenerative diseases.

There is also accumulated evidence suggesting that HLA, a critical component of the immune response, partly contributes to the risk of AD pathogenesis [216, 219, 220]. However, the underlying mechanism of HLA that influences AD-related pathophysiological functions is still unclear. Our preliminary survey of global IR neo-peptides revealed some HLA alleles have more potential in presenting unspliced intron-induced neo-peptide than other HLA alleles, consistent with existing literature results. The difference in IR neo-peptide presenting ability may partly explain how the HLA genotype affects the susceptibility of immune-related diseases like AD and Type I diabetes.

Our study highlights the potential of IR events as a valuable tool for evaluating the pathology of neurodegenerative disease, including AD and AUD. The correlation of IR with apoptosis and dsRNA molecules illustrated a novel role in neuron and other cell immune inflammation. The findings from our study might help develop a novel class of biomarkers and new strategies for the treatment of neurodegenerative disease.

Chapter 7 Conclusions and Discussions

7.1 Conclusions

Despite generating unique mature transcripts with different functions, alternative splicing, especially intron retention (IR), has emerged a novel role by mediating adaptive and innate immune responses. In this dissertation, we developed a computational pipeline for prioritizing tumor neoantigens derived from intron retention. We demonstrated the IR-neoAgs measured from RNA-Seq could be validated from Mass Spectrum proteomics data. In Chapter 4, we analyzed the IR-neoAg load in 892 multiple myeloma samples from the MMRF study. We found a higher IR-neoAg load was associated with unfavorable OS, consistent with mutation-derived neoantigen. This section suggests that the IR-neoAg load could function as a prognosis biomarker and targeting the spliceosome might be a promising strategy for MM treatment. In Chapter 5, we demonstrated high IR-neoAg load could predict favorable OS in TCGA pancreatic cancer cohort. Further analysis revealed patients with both high IR-neoAg load and low expression of inhibitory checkpoint genes had the longest overall survival time. Moreover, we found the combination of high IR-neoAg load and HLA class-I gene expression may be useful in identifying PDAC patients who might benefit from immune checkpoint blockade therapy. The discovery of the cancer studies demonstrated that intron retention neoantigen provides meaningful information for aiding cancer immunotherapy.

Our preliminary research on neurodegenerative disease suggested intron retention was significantly associated with AD pathological characters. Because the analysis is based on bulk RNA-Seq data, it is unclear the increased IR levels are resulted from neuronal cells or supporting cells such as astrocytes. Although the definition of neo-antigen or antigen

peptide is mainly related to cancer studies, we assume the mechanism can be applied to other diseases. The intron neo-peptides provide a unique angle for elucidating the connections between HLA genotype, immune inflammation, and disease vulnerability.

7.2 Future Directions

Our long-term goal is to utilize the IR and IR-neoAg as an applicable biomarker for clinical usage. We hope to develop a more accurate algorithm for identifying immunogenic IR-neoAgs. Our collaborator has experimentally augmented intron retention by spliceosome inhibition, which allows us to measure the effect of intron retention mediated immune response. On one side, the HLA-enriched Mass Spectrum proteomics can help uncover the ground truth of real neoantigen, which will improve the algorithm accuracy for intron neoantigen prediction. In addition, we can further validate whether the IR-neoAgs can trigger T cell response and augment cancer immunotherapy. On another end, we hope to check whether the intron-derived dsRNA mediated antiviral effect could be applied to other cancer types, which may inspire a new treatment strategy by combination spliceosome inhibition and checkpoint blockade.

We expect the IR and IR derived antigen to statistically correlate with AD pathological progress in mouse AD models, independent from other attributes including age, gender, APOE risk, and other known factors. As the analysis was based on transcriptional expression level analysis, it is hard to prove whether the IR will function as a risk factor in pushing the brain to the irrevocable AD track. Further experiments will be needed to estimate IR levels during AD pathological development. We will systematically evaluate the correlation between HLA-specific IR neo-peptides load and disease susceptibility, including but not limited to neurodegenerative diseases.

Another open question we want to focus on is what causes the increase of unspliced introns in malignancies. It might be a consequence of epigenetic changes such as methylation in response to environmental stimulations. Further research on the connection between alternative splicing and immune effects could illuminate our understanding of the propounding effect of alternative splicing and inspire new treatment strategies by targeting splicing.

References

1. Noble JD, Balmant KM, Dervinis C, De Los Campos G, Resende Jr MFR, Kirst M, Barbazuk WB: The Genetic Regulation of Alternative Splicing in Populus deltoides. *Frontiers in Plant Science* 2020, 11:590.
2. Oh J, Pradella D, Shao C, Li H, Choi N, Ha J, Ruggiero S, Fu X-D, Zheng X, Ghigna C: Widespread alternative splicing changes in metastatic breast Cancer cells. *Cells* 2021, 10(4):858.
3. Sowalsky AG, Xia Z, Wang L, Zhao H, Chen S, Bubley GJ, Balk SP, Li W: Whole transcriptome sequencing reveals extensive unspliced mRNA in metastatic castration-resistant prostate cancer. *Molecular Cancer Research* 2015, 13(1):98-106.
4. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, Fugmann T, Wong K-K, Van Allen EM: Intron retention is a source of neoepitopes in cancer. *Nature biotechnology* 2018, 36(11):1056-1058.
5. Zappasodi R, Merghoub T, Wolchok JD: Emerging concepts for immune checkpoint blockade-based combination therapies. *Cancer cell* 2018, 33(4):581-598.
6. Ward JP, Gubin MM, Schreiber RD: The role of neoantigens in naturally occurring and therapeutically induced immune responses to cancer. *Advances in immunology* 2016, 130:25-74.
7. Jiang T, Shi T, Zhang H, Hu J, Song Y, Wei J, Ren S, Zhou C: Tumor neoantigens: from basic research to clinical applications. *Journal of hematology & oncology* 2019, 12(1):1-13.
8. Efremova M, Finotello F, Rieder D, Trajanoski Z: Neoantigens generated by individual mutations and their role in cancer immunity and immunotherapy. *Frontiers in immunology* 2017, 8:1679.
9. McGrail DJ, Pilié PG, Rashid NU, Voorwerk L, Slagter M, Kok M, Jonasch E, Khasraw M, Heimberger AB, Lim B: High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Annals of Oncology* 2021, 32(5):661-672.
10. Middleton R, Gao D, Thomas A, Singh B, Au A, Wong JJL, Bomane A, Cosson B, Eyras E, Rasko JEJ: IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome biology* 2017, 18(1):1-11.

11. Shiraishi Y, Kataoka K, Chiba K, Okada A, Kogure Y, Tanaka H, Ogawa S, Miyano S: A comprehensive characterization of cis-acting splicing-associated variants in human cancer. *Genome research* 2018, 28(8):1111-1125.
12. Wong JJL, Ritchie W, Ebner OA, Selbach M, Wong JWH, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K: Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* 2013, 154(3):583-595.
13. Bowling EA, Wang JH, Gong F, Wu W, Neill NJ, Kim IS, Tyagi S, Orellana M, Kurley SJ, Dominguez-Vidaña R: Spliceosome-targeted therapies trigger an antiviral immune response in triple-negative breast cancer. *Cell* 2021, 184(2):384-403.
14. Dvinge H, Bradley RK: Widespread intron retention diversifies most cancer transcriptomes. *Genome medicine* 2015, 7(1):1-13.
15. Zhang Z, Zhou C, Tang L, Gong Y, Wei Z, Zhang G, Wang F, Liu Q, Yu J: ASNEO: Identification of personalized alternative splicing based neoantigens with RNA-seq. *Aging (Albany NY)* 2020, 12(14):14633-14648.
16. Kiyotani K, Toyoshima Y, Nakamura Y: Personalized immunotherapy in cancer precision medicine. *Cancer Biology & Medicine* 2021, 18(4):955.
17. Bauer MA, Ashby C, Wardell C, Boyle EM, Ortiz M, Flynt E, Thakurta A, Morgan G, Walker BA: Differential RNA splicing as a potentially important driver mechanism in multiple myeloma. *haematologica* 2021, 106(3):736.
18. Frankiw L, Baltimore D, Li G: Alternative mRNA splicing in cancer immunotherapy. *Nature Reviews Immunology* 2019, 19(11):675-687.
19. Hanahan D, Weinberg Robert A: Hallmarks of Cancer: The Next Generation. *Cell* 2011, 144(5):646-674.
20. Gonzalez H, Hagerling C, Werb Z: Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes & development* 2018, 32(19-20):1267-1284.
21. Zhang Z, Lu M, Qin Y, Gao W, Tao L, Su W, Zhong J: Neoantigen: A New Breakthrough in Tumor Immunotherapy. *Frontiers in Immunology* 2021, 12.
22. Schumacher Ton N, Schreiber Robert D: Neoantigens in cancer immunotherapy. *Science* 2015, 348(6230):69-74.

23. Akinleye A, Rasool Z: Immune checkpoint inhibitors of PD-L1 as cancer therapeutics. *Journal of Hematology & Oncology* 2019, 12(1):92.
24. Vormehr M, Türeci Ö, Sahin U: Harnessing Tumor Mutations for Truly Individualized Cancer Vaccines. *Annual Review of Medicine* 2019, 70(1):395-407.
25. Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, Oliveira G, Giobbie-Hurder A, Felt K, Gjini E *et al*: Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* 2019, 565(7738):234-239.
26. Peng M, Mo Y, Wang Y, Wu P, Zhang Y, Xiong F, Guo C, Wu X, Li Y, Li X *et al*: Neoantigen vaccine: an emerging tumor immunotherapy. *Molecular Cancer* 2019, 18(1):128.
27. Jiang T, Shi T, Zhang H, Hu J, Song Y, Wei J, Ren S, Zhou C: Tumor neoantigens: from basic research to clinical applications. *Journal of Hematology & Oncology* 2019, 12(1):93.
28. Kast F, Klein C, Umaña P, Gros A, Gasser S: Advances in identification and selection of personalized neoantigen/T-cell pairs for autologous adoptive T cell therapies. *OncoImmunology* 2021, 10(1):1869389.
29. Ghorani E, Rosenthal R, McGranahan N, Reading JL, Lynch M, Peggs KS, Swanton C, Quezada SA: Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Annals of Oncology* 2018, 29(1):271-279.
30. Rizvi Naiyer A, Hellmann Matthew D, Snyder A, Kvistborg P, Makarov V, Havel Jonathan J, Lee W, Yuan J, Wong P, Ho Teresa S *et al*: Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* 2015, 348(6230):124-128.
31. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, Sucker A, Hillen U, Geukes Foppen MH, Goldinger SM: Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 2015, 350(6257):207-211.
32. Lee C-H, Yelensky R, Jooss K, Chan TA: Update on Tumor Neoantigens and Their Utility: Why It Is Good to Be Different. *Trends in Immunology* 2018, 39(7):536-548.

33. Yakirevich E, Patel NR: Tumor mutational burden and immune signatures interplay in renal cell carcinoma. *Annals of Translational Medicine; Vol 8, No 6 (March 2020): Annals of Translational Medicine* 2020.
34. Bräunlein E, Krackhardt AM: Identification and Characterization of Neoantigens As Well As Respective Immune Responses in Cancer Patients. *Frontiers in Immunology* 2017, 8.
35. Wang X, Guo G, Guan H, Yu Y, Lu J, Yu J: Challenges and potential of PD-1/PD-L1 checkpoint blockade immunotherapy for glioblastoma. *Journal of Experimental & Clinical Cancer Research* 2019, 38(1):87.
36. Łuksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, Solovyov A, Rizvi NA, Merghoub T, Levine AJ, Chan TA: A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* 2017, 551(7681):517-520.
37. Li J, Wang Y, Rao X, Wang Y, Feng W, Liang H, Liu Y: Roles of alternative splicing in modulating transcriptional regulation. *BMC systems biology* 2017, 11(5):1-12.
38. Baralle FE, Giudice J: Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology* 2017, 18(7):437-451.
39. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J *et al*: GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* 2019, 47(D1):D766-D773.
40. Shen S, Park JW, Lu Z-x, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y: rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences* 2014, 111(51):E5593-E5601.
41. Kim HK, Pham MHC, Ko KS, Rhee BD, Han J: Alternative splicing isoforms in health and disease. *Pflügers Archiv-European Journal of Physiology* 2018, 470(7):995-1016.
42. Li Y, Sahni N, Pancsa R, McGrail DJ, Xu J, Hua X, Coulombe-Huntington J, Ryan M, Tychhon B, Sudhakar D: Revealing the determinants of widespread alternative splicing perturbation in cancer. *Cell reports* 2017, 21(3):798-812.
43. Groulx J-F, Boudjadi S, Beaulieu J-F: MYC Regulates α6 Integrin Subunit Expression and Splicing Under Its Pro-Proliferative ITGA6A Form in Colorectal Cancer Cells. *Cancers* 2018, 10(2).

44. Zhang Y, Qian J, Gu C, Yang Y: Alternative splicing and cancer: A systematic review. *Signal transduction and targeted therapy* 2021, 6(1):1-14.
45. Zhang Y, Qian J, Gu C, Yang Y: Alternative splicing and cancer: a systematic review. *Signal Transduction and Targeted Therapy* 2021, 6(1):78.
46. Xie R, Chen X, Chen Z, Huang M, Dong W, Gu P, Zhang J, Zhou Q, Dong W, Han J *et al*: Polypyrimidine tract binding protein 1 promotes lymphatic metastasis and proliferation of bladder cancer via alternative splicing of MEIS2 and PKM. *Cancer Letters* 2019, 449:31-44.
47. Babic I, Anderson Erik S, Tanaka K, Guo D, Masui K, Li B, Zhu S, Gu Y, Villa Genaro R, Akhavan D *et al*: EGFR Mutation-Induced Alternative Splicing of Max Contributes to Growth of Glycolytic Tumors in Brain Cancer. *Cell Metabolism* 2013, 17(6):1000-1008.
48. Anczuków O, Rosenberg AZ, Akerman M, Das S, Zhan L, Karni R, Muthuswamy SK, Krainer AR: The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nature Structural & Molecular Biology* 2012, 19(2):220-228.
49. Zhou X, Li X, Cheng Y, Wu W, Xie Z, Xi Q, Han J, Wu G, Fang J, Feng Y: BCLAF1 and its splicing regulator SRSF10 regulate the tumorigenic potential of colon cancer cells. *Nature Communications* 2014, 5(1):4581.
50. Duriez M, Mandouri Y, Lekbaby B, Wang H, Schnuriger A, Redelsperger F, Guerrera CI, Lefevre M, Fauveau V, Ahodantin J *et al*: Alternative splicing of hepatitis B virus: A novel virus/host interaction altering liver immunity. *Journal of Hepatology* 2017, 67(4):687-699.
51. Sheng J, Zhao Q, Zhao J, Zhang W, Sun Y, Qin P, Lv Y, Bai L, Yang Q, Chen L *et al*: SRSF1 modulates PTPMT1 alternative splicing to regulate lung cancer cell radioresistance. *EBioMedicine* 2018, 38:113-126.
52. Love JE, Hayden EJ, Rohn TT: Alternative Splicing in Alzheimer's Disease. *J Parkinsons Dis Alzheimers Dis* 2015, 2(2):6.
53. Ingelfinger JR, Jarcho JA: Increase in the Incidence of Diabetes and Its Implications. *New England Journal of Medicine* 2017, 376(15):1473-1474.

54. Dlamini Z, Mokoena F, Hull R: Abnormalities in alternative splicing in diabetes: therapeutic targets. *Journal of Molecular Endocrinology* 2017, 59(2):R93-R107.
55. Moore MJ, Wang Q, Kennedy CJ, Silver PA: An Alternative Splicing Network Links Cell-Cycle Control to Apoptosis. *Cell* 2010, 142(4):625-636.
56. Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O, Sander C, Caesar-Johnson SJ: Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer cell* 2018, 34(2):211-224.
57. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 2012, 7(3):562-578.
58. Katz Y, Wang ET, Airoldi EM, Burge CB: Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* 2010, 7(12):1009-1015.
59. Shen S, Park Juw W, Lu Z-x, Lin L, Henry Michael D, Wu Ying N, Zhou Q, Xing Y: rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences* 2014, 111(51):E5593-E5601.
60. Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y: A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* 2016, 5:e11752.
61. Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E: Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* 2015, 21(9):1521-1531.
62. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK: Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics* 2018, 50(1):151-158.
63. Li L, Zhang X, Wang X, Kim SW, Herndon JM, Becker-Hapak MK, Carreno BM, Myers NB, Sturmoski MA, McLellan MD *et al*: Optimized polyepitope neoantigen DNA vaccines elicit neoantigen-specific immune responses in preclinical models and in clinical translation. *Genome Medicine* 2021, 13(1):56.

64. Orenbuch R, Filip I, Comito D, Shaman J, Pe'er I, Rabadan R: arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* 2020, 36(1):33-40.
65. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, Griffith M: pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Medicine* 2016, 8(1):11.
66. Bjerregaard A-M, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC: MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunology, Immunotherapy* 2017, 66(9):1123-1130.
67. Kim S, Kim HS, Kim E, Lee MG, Shin EC, Paik S, Kim S: Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology* 2018, 29(4):1030-1036.
68. Zhou C, Wei Z, Zhang Z, Zhang B, Zhu C, Chen K, Chuai G, Qu S, Xie L, Gao Y *et al*: pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. *Genome Medicine* 2019, 11(1):67.
69. Wang T-Y, Wang L, Alam SK, Hoeppner LH, Yang R: ScanNeo: identifying indel-derived neoantigens using RNA-Seq data. *Bioinformatics* 2019, 35(20):4159-4161.
70. Kracht MJL, Zaldumbide A, Roep BO: Neoantigens and microenvironment in type 1 diabetes: lessons from antitumor immunity. *Trends in Endocrinology & Metabolism* 2016, 27(6):353-362.
71. Kolb G, Eckle I, Heidtmann HH, Neurath F, Havemann K: Neoantigenic group on Fc fragments in rheumatoid arthritis synovial fluids. *Scandinavian Journal of Rheumatology* 1988, 17(sup75):179-189.
72. De Mattos-Arruda L, Blanco-Heredia J, Aguilar-Gurrieri C, Carrillo J, Blanco J: New emerging targets in cancer immunotherapy: the role of neoantigens. *ESMO open* 2019, 4:e000684.
73. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS: Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* 2015, 348(6230):124-128.
74. Van Allen Eliezer M, Miao D, Schilling B, Shukla Sachet A, Blank C, Zimmer L, Sucker A, Hillen U, Geukes Foppen Marnix H, Goldinger Simone M *et al*: Genomic

correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 2015, 350(6257):207-211.

75. Lazdun Y, Si H, Creasy T, Ranade K, Higgs BW, Streicher K, Durham NM: A New Pipeline to Predict and Confirm Tumor Neoantigens Predict Better Response to Immune Checkpoint Blockade. *Molecular Cancer Research* 2021, 19(3):498-506.
76. Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, Oliveira G, Giobbie-Hurder A, Felt K, Gjini E: Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* 2019, 565(7738):234-239.
77. Hu Z, Leet DE, Allesøe RL, Oliveira G, Li S, Luoma AM, Liu J, Forman J, Huang T, Iorgulescu JB: Personal neoantigen vaccines induce persistent memory T cell responses and epitope spreading in patients with melanoma. *Nature medicine* 2021, 27(3):515-525.
78. Liu Y, Gonzàlez-Porta M, Santos S, Brazma A, Marioni JC, Aebersold R, Venkitaraman AR, Wickramasinghe VO: Impact of alternative splicing on the human proteome. *Cell reports* 2017, 20(5):1229-1241.
79. Shen L, Zhang J, Lee H, Batista MT, Johnston SA: RNA transcription and splicing errors as a source of cancer frameshift neoantigens for vaccines. *Scientific reports* 2019, 9(1):1-13.
80. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013, 29(1):15-21.
81. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, 25(9):1105-1111.
82. Anders S, Pyl PT, Huber W: HTSeq—a Python framework to work with high-throughput sequencing data. *bioinformatics* 2015, 31(2):166-169.
83. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M: 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 research* 2020, 48(W1):W449-W454.
84. Chowell D, Krishna S, Becker PD, Cocita C, Shu J, Tan X, Greenberg PD, Klavinskis LS, Blattman JN, Anderson KS: TCR contact residue hydrophobicity is a hallmark of

immunogenic CD8+ T cell epitopes. *Proceedings of the National Academy of Sciences* 2015, 112(14):E1754-E1762.

85. Dintzis HM, Dintzis RZ, Vogelstein B: Molecular determinants of immunogenicity: the immunon model of immune response. *Proceedings of the National Academy of Sciences* 1976, 73(10):3671-3675.
86. Liu MKP, Hawkins N, Ritchie AJ, Ganusov VV, Whale V, Brackenridge S, Li H, Pavlicek JW, Cai F, Rose-Abrahams M: Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *The Journal of clinical investigation* 2012, 123(1).
87. Zeng J, Treutlein HR, Rudy GB: Predicting sequences and structures of MHC-binding peptides: a computational combinatorial approach. *Journal of Computer-Aided Molecular Design* 2001, 15(6):573-586.
88. Patronov A, Doytchinova I: T-cell epitope vaccine design by immunoinformatics. *Open Biology*, 3(1):120139.
89. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M: AAindex: amino acid index database, progress report 2008. *Nucleic acids research* 2007, 36(suppl_1):D202-D205.
90. Stranzl T, Larsen MV, Lundegaard C, Nielsen M: NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 2010, 62(6):357-368.
91. Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, Keşmir C, Peters B: Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS computational biology* 2013, 9(10):e1003266.
92. Tan X, Li D, Huang P, Jian X, Wan H, Wang G, Li Y, Ouyang J, Lin Y, Xie L: dbPepNeo: a manually curated database for human tumor neoantigen peptides. *Database* 2020, 2020.
93. Genomes Project C: A global reference for human genetic variation. *Nature* 2015, 526(7571):68.
94. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 2010, 38(16):e164-e164.

95. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, Hodi FS, Martín-Algarra S, Mandal R, Sharfman WH: Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell* 2017, 171(4):934-949.
96. Hugo W, Zaretsky JM, Sun LU, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G: Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* 2016, 165(1):35-44.
97. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C: The sequence read archive. *Nucleic acids research* 2010, 39(suppl_1):D19-D21.
98. Hutter C, Zenklusen JC: The cancer genome atlas: creating lasting value beyond its data. *Cell* 2018, 173(2):283-285.
99. Consortium GT: The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015, 348(6235):648-660.
100. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The sequence alignment/map format and SAMtools. *Bioinformatics* 2009, 25(16):2078-2079.
101. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* 2017, 14(4):417-419.
102. Barsnes H, Vaudel M: SearchGUI: a highly adaptable common interface for proteomics search and de novo engines. *Journal of proteome research* 2018, 17(7):2552-2555.
103. Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, Sickmann A, Martens L, Barsnes H: PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature biotechnology* 2015, 33(1):22-24.
104. Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L: SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X! Tandem searches. *Proteomics* 2011, 11(5):996-999.
105. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA: Profiling tumor infiltrating immune cells with CIBERSORT. In: *Cancer systems biology*. Springer; 2018: 243-259.
106. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, Hartigan CR, Zhang W, Braun DA, Ligon KL *et al*: A large peptidome dataset improves HLA class I

epitope prediction across most of the human population. *Nature Biotechnology* 2020, 38(2):199-209.

107. Durgeau A, Virk Y, Corgnac S, Mami-Chouaib F: Recent Advances in Targeting CD8 T-Cell Immunity for More Effective Cancer Immunotherapy. *Frontiers in Immunology* 2018, 9.
108. Ladomersky E, Zhai L, Lauing KL, Bell A, Xu J, Kocherginsky M, Zhang B, Wu JD, Podojil JR, Platanias LC *et al*: Advanced Age Increases Immunosuppression in the Brain and Decreases Immunotherapeutic Efficacy in Subjects with Glioblastoma. *Clinical Cancer Research* 2020, 26(19):5232-5245.
109. Farkona S, Diamandis EP, Blasutig IM: Cancer immunotherapy: the beginning of the end of cancer? *BMC Medicine* 2016, 14(1):73.
110. McGrail DJ, Pilié PG, Rashid NU, Voorwerk L, Slagter M, Kok M, Jonasch E, Khasraw M, Heimberger AB, Lim B *et al*: High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Annals of Oncology* 2021, 32(5):661-672.
111. Cowan AJ, Allen C, Barac A, Basaleem H, Bensenor I, Curado MP, Foreman K, Gupta R, Harvey J, Hosgood HD: Global burden of multiple myeloma: a systematic analysis for the global burden of disease study 2016. *JAMA oncology* 2018, 4(9):1221-1227.
112. Mohamed A, Collins J, Jiang H, Molendijk J, Stoll T, Torta F, Wenk MR, Bird RJ, Marlton P, Mollee P: Concurrent lipidomics and proteomics on malignant plasma cells from multiple myeloma patients: Probing the lipid metabolome. *PLoS One* 2020, 15(1):e0227455.
113. Wei SC, Duffy CR, Allison JP: Fundamental mechanisms of immune checkpoint blockade therapy. *Cancer discovery* 2018, 8(9):1069-1086.
114. McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT: Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 2016, 351(6280):1463-1469.
115. Schumacher TN, Schreiber RD: Neoantigens in cancer immunotherapy. *Science* 2015, 348(6230):69-74.

116. Bailey P, Chang DK, Forget M-A, Lucas FAS, Alvarez HA, Haymaker C, Chattopadhyay C, Kim S-H, Ekmekcioglu S, Grimm EA: Exploiting the neoantigen landscape for immunotherapy of pancreatic ductal adenocarcinoma. *Scientific reports* 2016, 6(1):1-8.
117. Shien K, Papadimitrakopoulou VA, Wistuba II: Predictive biomarkers of response to PD-1/PD-L1 immune checkpoint inhibitors in non–small cell lung cancer. *Lung Cancer* 2016, 99:79-87.
118. Zhang J, Caruso FP, Sa JK, Justesen S, Nam D-H, Sims P, Ceccarelli M, Lasorella A, Iavarone A: The combination of neoantigen quality and T lymphocyte infiltrates identifies glioblastomas with the longest survival. *Communications biology* 2019, 2(1):1-10.
119. Ren Y, Cherukuri Y, Wickland DP, Sarangi V, Tian S, Carter JM, Mansfield AS, Block MS, Sherman ME, Knutson KL: HLA class-I and class-II restricted neoantigen loads predict overall survival in breast cancer. *Oncoimmunology* 2020, 9(1):1744947.
120. Miller A, Asmann Y, Cattaneo L, Braggio E, Keats J, Auclair D, Lonial S, Russell SJ, Stewart AK: High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. *Blood cancer journal* 2017, 7(9):e612-e612.
121. Perumal D, Imai N, Laganà A, Finnigan J, Melnekoff D, Leshchenko VV, Solovyov A, Madduri D, Chari A, Cho HJ: Mutation-derived neoantigen-specific T-cell responses in multiple myeloma. *Clinical Cancer Research* 2020, 26(2):450-464.
122. Hoyos LE, Abdel-Wahab O: Cancer-specific splicing changes and the potential for splicing-derived neoantigens. *Cancer Cell* 2018, 34(2):181-183.
123. Pan Y, Lee AH, Yang HT, Wang Y, Xu Y, Kadash-Edmondson KE, Phillips J, Champhekar A, Puig C, Ribas A: IRIS: Big data-informed discovery of cancer immunotherapy targets arising from pre-mRNA alternative splicing. *bioRxiv* 2019:843268.
124. Nilsen TW, Graveley BR: Expansion of the eukaryotic proteome by alternative splicing. *Nature* 2010, 463(7280):457-463.
125. Chabot B, Shkreta L: Defective control of pre–messenger RNA splicing in human disease. *Journal of Cell Biology* 2016, 212(1):13-27.

126. Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV: Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes & development* 2012, 26(11):1209-1223.
127. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ: Widespread intron retention in mammals functionally tunes transcriptomes. *Genome research* 2014, 24(11):1774-1786.
128. Jacob AG, Smith CWJ: Intron retention as a component of regulated gene expression programs. *Human genetics* 2017, 136(9):1043-1057.
129. Brogna S, Wen J: Nonsense-mediated mRNA decay (NMD) mechanisms. *Nature structural & molecular biology* 2009, 16(2):107-113.
130. Apcher S, Daskalogianni C, Lejeune F, Manoury B, Imhoos G, Heslop L, Fåhraeus R: Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation. *Proceedings of the National Academy of Sciences* 2011, 108(28):11572-11577.
131. Palumbo A, Avet-Loiseau H, Oliva S, Lokhorst HM, Goldschmidt H, Rosinol L, Richardson P, Caltagirone S, Lahuerta JJ, Facon T: Revised international staging system for multiple myeloma: a report from International Myeloma Working Group. *Journal of clinical oncology* 2015, 33(26):2863.
132. Walker BA, Mavrommatis K, Wardell CP, Ashby TC, Bauer M, Davies F, Rosenthal A, Wang H, Qu P, Hoering A: A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia* 2019, 33(1):159-170.
133. Teoh PJ, An O, Chung T-H, Chooi JY, Toh SHM, Fan S, Wang W, Koh BTH, Fullwood MJ, Ooi MG: Aberrant hyperediting of the myeloma transcriptome by ADAR1 confers oncogenicity and is a marker of poor prognosis. *Blood, The Journal of the American Society of Hematology* 2018, 132(12):1304-1317.
134. Agirre X, Meydan C, Jiang Y, Garate L, Doane AS, Li Z, Verma A, Paiva B, Martín-Subero JI, Elemento O: Long non-coding RNAs discriminate the stages and gene regulatory states of human humoral immune response. *Nature communications* 2019, 10(1):1-16.

135. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012, 483(7391):603-607.
136. Ritchie ME, Phipson B, Wu DI, Hu Y, Law CW, Shi W, Smyth GK: limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 2015, 43(7):e47-e47.
137. Sedgwick P: Multiple significance tests: the Bonferroni correction. *Bmj* 2012, 344.
138. Yu G, Wang L-G, Han Y, He Q-Y: clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* 2012, 16(5):284-287.
139. Hänzelmann S, Castelo R, Guinney J: GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 2013, 14(1):7.
140. Stablein DM, Carter Jr WH, Novak JW: Analysis of survival data with nonproportional hazard functions. *Controlled clinical trials* 1981, 2(2):149-159.
141. Ahmad IA: A class of mann—whitney—wilcoxon type statistics. *The American Statistician* 1996, 50(4):324-327.
142. Greipp PR, Miguel JS, Durie BGM, Crowley JJ, Barlogie B, Bladé J, Boccadoro M, Child JA, Avet-Loiseau H, Kyle RA: International staging system for multiple myeloma. *Journal of clinical oncology* 2005, 23(15):3412-3420.
143. Chretien M-L, Corre J, Lauwers-Cances V, Magrangeas F, Cleynen A, Yon E, Hulin C, Leleu X, Orsini-Piocelle F, Blade J-S *et al*: Understanding the role of hyperdiploidy in myeloma prognosis: which trisomies really matter? *Blood* 2015, 126(25):2713-2719.
144. Zelle-Rieser C, Thangavadivel S, Biedermann R, Brunner A, Stoitzner P, Willenbacher E, Greil R, Jöhrer K: T cells in multiple myeloma display features of exhaustion and senescence at the tumor site. *Journal of hematology & oncology* 2016, 9(1):1-12.
145. Buchner M, Müschen M: Targeting the B cell receptor signaling pathway in B lymphoid malignancies. *Current opinion in hematology* 2014, 21(4):341.
146. Chen L: Co-inhibitory molecules of the B7–CD28 family in the control of T-cell immunity. *Nature Reviews Immunology* 2004, 4(5):336-347.

147. Dufva O, Pölönen P, Brück O, Keränen MAI, Klievink J, Mehtonen J, Huuhtanen J, Kumar A, Malani D, Siitonen S *et al*: Immunogenomic Landscape of Hematological Malignancies. *Cancer Cell* 2020, 38(3):380-399.e313.
148. Zhao R, Chinai Jordan M, Buhl S, Scandiuzzi L, Ray A, Jeon H, Ohaegbulam Kim C, Ghosh K, Zhao A, Scharff Matthew D *et al*: HHLA2 is a member of the B7 family and inhibits human CD4 and CD8 T-cell function. *Proceedings of the National Academy of Sciences* 2013, 110(24):9879-9884.
149. Lopes R, Caetano J, Ferreira B, Barahona F, Carneiro EA, João C: The Immune Microenvironment in Multiple Myeloma: Friend or Foe? *Cancers* 2021, 13(4).
150. Jensen MA, Wilkinson JE, Krainer AR: Splicing factor SRSF6 promotes hyperplasia of sensitized skin. *Nature Structural & Molecular Biology* 2014, 21(2):189-197.
151. Jiménez-Vacas JM, Herrero-Aguayo V, Montero-Hidalgo AJ, Gómez-Gómez E, Fuentes-Fayos AC, León-González AJ, Sáez-Martínez P, Alors-Pérez E, Pedraza-Arévalo S, González-Serrano T *et al*: Dysregulation of the splicing machinery is directly associated to aggressiveness of prostate cancer. *EBioMedicine* 2020, 51:102547.
152. Liu J, Huang B, Xiao Y, Xiong HM, Li J, Feng DQ, Chen XM, Zhang HB, Wang XZ: Aberrant Expression of Splicing Factors in Newly Diagnosed Acute Myeloid Leukemia. *Oncology Research and Treatment* 2012, 35(6):335-340.
153. Rahman MA, Krainer AR, Abdel-Wahab O: SnapShot: splicing alterations in cancer. *Cell* 2020, 180(1):208-208.
154. Zhang D, Hu Q, Liu X, Ji Y, Chao H-P, Liu Y, Tracz A, Kirk J, Buonamici S, Zhu P *et al*: Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. *Nature Communications* 2020, 11(1):2089.
155. Tan DJ, Mitra M, Chiu AM, Coller HA: Intron retention is a robust marker of intertumoral heterogeneity in pancreatic ductal adenocarcinoma. *NPJ genomic medicine* 2020, 5(1):1-17.
156. Yang RY, Quan J, Sodaei R, Aguet F, Segrè AV, Allen JA, Lanz TA, Reinhart V, Crawford M, Hasson S *et al*: A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. *bioRxiv* 2018:311563.

157. Merryman RW, Armand P, Wright KT, Rodig SJ: Checkpoint blockade in Hodgkin and non-Hodgkin lymphoma. *Blood Advances* 2017, 1(26):2643-2654.
158. Kawano Y, Zavidij O, Park J, Moschetta M, Kokubun K, Mouhieddine TH, Manier S, Mishima Y, Murakami N, Bustoros M *et al*: Blocking IFNAR1 inhibits multiple myeloma–driven Treg expansion and immunosuppression. *The Journal of Clinical Investigation* 2018, 128(6):2487-2499.
159. Gu SS, Zhang W, Wang X, Jiang P, Traugh N, Li Z, Meyer C, Stewig B, Xie Y, Bu X *et al*: Therapeutically Increasing MHC-I Expression Potentiates Immune Checkpoint Blockade. *Cancer Discovery* 2021, 11(6):1524-1541.
160. Lu SX, De Neef E, Thomas JD, Sabio E, Rousseau B, Gigoux M, Knorr DA, Greenbaum B, Elhanati Y, Hogg SJ *et al*: Pharmacologic modulation of RNA splicing enhances anti-tumor immunity. *Cell* 2021, 184(15):4032-4047.e4031.
161. Brahmer JR, Tykodi SS, Chow LQM, Hwu W-J, Topalian SL, Hwu P, Drake CG, Camacho LH, Kauh J, Odunsi K: Safety and activity of anti–PD-L1 antibody in patients with advanced cancer. *New England Journal of Medicine* 2012, 366(26):2455-2465.
162. Ji R-R, Chasalow SD, Wang L, Hamid O, Schmidt H, Cogswell J, Alaparthy S, Berman D, Jure-Kunkel M, Siemers NO: An immune-active tumor microenvironment favors clinical response to ipilimumab. *Cancer Immunology, Immunotherapy* 2012, 61(7):1019-1031.
163. Goodman AM, Castro A, Pyke RM, Okamura R, Kato S, Riviere P, Frampton G, Sokol E, Zhang X, Ball ED: MHC-I genotype and tumor mutational burden predict response to immunotherapy. *Genome medicine* 2020, 12(1):1-13.
164. Valero C, Lee M, Hoen D, Wang J, Nadeem Z, Patel N, Postow MA, Shoushtari AN, Plitas G, Balachandran VP: The association between tumor mutational burden and prognosis is dependent on treatment context. *Nature genetics* 2021, 53(1):11-15.
165. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, Barron DA, Zehir A, Jordan EJ, Omuro A: Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature genetics* 2019, 51(2):202-206.

166. Hendriks LE, Rouleau E, Besse B: Clinical utility of tumor mutational burden in patients with non-small cell lung cancer treated with immunotherapy. *Translational lung cancer research* 2018, 7(6):647.
167. Orth M, Metzger P, Gerum S, Mayerle J, Schneider G, Belka C, Schnurr M, Lauber K: Pancreatic ductal adenocarcinoma: Biological hallmarks, current status, and future perspectives of combined modality treatment approaches. *Radiation Oncology* 2019, 14(1):1-20.
168. Amin S, Baine MJ, Meza JL, Lin C: The Association of the Sequence of Immunotherapy With the Survival of Unresectable Pancreatic Adenocarcinoma Patients: A Retrospective Analysis of the National Cancer Database. *Frontiers in Oncology* 2020:1518.
169. Principe DR, Korc M, Kamath SD, Munshi HG, Rana A: Trials and tribulations of pancreatic cancer immunotherapy. *Cancer Letters* 2021, 504:1-14.
170. Maleki Vareki S: High and low mutational burden tumors versus immunologically hot and cold tumors and response to immune checkpoint inhibitors. *Journal for immunotherapy of cancer* 2018, 6(1):1-5.
171. Balachandran VP, Łuksza M, Zhao JN, Makarov V, Moral JA, Remark R, Herbst B, Askan G, Bhanot U, Senbabaoglu Y: Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* 2017, 551(7681):512-516.
172. Hegde S, Krisnawan VE, Herzog BH, Zuo C, Breden MA, Knolhoff BL, Hogg GD, Tang JP, Baer JM, Mpoy C: Dendritic cell paucity leads to dysfunctional immune surveillance in pancreatic cancer. *Cancer cell* 2020, 37(3):289-307.
173. Wang J, Dumartin L, Mafficini A, Ulug P, Sangaralingam A, Alamiry NA, Radon TP, Salvia R, Lawlor RT, Lemoine NR: Splice variants as novel targets in pancreatic ductal adenocarcinoma. *Scientific reports* 2017, 7(1):1-13.
174. Waddell N, Pajic M, Patch A-M, Chang DK, Kassahn KS, Bailey P, Johns AL, Miller D, Nones K, Quek K: Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* 2015, 518(7540):495-501.
175. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N: The genotype-tissue expression (GTEx) project. *Nature genetics* 2013, 45(6):580-585.

176. Badea L, Herlea V, Dima SO, Dumitrascu T, Popescu I: Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia-the authors reported a combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepato-gastroenterology* 2008, 55(88):2016.
177. Pei H, Li L, Fridley BL, Jenkins GD, Kalari KR, Lingle W, Petersen G, Lou Z, Wang L: FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. *Cancer cell* 2009, 16(3):259-266.
178. Zhang G, Schetter A, He P, Funamizu N, Gaedcke J, Ghadimi BM, Ried T, Hassan R, Yfantis HG, Lee DH: DPEP1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma. *PloS one* 2012, 7(2):e31507.
179. Yang S, He P, Wang J, Schetter A, Tang W, Funamizu N, Yanaga K, Uwagawa T, Satoskar AR, Gaedcke J: A novel MIF signaling pathway drives the malignant character of pancreatic cancer by targeting NR3C2. *Cancer research* 2016, 76(13):3838-3850.
180. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA: Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* 2015, 12(5):453-457.
181. Hoshida Y, Brunet J-P, Tamayo P, Golub TR, Mesirov JP: Subclass mapping: identifying common subtypes in independent disease data sets. *PloS one* 2007, 2(11):e1195.
182. Roh W, Chen P-L, Reuben A, Spencer CN, Prieto PA, Miller JP, Gopalakrishnan V, Wang F, Cooper ZA, Reddy SM: Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Science translational medicine* 2017, 9(379):eaah3560.
183. Shen R, Li P, Li B, Zhang B, Feng L, Cheng S: Identification of distinct immune subtypes in colorectal cancer based on the stromal compartment. *Frontiers in oncology* 2020:1497.

184. Goel MK, Khanna P, Kishore J: Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research* 2010, 1(4):274.
185. Dexter F: Wilcoxon-Mann-Whitney test used for data that are not normally distributed. In., vol. 117: LWW; 2013: 537-538.
186. Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, Robertson AG, Cherniack AD, Gupta M, Getz G: Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell* 2017, 32(2):185-203.
187. Lin X, Ye L, Wang X, Liao Z, Dong J, Yang Y, Zhang R, Li H, Li P, Ding L: Follicular Helper T Cells Remodel the Immune Microenvironment of Pancreatic Cancer via Secreting CXCL13 and IL-21. *Cancers* 2021, 13(15):3678.
188. Bruno TC: New predictors for immunotherapy responses sharpen our view of the tumour microenvironment. In.: Nature Publishing Group; 2020.
189. Sautès-Fridman C, Petitprez F, Calderaro J, Fridman WH: Tertiary lymphoid structures in the era of cancer immunotherapy. *Nature Reviews Cancer* 2019, 19(6):307-325.
190. Ahn S, Lee J-c, Shin DW, Kim J, Hwang J-H: High PD-L1 expression is associated with therapeutic response to pembrolizumab in patients with advanced biliary tract cancer. *Scientific reports* 2020, 10(1):1-8.
191. Patel SP, Kurzrock R: PD-L1 expression as a predictive biomarker in cancer immunotherapy. *Molecular cancer therapeutics* 2015, 14(4):847-856.
192. Vareki SM, Garrigós C, Duran I: Biomarkers of response to PD-1/PD-L1 inhibition. *Critical reviews in oncology/hematology* 2017, 116:116-124.
193. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS: Genetic basis for clinical response to CTLA-4 blockade in melanoma. *New England Journal of Medicine* 2014, 371(23):2189-2199.
194. van Rooij N, van Buuren MM, Philips D, Velds A, Toebes M, Heemskerk B, van Dijk LJA, Behjati S, Hilkmann H, El Atmioui D: Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 2013, 31(32).

195. Yarchoan M, Hopkins A, Jaffee EM: Tumor mutational burden and response rate to PD-1 inhibition. *The New England journal of medicine* 2017, 377(25):2500.
196. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L: Signatures of mutational processes in human cancer. *Nature* 2013, 500(7463):415-421.
197. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, Schrock A, Campbell B, Shlien A, Chmielecki J: Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome medicine* 2017, 9(1):1-14.
198. Chan TA, Yarchoan M, Jaffee E, Swanton C, Quezada SA, Stenzinger A, Peters S: Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Annals of Oncology* 2019, 30(1):44-56.
199. Stenzinger A, Allen JD, Maas J, Stewart MD, Merino DM, Wempe MM, Dietel M: Tumor mutational burden standardization initiatives: recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions. *Genes, Chromosomes and Cancer* 2019, 58(8):578-588.
200. Kleeff J, Korc M, Apte M, La Vecchia C, Johnson CD, Biankin AV, Neale RE, Tempero M, Tuveson DA, Hruban RH: Pancreatic cancer. *Nature reviews Disease primers* 2016, 2(1):1-22.
201. Erkkinen MG, Kim M-O, Geschwind MD: Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases. *Cold Spring Harbor perspectives in biology* 2018, 10(4):a033118.
202. DeTure MA, Dickson DW: The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration* 2019, 14(1):32.
203. Burchi E, Makris N, Lee MR, Pallanti S, Hollander E: Compulsivity in Alcohol Use Disorder and Obsessive Compulsive Disorder: Implications for Neuromodulation. *Frontiers in Behavioral Neuroscience* 2019, 13.
204. Yoo ER, Cholankeril G, Ahmed A: Treating Alcohol Use Disorder in Chronic Liver Disease. *Clinical Liver Disease* 2020, 15(2):77-80.
205. Kwon HS, Koh S-H: Neuroinflammation in neurodegenerative disorders: the roles of microglia and astrocytes. *Translational Neurodegeneration* 2020, 9(1):42.

206. Gate D, Saligrama N, Leventhal O, Yang AC, Unger MS, Middeldorp J, Chen K, Lehallier B, Channappa D, De Los Santos MB *et al*: Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. *Nature* 2020, 577(7790):399-404.
207. Liu EY, Cali CP, Lee EB: RNA metabolism in neurodegenerative disease. *Dis Model Mech* 2017, 10(5):509-518.
208. Ong C-T, Adusumalli S: Increased intron retention is linked to Alzheimer's disease. *Neural Regen Res* 2020, 15(2):259-260.
209. Adusumalli S, Ngian Z-K, Lin W-Q, Benoukraf T, Ong C-T: Increased intron retention is a post-transcriptional signature associated with progressive aging and Alzheimer's disease. *Aging Cell* 2019, 18(3):e12928.
210. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, Ming C, Neff R, Ma W, Fullard JF *et al*: The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Scientific Data* 2018, 5(1):180185.
211. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA: Religious Orders Study and Rush Memory and Aging Project. *J Alzheimers Dis* 2018, 64(s1):S161-S189.
212. Hyun J, Sun Z, Ahmadi AR, Bangru S, Chembazhi UV, Du K, Chen T, Tsukamoto H, Rusyn I, Kalsotra A *et al*: Epithelial splicing regulatory protein 2–mediated alternative splicing reprograms hepatocytes in severe alcoholic hepatitis. *The Journal of Clinical Investigation* 2020, 130(4):2129-2145.
213. Kisby BR, Farris SP, McManus MM, Varodayan FP, Roberto M, Harris RA, Ponomarev I: Alcohol Dependence in Rats Is Associated with Global Changes in Gene Expression in the Central Amygdala. *Brain Sciences* 2021, 11(9).
214. Van Booven D, Mengying L, Sunil Rao J, Blokhin IO, Dayne Mayfield R, Barbier E, Heilig M, Wahlestedt C: Alcohol use disorder causes global changes in splicing in the human brain. *Translational Psychiatry* 2021, 11(1):2.
215. Donadoni M, Cicalese S, Sarkar DK, Chang SL, Sariyer IK: Alcohol exposure alters pre-mRNA splicing of antiapoptotic Mcl-1L isoform and induces apoptosis in neural progenitors and immature neurons. *Cell Death & Disease* 2019, 10(6):447.

216. Wang Z-X, Wan Q, Xing A: HLA in Alzheimer's Disease: Genetic Association and Possible Pathogenic Roles. *NeuroMolecular Medicine* 2020, 22(4):464-473.
217. Steele NZR, Carr JS, Bonham LW, Geier EG, Damotte V, Miller ZA, Desikan RS, Boehme KL, Mukherjee S, Crane PK *et al*: Fine-mapping of the human leukocyte antigen locus as a risk factor for Alzheimer disease: A case–control study. *PLOS Medicine* 2017, 14(3):e1002272.
218. Ma SL, Tang NLS, Tam CWC, Lui VWC, Suen EWC, Chiu HFK, Lam LCW: Association between HLA-A Alleles and Alzheimer's Disease in a Southern Chinese Community. *Dementia and Geriatric Cognitive Disorders* 2008, 26(5):391-397.
219. Guerini FR, Tinelli C, Calabrese E, Agliardi C, Zanzottera M, De Silvestri A, Franceschi M, Grimaldi LME, Nemni R, Clerici M: HLA-A*01 is Associated with Late Onset of Alzheimer's Disease in Italian Patients. *International Journal of Immunopathology and Pharmacology* 2009, 22(4):991-999.
220. Lehmann DJ, Barnardo MCNM, Fuggle S, Quiroga I, Sutherland A, Warden DR, Barnetson L, Horton R, Beck S, Smith AD: Replication of the association of HLA-B7 with Alzheimer's disease: a role for homozygosity? *Journal of Neuroinflammation* 2006, 3(1):33.

Curriculum Vitae

Chuanpeng Dong

Education

2017 – 2022 Indiana University–Purdue University Indianapolis, Indianapolis, IN

Ph.D. in Informatics with Bioinformatics Specification

Minor: Cancer biology

2014 – 2017 Fudan University, Shanghai, China

M.S. in Biochemistry and Molecular Biology (bioinformatics)

2008 – 2012 Jilin University, Changchun, China

B.S. in Animal Science

Research Teaching and Work Experience

2017 – 2022 Indiana University School of Medicine, Medical & Molecular Genetics

Research Assistant in Dr. Yunlong Liu's Lab

2019 Fall Teaching assistant, Next Generation Sequencing(I-590)

2018 Summer HIM-M110 Computer Concepts for Health Information, Instructor

2014 – 2017 Fudan University, Shanghai Cancer Center & IBS

Research Assistant in Bioinformatics Dr. Lei Liu's Lab

2012 – 2014 Shanghai Hengrui Pharmaceuticals Co., LTD

Research Assistant in the Department of Pharmacology

Conference

2019 Jun International Conference on Intelligent Biology and Medicine (ICIBM 2019). Columbus, Ohio, USA

2019 Dec 62nd American Society of Hematology Annual Meeting and Exposition (ASH 2019). Orlando, Florida, USA

2021 Oct 2nd CCBB retreat, Indiana University School of Medicine. Indianapolis, Indiana, USA

2022 Jun 82nd American Diabetes Association Scientific Sessions. New Orleans, Louisiana, USA

Publications

Dong, C., Cesarano, A., Bombaci, G., Reiter, J. L., Yu, C. Y., Wang, Y., ... & Liu, Y. (2021). Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. Oncogene, 40(42), 6130-6138.

Dong, Chuanpeng, et al. "Intron-Retention Neoantigen Load Predicts Favorable Prognosis in Pancreatic Cancer." JCO Clinical Cancer Informatics 6 (2022): e2100124.

Tsai, A. P., Dong, C., Lin, P. B. C., Messenger, E. J., Casali, B. T., Moutinho, M., ... & Nho, K. (2022). PLCG2 is associated with the inflammatory response and is induced by amyloid plaques in Alzheimer's disease. Genome Medicine, 14(1), 1-13.

Johnson, T. S., Yu, C. Y., Huang, Z., Xu, S., Wang, T., Dong, C., ... & Zhang, J. (2022). Diagnostic Evidence GAuge of Single cells (DEGAS): a flexible deep transfer learning framework for prioritizing cells in relation to disease. Genome medicine, 14(1), 1-23.

Wu, W., Syed, F., Simpson, E., Lee, C. C., Liu, J., Chang, G., Dong, C,... & Evans-Molina, C. (2022). Impact of Proinflammatory Cytokines on Alternative Splicing Patterns in Human Islets. Diabetes, 71(1), 116-127.

Tsai, A. P. Y., Lin, P. B. C., Dong, C., Moutinho, M., Liu, Y., Bissel, S. J., ... & Landreth, G. E. (2021). Impact of PLCG2 expression on Microglial Biology and Disease Pathogenesis in Alzheimer's Disease. Alzheimer's & Dementia, 17, e058740.

Ma, B., Jiang, H., Luo, Y., Liao, T., Xu, W., Wang, X., ...Dong, C & Wang, Y. (2021). Tumor-Infiltrating Immune-Related Long Non-Coding RNAs Indicate Prognoses and Response to PD-1 Blockade in Head and Neck Squamous Cell Carcinoma. Frontiers in immunology, 12.

Liu, S., Sun, X., Li, K., Zha, R., Feng, Y., Sano, T., Dong, C.,... & Yokota, H. (2021). Generation of the tumor-suppressive secretome from tumor cells. Theranostics, 11(17), 8517.

Liu, J., Dong, C., Liu, Y., & Wu, H. (2021). CGPE: an integrated online server for C ancer G ene and P athway E xploration. Bioinformatics, 37(15), 2201-2202.

Tsai, A. P., Lin, P. B. C., Dong, C., Moutinho, M., Casali, B. T., Liu, Y., ... & Nho, K. (2021). INPP5D expression is associated with risk for Alzheimer's disease and induced by plaque-associated microglia. Neurobiology of disease, 153, 105303.

Tsai, Andy Po-Yi, et al. PLCG2 expression is associated with plaque-associated microglia in Alzheimer's disease. Alzheimer's & Dementia 17 (2021): e054755.

Dong, C., Liu, J., Chen, S. X., Dong, T., Jiang, G., Wang, Y., ... & Liu, Y. (2020). Highly robust model of transcription regulator activity predicts breast cancer overall survival. BMC medical genomics, 13(5), 1-10.

Huang, M., Kim, H. G., Zhong, X., Dong, C., Zhang, B., Fang, Z., ... & Dong, X. C. (2020). Sestrin 3 protects against diet-induced nonalcoholic steatohepatitis in mice through

suppression of transforming growth factor β signal transduction. Hepatology, 71(1), 76-92.

Zheng, S., Luo, X., Dong, C., Zheng, D., Xie, J., Zhuge, L., ... & Chen, H. (2018). A B7-CD28 family based signature demonstrates significantly different prognoses and tumor immune landscapes in lung adenocarcinoma. International journal of cancer, 143(10), 2592-2601.

Liu, G., Zhan, X., Dong, C., & Liu, L. (2017). Genomics alterations of metastatic and primary tissues across 15 cancer types. Scientific reports, 7(1), 1-9.

Hou, G., Dong, C., Dong, Z., Liu, G., Xu, H., Chen, L., ... & Zhou, W. (2017). Upregulate KIF4A enhances proliferation, invasion of hepatocellular carcinoma and indicates poor prognosis across human cancer types. Scientific reports, 7(1), 1-10.

Liu, G., Dong, C., & Liu, L. (2016). Integrated multiple "-omics" data reveal subtypes of hepatocellular carcinoma. PloS one, 11(11), e0165457.