Scientific Reports

https://doi.org/10.1038/s41598-026-43687-2

Article in Press

Identification of immunogenic neoantigens from intron retention in colorectal cancer

Received: 14 October 2025

Accepted: 5 March 2026

Published online: 09 March 2026

Cite this article as: Manoharan T., Kee B.B.R., Cheng C.Z.M. *et al.* Identification of immunogenic neoantigens from intron retention in colorectal cancer. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-43687-2

Thamizhanban Manoharan, Brandon Bing Rui Kee, Cyrus Zai Ming Cheng, Malcolm Kaiheng Choy, Bei En Siew, Wai-Kit Cheong, Kai-yin Lee, Ian Jse-Wei Tan, Bettina Lieske, Choon Kong Yap, Iain Bee Huat Tan, Ker-Kan Tan, Kar Tong Tan & Gloryn Chia

Identification of immunogenic neoantigens from Intron Retention in Colorectal Cancer

Thamizhanban Manoharan[1,2,6], Brandon Bing Rui Kee[2,6], Cyrus Zai Ming Cheng[1,2], Malcolm Choy Kaiheng[2], Bei En Siew[3,4], Wai-Kit Cheong[4], Kai-yin Lee[4], Ian Jse-Wei Tan[4], Bettina Lieske[4], Yap Choon Kong[5], Iain Bee Huat Tan[5], Ker-Kan Tan[3,4], Kar Tong Tan[2], Gloryn Chia[1,2*]

[1]Institute for Health Innovation & Technology (iHealthtech), National University of Singapore, Singapore

[2]Department of Pharmacy, Faculty of Science, National University of Singapore, Singapore

[3]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[4]Division of Colorectal Surgery, Department of Surgery, National University Hospital, Singapore

[5]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR)

[6]These authors contributed equally: Thamizhanban Manoharan, Brandon Kee Bing Rui

[*]Corresponding author. Email: phagcl@nus.edu.sg; Address: Block S9, Level 15, 4 Science Drive 2, Singapore, 117544

ABSTRACT

Intron retention (IR) is an underexplored source of alternative neoantigens in cancer. Unlike mutation-derived neoantigens, IR-derived neoantigens can arise even in malignancies with low tumour mutational burden (TMB), such as microsatellite-stable colorectal cancer (MSS-CRC), making them a potentially important source of tumour antigens. Here, we provide experimental evidence that IR-derived neoantigens are expressed in tumours and are recognized by T cells, with approximately 30% of predicted epitopes eliciting measurable $CD8^+$ T cell responses. We applied a bioinformatics pipeline to RNA sequencing data from 23 CRC patients, identifying 49 patient-specific and 24 shared IR-derived neoantigens, the latter present in ~30% of the cohort. Notably, most shared neoantigens exhibited high binding affinity to the predominant HLA alleles, highlighting their potential as broadly targetable vaccine candidates. Together, these findings establish IR-derived neoantigens as a validated and clinically relevant class of tumour antigens, expanding opportunities for personalized and off-the-shelf immunotherapies in CRC and other low-TMB cancers.

INTRODUCTION

Neoantigen cancer vaccines are highly personalized as neoantigens are predicted from patient's tumour samples. These neoantigens are defined as tumour specific proteins that are absent in normal tissue but exist in tumour cells[1]. By vaccinating the patients against their neoantigens, cancer vaccines aim to prime the adaptive immune system to target the tumours and boost antitumour immunity[2]. Neoantigens used for cancer vaccines generation are usually predicted from single nucleotide polymorphisms (SNP) such as Single Nucleotide Variant (SNV) mutations, insertions and deletions (indels). However, SNPs rarely give rise to immunogenic neoantigens and are highly patient-specific[1]. Moreover, low Tumour Mutational Burden (TMB) tumours, such as microsatellite stable colorectal cancer (MSS CRC), harbour few SNP-derived neoantigens, thereby limiting the effectiveness of SNP-based neoantigen vaccines[3].

Multiple studies have demonstrated that neoantigens can arise from diverse sources, including alternative splicing, human endogenous retroviruses, and gene fusions[4,5]. By exploring these alternative sources, the repertoire of targetable antigens can be expanded, providing opportunities to develop more effective cancer vaccines for tumours with low mutational burden, such as MSS-CRC[1]. Importantly, many of these alternative neoantigens can be upregulated across patients and can be shared among different tumour types, offering a broader and more universal set of targets than patient-specific SNP-derived neoantigens[6].

Intron retention (IR) is one such alternative source of neoantigens. IR occurs when intronic sequences are retained in mature mRNA after splicing, and in the cytoplasm these transcripts can escape nonsense-mediated decay (NMD). When these retained intron transcripts are translated, they can give rise to alternative protein isoforms not normally produced in healthy cells, thereby serving as a potential source of neoantigens. Under normal conditions, introns are excised from exonic regions through cleavage at 5′ splice sites[7]. Spliceosome complexes, composed of small nuclear ribonucleoproteins (snRNPs), recognize the 5′ splice site and form a lariat structure with the branch point and 3′ splice site, thereby excising introns. Pre-mRNA splicing is mediated by multiple factors, including the activity of the six snRNP-containing spliceosome complexes and their affinity for splice sites. Studies have shown that mutations in these splice sites disrupt normal splicing and increase intron retention[8,9].

Reports have shown that intron retention occurs at markedly higher rates in most cancer tissues as compared to the normal tissues[10]. Given the elevated frequency of intron retention (IR) in cancer cells, we hypothesized that this would lead to an increased repertoire of IR-derived peptides that could serve as alternative source of neoantigens. Thus far, multiple bioinformatic pipelines have been developed to detect retained introns and predict corresponding neoepitopes[11,12]. Using one such published pipeline, in this study, we have systematically identified and validated IR-derived neoantigens in colorectal cancer (CRC). By integrating RNA sequencing data from 23 CRC patients, we predicted both patient-specific and shared IR-derived neoantigens, confirmed their expression by RT-PCR, and evaluated their immunogenicity using in vitro immune assays. Our findings establish IR-derived neoantigens as a promising class of targets for cancer immunotherapy, particularly in low–tumour mutational burden (TMB) cancers such as CRC.

RESULTS

To assess the presence of tumour-specific IR events in colorectal cancer, we applied IRFinder to RNA-sequencing data from 23 paired tumour and normal CRC samples, identifying tumour-specific IR events through differential expression analysis. The binding affinity of these sequences to the HLA alleles expressed by the patients were predicted by netMHCpan4.1. 9-mer peptides translated from these sequences were subsequently queried against PepQuery 2.0. Peptides that have the

potential to be translated with high confidence in CRC datasets but not found in healthy tissue datasets were selected. These potentially translatable peptides were further screened to ensure that they were not part of already existing proteins, using NCBI tblastn (**Fig 1A**).

We observed that for each patient sample, there were more intron retention events in the tumour as compared to their matched normal (**Fig 1B**). On average, each tumour sample has 146 unique IR events that were not found in the matched-normal sample while the normal samples have 102 unique intron retention events (**Fig 1C**). Previous published studies also reported the identification of unique IR events in tumour samples. A significant number of IR events were upregulated in tumours, although also detectable in normal samples, and were classified as tumour-associated IR events (**Supplementary Fig 1A**). On average, there were a total of 5405 increased differential and 5059 decreased differential retention events in the tumour and normal, respectively.

Further analysis of these tumour-associated IR events showed that there are 2637 upregulated introns and 4395 downregulated introns among the patient population ($n = 23$, $p < 0.01$) (**Fig 1D**). This shows that differentially expressed IR events are common throughout the patient population. Analysis of the top 50 differentially expressed introns revealed distinct expression clustering between tumour and normal samples (**Fig 1E**).

To determine the functions of genes affected by IR, we performed gene ontology analysis on 695 genes containing 2,637 upregulated introns and

1,014 genes containing 4,395 downregulated introns. Gene ontology analysis revealed that genes affected by upregulated introns were predominantly involved in extracellular matrix organization, cell division, cell cycle checkpoints, and DNA replication (**Fig 2A**) ($p < 0.05$). The increase in intron retention within these processes suggests reduced levels of functional mRNA and, consequently, fewer proteins necessary for their execution, indicating that these pathways are impaired in tumour cells[13]. Such disruption may represent one mechanism enabling uncontrolled tumour proliferation. To assess disease associations of the upregulated introns, we repeated the analysis using disease ontology and found that colon adenocarcinoma was among the eleven identified diseases **(Fig 2B)**, indicating that a substantial proportion of upregulated introns are correlated with CRC. On the other hand, the gene ontology for the downregulated introns has an overrepresentation of processes involved with the skeletal muscle (**Fig 2C**). Consistently, most of the genes with the downregulated introns seem to be involved in diseased such as muscular dystrophy (**Fig 2D**). Thus, while downregulated introns appear less directly relevant to tumour progression, the upregulation of introns shows a clear association with CRC.

Multiple studies have hypothesized that the splice site strength of an intron might be the cause of differential expression in different genes[14]. To test this, the splice strength of all introns was calculated using MaxEntScan[15] and plotted accordingly (**Supplementary Fig 1B**). Mann-Whitney U test was done to determine if there were any significant

differences between the two populations of upregulated and downregulated intron retention events. Interestingly, there appears to be no difference between the population of upregulated and downregulated intron retention events ($p < 0.05$, $\mu = 0.2$). This similarity indicates that splice site strength alone does not account for the observed differential expression of intron retention events.

From the unique peptides identified across all colorectal cancer patients using NCBI tblastn, we selected peptides from seven patients for subsequent expression validation and immunogenicity testing (**Table 1**). We shortlisted peptides from patients for whom tumour organoids were successfully derived. RT-PCR primers were designed with one primer in the exon and the other in the retained intron at both the 5' and 3' ends to selectively detect intron expression (**Supplementary Fig 1C**). Of the four tumour samples analysed, two (CV25 and CV20) exhibited significantly higher intron expression (encoding IR1 and IR8, respectively) in tumour tissue compared with matched normal samples **(Fig 2E)**. To verify that the correct intron-exon regions were amplified, agarose gel electrophoresis was performed on CV25-derived PCR products, followed by gel extraction, purification, and Sanger sequencing. The resulting DNA bands corresponded to the expected amplicon sizes **(Supplementary Fig 1D)**. Importantly, Sanger sequencing confirmed that the intron–exon junctions were specifically amplified during qPCR, indicating that the signal was not due to off-target amplification **(Supplementary Fig 1E)**. Taken together, the qPCR data demonstrates that the predicted peptides are expressed in

over half of the patients tested, validating that we can identify IR events that are truly expressed in patients.

Table 1: Selected 9-mer peptides for in vitro validation

Patient	Peptide Sequence	Gene	HLA	HLA-binding	Rank (%)
CV25	LSCSPMMRK **(IR1)**	LRRC8B	A*11:01	0.641	0.201
	QLLGPWVFK **(IR2)**	OPLAH		0.817	0.081
CV30	ILAFIAPLK **(IR3)**	CHD7	A*11:01	0.526	0.343
	NLFQVMHIK **(IR4)**	FANCB		0.404	0.533
CV31	AVVAMVTLM **(IR5)**	PLEKHG5	A*34:01	0.214	0.873
CV23	MWSFIHNNL **(IR6)**	WDR19	A*24:07	0.334	0.382
	RQDPAPQQV **(IR7)**	FERMT3	A*02:01, A*24:07	0.0645	1.60
CV20	SLPPGLRGT **(IR8)**	RNF123	A*02:07	0.165	0.820
CV21	AVLHGRLFL **(IR9)**	TPCN2	A*02:06	0.259	0.898
CV26	TGRGVLCLR **(IR10)**	ITGAM	A*31:01	0.483	0.497
	SLLASSPAR **(IR11)**	CDH24	A*11:01, A*31:01	0.490	0.486

To determine the immunogenicity of these peptides, we stimulated naïve $CD8^+$ T cells from healthy donor peripheral blood mononuclear cells (PBMCs) using autologous monocyte-derived dendritic cells (moDCs) pulsed with the predicted IR neoantigen peptides. Of the peptides that were tested, three of the peptides were determined to be immunogenic; IR3, IR7 and IR9 and these peptides had high binding affinity to HLA-

A*11:01, HLA-A*02:01 and HLA-A*02:06 respectively. As such K562 aAPCs and GZMB reporter cell lines were engineered to express the specific HLA haplotypes before proceeding with the immunogenicity validation for each of the peptides. Antigen recognition was quantified by measuring interferon-γ (IFNγ) secretion from stimulated T cells using an ELISPOT assay. T cells trained against moDCs pulsed with IR peptides were subsequently restimulated with K562 cells expressing HLA-A*11:01, HLA-A*02:01 and HLA-A*02:06 and loaded individually with IR3, IR7 and IR9 peptides respectively. Compared to unpulsed and irrelevant peptide pulsed aAPCs, aAPCs pulsed with IR9 peptides readily stimulated the IFNγ secretion of trained T cells. When IR9 peptide pulsed aAPCs were cocultured with control T cells not trained against these peptides, IFNγ secretion remained minimal (**Fig 3A & 3D**). Next, we quantified the GZMB release from IR7-specific T cells and IR9-specific T cells in response to IR7 and IR9 peptides presented by HLA-A*02:01 and HLA-A*02:06 expressing GZMB reporter cell lines. Co-culture of reporter cells pulsed with IR peptides and their trained T cells showed a significant increase in IFP^+ population indicating the functional recognition of the neoepitopes by the expanded T cells. As controls, reporter cells either not pulsed with peptides or pulsed with irrelevant EBV peptide showed significantly lower IFP positive population. Similarly, when the GZMB reporters were cocultured with the control T cells, there was a significantly lower population of IFP positive population. (Top-panel:IR9, Bottom-panel:IR7) This suggests that the IR T cells specifically recognized the IR neoepitopes (**Fig 3B & 3D)**. Trained T cells were also restimulated with IR peptides to

assess CD137 expression, a marker of T cell activation. As expected, the IR peptides resulted in significantly higher CD137 expression relative to the unpulsed and irrelevant controls as well as when cocultured with control T cells (Top-panel:IR9, Bottom-panel:IR7) (**Fig 3C & 3D**). Tetramer staining revealed that approximately 2.56 % of the IR7 specific T cells exhibited specificity towards IR7 neoepitopes, which was absent in the untrained control T cells (**Fig 3E**). The same trend was observed when tetramer staining was performed with IR3 specific T cells across two different healthy donors (**Supplementary Fig 1D**) and for IR9 specific T cells (**Fig 3F**). To assess translational evidence for the predicted IR peptides, we analyzed IR7 and IR9 using PepQuery and identified one spectrum for each peptide that passed all quality filters (**Fig 3G & 3H**). This suggests that IR7 and IR9 peptides can be translated. Taken together, we have demonstrated that neoantigens produced from intron retention, namely IR3, IR7, and IR9, can elicit CD8+ T cell responses and stimulate expansion, hence, are potential candidates for cancer vaccines.

In addition to patient-specific IR events, we investigated whether shared IR events could be identified among CRC patients. We detected 24 neoantigens present in at least six patients, accounting for ~30% of the 23 patients analyzed (**Table 2**). Among these patients, the most prevalent HLA alleles were HLA-A11:01, HLA-A02:07, and HLA-A*33:03 (**Supplementary Fig 1D**). Notably, the majority of predicted neoantigens showed predicted binding affinity to these three dominant alleles in the cohort. These findings demonstrate that IR-derived neoantigens can be shared across patients. Importantly, many exhibited high binding affinity

to the most prevalent HLA alleles, suggesting their potential utility in developing vaccines applicable to multiple patients.

Table 2: Predicted shared IR neoantigens and their HLA binding motifs

HLA binding motifs	Sequence
HLA-A11:01	QVALLGLGR
HLA-A11:01	FTQRVSAKR
HLA-A11:01	RARPLGHLR
HLA-A02:07	SAPDLISPF
HLA-A02:01, HLA-A02:06, HLA-A02:07	LLPTSFPHL
HLA-A02:07	VTGRLFIHL
HLA-A02:01, HLA-A02:06, HLA-A02:07	LQTPVSQTV
HLA-A11:01	RALFHDIGR
HLA-A11:01	HTHSCRAPR
HLA-A02:01, HLA-A02:06, HLA-A02:07	SQSPHLPSA
HLA-A02:01, HLA-A02:06, HLA-A02:07	RVTGATPNL
HLA-A11:01	VTGATPNLR
HLA-A11:01	GTLASPPSR
HLA-A02:06	RTMSVKVGA
HLA-A02:01, HLA-A02:06, HLA-A02:07	RLAHGPSSL
HLA-A11:01	QTRRGLAAR
HLA-A11:01	WSAANQTGK
HLA-A02:06	AVAFWKTPV
HLA-A02:01, HLA-A02:06, HLA-A02:07	VLWSVATPI
HLA-A02:01	FLLSYVQDS
HLA-A02:01, HLA-A02:06, HLA-A02:07	VLFHSLSLL
HLA-A11:01	TESGVMVNK
HLA-A11:01	LSLSVPALK

HLA-A11:01	VSLAGGPPR

DISCUSSION

In this study, we demonstrated that intron retention (IR) generates alternative immunogenic neoantigens that can be identified from RNA sequencing data of CRC patients. At least half of the predicted neoantigens were confirmed to be expressed in patient samples, and approximately 30% elicited in vitro T cell responses based on our prediction and validation pipeline. Our work delivers both computational and experimental evidence that IR gives rise to bona fide immunogenic neoantigens. We utilised the public CRC MS/MS dataset to approximate the translation potential of neoantigens. However, additional evidence is required to confirm their translation and validate their suitability as therapeutic targets. Ultimately, patient-specific protein MS/MS analyses may be necessary to determine whether the predicted 9-mer peptides are indeed translated.

An additional consideration for improving the accuracy of intron retention-derived neoantigen identification is the choice of RNA-seq library

preparation. For the detection of IR and other alternative splicing events, poly(A)-selected RNA sequencing may be advantageous, as it reduces the contribution of unspliced pre-mRNA transcripts that still contain introns. This approach could help minimize potential false-positive IR calls and increase confidence that predicted peptides originate from mature transcripts expressed in tumour cells. Future studies systematically comparing ribosomal depletion-based and poly(A)-selected RNA-seq strategies will be valuable for optimizing the precision of IR-based neoantigen discovery.

Another critical consideration for the clinical translation of IR-derived neoantigens is the possibility that predicted epitopes may also be expressed in non-malignant tissues. Comprehensive assessment of this risk would require systematic evaluation across large cohorts of healthy donors and reference datasets, including specialized cell types such as medullary thymic epithelial cells (mTECs), which play a key role in central tolerance. While such analyses were beyond the scope of the present study due to data availability constraints, incorporating healthy tissue and mTEC-derived proteomic and transcriptomic datasets, as described in previous work,[16] will be an important direction for future refinement of IR neoantigen prioritization pipelines.

Importantly, we found that IR events can generate not only patient-specific but also shared neoantigens across multiple CRC samples. For instance, the predicted peptides (Table 2) were consistently retained in several patients, demonstrating that recurrent IR events can serve as a foundation

for shared antigen discovery. Refining prediction pipelines to better capture such recurrent events will be crucial for identifying neoantigens with broader clinical applicability. The identification of shared IR-derived neoantigens is particularly significant for microsatellite-stable CRC, where mutation-derived epitopes are scarce, as it highlights the potential for developing vaccines that extend beyond personalized approaches to off-the-shelf immunotherapies capable of benefiting larger patient populations.

In conclusion, this study provides the experimental validation that IR-derived peptides can drive $CD8^+$ T cell recognition and expansion, establishing IR-derived neoantigens as a novel and clinically relevant repertoire of tumour antigens. Our results are also consistent with recently published studies which suggest that IR-derived peptides have a great potential in serving as alternative sources of neoantigens[17,18]. These findings open new opportunities to expand the range of targetable antigens for cancer vaccine development, especially in low-TMB tumours where conventional strategies are limited.

METHODS

DNA and RNA extraction

DNA and RNA from colorectal cancer (CRC) patient samples were isolated using the AllPrep DNA/RNA Mini Kit (Qiagen, #80204) as previously described[19]. Genomic DNA from human peripheral blood mononuclear cells (PBMCs) was purified with the Wizard Genomic DNA Purification Kit (Promega, #A1125). Total RNA from cultured cell lines was extracted using the E.Z.N.A. Total RNA Kit I (Omega Bio-Tek, #R6834). Cell-free RNA from plasma samples was isolated using the QIAamp Circulating Nucleic Acid Kit (Qiagen, #55114).

Plasmid

Lentiviral packaging plasmids pMDLg/pRRE (Addgene #12251), pMD2.G (Addgene #12259), and pRSV-Rev (Addgene #12253) were obtained from Didier Trono. The lentiviral backbone pLV-EF1α-IRES-Neo was provided by Tobias Meyer (Addgene #85139), and pcDNA3.1 iCasper T2A HO1 was obtained from Xiaokun Shu (Addgene #64278). The AAVS1-CAG-hrGFP construct was acquired from Su-Chun Zhang (Addgene #52344), while AAVS1-TALEN-L and AAVS1-TALEN-R were gifts from Danwei Huangfu (Addgene #59025 and #59026). The gBlock encoding HLA-A*11:01, HLA-A*02:01 and HLA-A*02:06 coding sequences was cloned using EcoRI (New England Biolabs #R3101L) and BamHI (New England Biolabs #R3136L) into pLV-EF1a-IRES-Neo to generate lentivirus.

The GZMB reporter and cr-ICAD (caspase-resistant inhibitor of caspase-activated DNase) sequences, synthesized by IDT, were cloned as follows: the GZMB reporter[19] was inserted into pcDNA3.1 iCasper T2A HO1,

replacing the caspase cleavage site with a GZMB-specific site via BamHI and EcoRI digestion. The resulting GZMB sequence was amplified, digested with SalI and EcoRV, and inserted into AAVS1-CAG-hrGFP to generate AAVS1-CAG-GZMB Reporter[19]. The cr-ICAD gBlock was directly cloned into AAVS1-CAG-hrGFP using SalI and EcoRV to produce AAVS1-CAG-cr-ICAD.

RT-qPCR and Sanger Sequencing

Complementary DNA (cDNA) was generated using the GoScript Reverse Transcription System Kit (Promega, #A5000) from 3 µg of total RNA extracted from tumour and adjacent normal tissue samples. RT-qPCR was carried out with SYBR Select Master Mix (Thermo Fisher Scientific, #4472919), using 50 ng of cDNA in a 10 µL reaction. The PCR products were separated via agarose gel electrophoresis, then excised and purified with the E.Z.N.A. Gel Extraction Kit (Omega Bio-Tek, #D2500). The purified DNA was subsequently sent for Sanger sequencing with the same RT-qPCR primers to verify intron-exon junctions. RT-qPCR reactions were performed in triplicate for each intron-exon primer set, with threshold cycle (Ct) values calculated using the Design and Analysis Software (Thermo Fisher Scientific, V.2.6). The mean Ct values were analyzed using the comparative Ct method and normalized to the housekeeping gene beta actin.

HLA haplotyping

To obtain high resolution (4-digits) haplotype for the HLA-A, exons 2 and 3 were amplified using GoTaq G2 Green Master Mix (Promega, #M7822) from healthy donor PBMC DNA that was extracted using the Wizard Genomic DNA Purification Kit (Promega, #A1120). Primers used were forward primer HLA-A-F: GAAACSGCCTCTGYGGGGAGAAGCAA and reverse primer HLA-A-R: TGTTGGTCCCAATTGTCTCCCCTC. A 985-bp PCR product was confirmed by agarose gel electrophoresis and purified using the E.Z.N.A Cycle Pure Kit (Omega Bio-Tek, #D6492-02). The purified PCR product was sequenced by Bio Basic Asia (Singapore) using the abovementioned primers. The sequencing data were analyzed using SOAPTyping[21] and filtered based on the most common haplotypes in Singapore.

Generation of lentivirus

HEK293T cells were seeded in a T175 flask a few days before transfection to allow them to grow to 70% confluency. 40µg of an appropriate transfer plasmid (HLA-A*02:06, HLA-A*02:01 and HLA-A*11:01), pMDLg/pRRE, pMD2.G, and pRSV-Rev (mass ratio 4:2:1:1) were transfected using Lipofectamine 3000 Transfection Reagent (Thermo Fisher Scientific # L3000015) into HEK293T cells. Media were changed 6 hrs after transfection. The media containing virus was collected at 72 hrs after transfection, and the virus was pelleted down using an ultracentrifuge (100,000 g, 2 hrs at 4 °C). HLA-A*02:06, HLA-A*02:01 and HLA-A*11:01 virus pellets were resuspended in PBS and kept in -80 °C[20].

Generation of aAPC

The HLA-null K562 cells were used to generate aAPC as previously described[19]. Cells were transduced with CD80 (Origene, #RC206540L1V), CD86 (Origene, #RC217341L1V). Then HLA-A*11:01, HLA-A*02:06 and HLA-A*02:01 viruses were used to transduce the K562 aAPCs to generate three separate aAPCs with different HLA alleles. Briefly, cells were mixed with appropriate amounts of virus and 8 µg/mL polybrene (Merck #TR-1003-G), centrifuged at 800 G for 1 hr at 32 °C and cultured for 1 day. Media containing virus were replaced the next day to expand cells. CD80, CD86 and HLA triple positive cells were sorted using BD FACSAria Fusion Cell Sorter.

Generation of GZMB reporter cells

HEK293T HLA KO cells were transfected with AAVS1-TALEN-L, AAVS1-TALEN-R plasmid and AAVS1-CAG-cr-ICAD (mass ratio 1:1:1) using the Lipofectamine 3000. The transfected cells were then selected with puromycin (Thermo Fisher Scientific #A1113803) at 1µg/ml to get stable cells expressing cr-ICAD. The HEK293T-HLA KO-cr-ICAD cells were then transfected with AAVS1-TALEN-L, AAVS1-TALEN-R plasmid and AAVS1-CAG-GZMB Reporter (mass ratio 1:1:1).[18] Stable cells expressing the GZMB Reporter were selected by sorting for GFP positive cells. Lastly, HEK293T-HLA KO-cr-ICAD-GZMB reporter cells were transduced with the

HLA-A*11:01, HLA-A*02:01 or HLA-A*02:06 virus to generate three separate GZMB reporter cells lines with different HLA. Stable cells expressing HLA-A*11:01 were selected by staining for HLA and sorted[20].

Monocyte-derived dendritic cell (moDCs)-naïve CD8$^+$ T cells co-culture

The EasySep Human Monocyte Isolation Kit (STEMCELL, #19359) was used to isolate monocytes from healthy donor PBMCs. ImmunoCult^{TM} Dendritic Cell Culture Kit (STEMCELL, #10985) was used to differentiate and mature these monocytes in Dendritic Cells (DCs) as previously described[47]. During this maturation step, nine-mer neoantigen peptides, that were ordered from Genescript, were added to the DCs at 100 µg/ml during maturation. These DCs were gently lifted from the wells using PBS containing 2% Human Serum (HS, Sigma #H3667) and 1% P/S by pipetting up and down, after 22 hours of maturation. EasySepTM Human Naïve CD8+ T cell Isolation Kit (STEMCELL #19258) was used to isolate naïve CD8+ T cells from matched PBMCs. The mature DCs from the previous step were then co-cultured with these naïve CD8+ T cells at a 1:4 ratio (DC: T cells) in the coculture medium. (RPMI 1640 with 10% HS, 1% P/S, 30 ng/ml IL-21 (Miltenyi #130-095-768), 40 ng/ml IL-7 (Miltenyi #130-095-361), and 40 ng/ml IL-15 (Miltenyi #130-095-762)) in a 48-well plate. After three days of coculture, on day four, the expansion media (RPMI 1640 with 10% HS, 2% P/S, 200 IU/ml IL-2 (Miltenyi #130-097-744), plus 40 ng/ml IL-7 and IL-15) was added to the co-cultured cells. Expansion

medium was replenished every 2 to 3 days to promote neoantigen-specific CD8+ T cell expansion, for 14 days[20].

GZMB reporter assay

One day before GZMB coculture assay, GZMB reporter cells were seeded in a 24-well plate in an appropriate cell density and neoantigen peptides of interest were added to the reporter cells at a final concentration of 100 µg/ml. On the day of coculture, neoantigen-specific T cells were added to the corresponding wells at a ratio of 1:4 (reporter to T cell) and incubated at 37°C for 6 hours. Then the co-cultured cells were collected, centrifuged, and the cell pellet was resuspended in flow buffer (1% BSA in PBS) for flow cytometry analysis[20].

ELISPOT assay

ELISPOT was performed following manufacturer's protocol and as previously described[20]. Briefly, the strip plate from the Human IFNγ ELISPOT PLUS kit (Mabtech, #3420-4APT-10) was washed with PBS and blocked with PBS containing 10% FBS for 30 minutes at 37°C. During the coculture, ten thousand neoantigen-expanded T cells were stimulated with 2,000 K562 aAPCs in either the absence (control) or presence of neoantigen or EBV peptides in the wells of the strip plate for 24 hours. Afterwhich, the cocultured cells were discarded, and the wells washed with PBS for a minimum of five times. Then biotin-conjugated anti-IFNγ primary antibody diluted in PBS with 0.5% FBS, was added to the wells

and incubated for 2 hours at room temperature. After incubation, the wells were washed with PBS and incubated with Streptavidin-ALP secondary antibody diluted in PBS with 0.5% FBS for 1 hour. Subsequently, BCIP/NBT-plus substrate was added and allowed to incubate at room temperature for 15 minutes. After drying overnight, the spots were scanned and analyzed using the CTL ImmunoSpot S6 Entry M2 Analyzer[20].

Tetramer staining

Tetramer staining was performed following the manufacturer's protocol and as previously described[20]. Flex-T HLA-A*11:01 monomers UVX (BioLegend, #280007) were assembled into peptide-loaded tetramers according to the manufacturer's instructions. T cells were stained with these peptide-loaded tetramers and an anti-CD8 antibody (BioLegend, #344718). After staining, the cells were pelleted by centrifugation. The cell pellets were then resuspended in flow buffer containing propidium iodide (PI, Sigma-Aldrich, #P4170, diluted 1:1000), a live/dead marker, and analyzed using a flow cytometer[20].

CD137 expression

Peptide-pulsed aAPCs were co-cultured with their corresponding neoantigen-expanded T cells at a 1:4 ratio (aAPC:T cell) for 24 hours as previously described[47]. Following incubation, cells were stained on ice for 30 minutes with LIVE/DEAD Fixable Near-IR dye, anti-CD8, and anti-

CD137 antibodies (BioLegend, #309810; 1:200 dilution). After three washes, samples were analyzed by flow cytometry.

Tumour-Matched Sample Analysis

Tumour-matched samples were obtained from National University of Singapore (NUS) and Genome Institute of Singapore (GIS)[22], comprising 23 patient samples in total. For each patient, IRFinder (v1.3.1) was used to identify intron retention in tumour and normal samples. Differential intron retention (IR) analysis was performed without replicates using IRFinder, based on Audic-Claverie test. Retained introns were considered for downstream analysis if:

1. IRratio in normal = 0
2. 0.05 < IRratio in tumour < 0.7
3. No warnings for the intron in either sample

Translation and Immunogenicity Analysis

Pileup was performed on the reads to determine the sequences covering retained introns, focusing on regions flanking the 5’ and 3’ splice junctions. For introns retained from the 5’ splice site, the reading frame was derived from the preceding exon, and the sequence was translated until a stop codon was encountered. For introns retained from the 3’ splice site, translation began at the first start codon derived from the pileup sequence and continued until the 3’ splice site or a stop codon was reached.

If HLA typing was not provided, arcasHLA (v0.6.0) was used to determine the patient's HLA.

Immunogenicity of translated sequences was assessed using netMHCpan (v4.1) with a k-mer length of 9. Peptides with a threshold ≤ 2% were retained.

PepQuery

Filtered immunogenic peptides were queried against two sets of MS/MS datasets using PepQuery(v2.0):

CRC-related datasets:

- CPTAC_Prospective_Colon_PNNL_Phosphoproteome_PDC000117
- CPTAC_Prospective_Colon_PNNL_Proteome_PDC000116
- CPTAC_Prospective_Colon_VU_Proteome_PDC000109
- CPTAC_TCGA_Colon_Cancer_Proteome_PDC000111

Healthy tissue datasets:

- Deep_29_healthy_human_tissues_PXD010154
- GTEx_32_Tissues_Proteome_PXD016999

Only neoantigens that are classified as confident or C7, per PepQuery algorithm, is retained. Peptides that were confidently identified in CRC datasets but not found in healthy tissue datasets were validated using tblastn to confirm unique genomic origins.

Bioinformatics Analysis

IRFinder's output was used to calculate the intron depths for the first and last 50 bp of each intron, averaged per tumour-matched sample. Differential expression analysis was performed using DESeq2 (v1.46.0) in R (v4.4.2) with p = 0.05. NA results were excluded, and significantly differentially expressed introns were identified with:

- |log2 fold change| > 2
- padj < 0.05

Visualization included a volcano plot in R and a heatmap of differentially expressed introns using ComplexHeatmap (v2.22.0). Genes were grouped into upregulated and downregulated categories based on log2 fold change value. Gene enrichment analysis was conducted using clusterProfiler (v4.14.0) with Org.Hs.eg.db (v3.20.0). Disease enrichment analysis was performed with DOSE (v4.0.0).

Statistics

Statistical significance was assessed with paired or unpaired two-tailed Student's t-tests using Prism V.8 (GraphPad) or R. For all figures: *p<0.05; **p<0.01; ***p<0.001; ****p<0.001 .

Data Availability

The RNA sequencing data generated in this study have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number

GSE292858. The data are publicly available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE292858.

Code Availability

The bioinformatics pipeline code can be found in a GitHub Repository (https://github.com/kingofburg/IRANP).

DECLARATIONS

Ethics approval

All the experiments involving human PBMCs and clinical samples were approved by Institutional Review Board of National University of Singapore (NUS) (Approval #LH-20-026E) and National Healthcare Group Domain Specific Review Board (Approval #2020/01343). Written informed consent was obtained from all participants before their participation in the study. All methods were performed in accordance with the relevant guidelines and regulations.

Funding Declaration

This work was supported by grants from the National Research Foundation (NRF-NRFF12-2020-0007), the Ministry of Education (T2EP30123-0038), the Singapore Ministry of Health's National Medical Research Council

(NMRC) (OF-IRG23jul-0080), the NMRC Open Fund-Large Collaborative Grant ("OF-LCG") (MOH-001573), NRF Competitive Research Programme (CRP28) and the Institute for Health Innovation & Technology (iHealthtech). T.M. was supported by NUS Research Scholarship from the Singapore Ministry of Education.

Competing Interests

The authors declare no competing interests.

Acknowledgements

We thank Shyam Prabhakar for providing access to the RNA and DNA sequencing data previously published in Joanito et al., Nature Genetics (2022). We thank Teo Hong Kai and Liu Beijia from our laboratory for providing the HLA-A*02:06 and HLA-A*02:01 aAPCs and GZMB reporter cell lines, respectively, for the immunology experiments. We thank Dr. Renyi for cloning the HLA allele-specific coding sequences into the expression plasmid. We are grateful to the patients, their families, and the medical teams involved in the clinical trial. We also extend our thanks to the Singapore Health Sciences Authority for providing healthy donor blood samples.

Authors' contributions

T.M and G.C conceived the study. T.M conducted all the RT-PCR and immunology experiments. B.K.B.R performed all the bioinformatics analysis. C.Z.M.C did the haplotyping experiments. M.C.K assisted T.M with immunology experiments. S. B. E., W.C, K.L., I.J.T., B.L., I.B.H.T., K. K. T. and Y.C.K provided clinical samples. T.M, B.K.B.R and G.C wrote the manuscript. T.K.T. contributed to bioinformatics guidance and manuscript preparation. G.C supervised the entire study and edited the manuscript.

Figure Legends

Figure 1: Identification of Intron Retention (IR) neoantigens from CRC clinical samples

(A) Flow of computational prediction pipeline. **(B)** Percentage of intron reads to the total mapped reads in individual patients. **(C)** Spread of tumour-unique and normal-unique Intron Retention (IR) events. **(D)** Volcano plot showing differentially expressed IR events between tumour and paired normal samples. The x-axis represents the log2 fold change, and the y-axis represents the −□□□10 adjusted p-value. Statistical significance was determined using the DESeq2 package, p-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure, with significant introns (adj. p-value < 0.05, |log2FoldChange| > 2) highlighted. **(E)** Heatmap showing the expression of the top 50 differentially expressed IR events (Adj. p-value < 0.05).

Figure 2: Gene ontology and expression levels of predicted IR neoantigens

(A) Top 15 enriched gene ontology (GO) of cellular processes for genes which contains upregulated introns (n = 695). **(B)** Top 10 enriched disease ontology (DO) for genes which contains upregulated introns (n = 695). **(C)** Top 15 enriched gene ontology (GO) of cellular processes for genes which contains downregulated introns (n = 1094). **(D)** Top 10 enriched disease ontology (DO) for genes which contains downregulated introns (n = 1094). GO panels are generated by GSEA package, significance is reported as nominal p-values (see colour scale, adj. p-value < 0.05). Disease enrichment significance was assessed using DOSE package, using

a hypergeometric test with Benjamini–Hochberg FDR correction. In the dot plot, point size is proportional to the gene count per DO term, and point colour indicates the adjusted p-value (BH) (see color scale, adj. p-value< 0.05) (**E)** Expression (relative fold change) of retained introns in CRC patient samples encoding the IR peptides (CV25: IR1, CV20: IR8, CV31: IR5 and CV26: IR11) by RT-qPCR. S2F1R1: Intronic sequence 2 amplified by forward primer 1 & reverse primer 1. S2F2R2: Intronic sequence 2 amplified by forward primer 2 & reverse primer 2. Unpaired Student's t test was used to determine significance (n = 3, * $p < 0.05$, ** $p < 0.01$).

Figure 3: Immunogenicity and translational validation of the predicted IR neoantigens from CRC patients

(A) ELISPOT quantification of IFNγ secretion by IR9 specific T cells after stimulation with HLA-A*02:06 expressing aAPCs, in the presence (leftmost panel) or absence of IR9 peptide or in the presence of irrelevant EBV peptide as well as when cocultured with control T cells. **(B)** Representative FACS plots of GZMB reporter validation in IR9 specific T cells (Top row) and in IR7 specific T cells (bottom row) in the presence (leftmost panel) or absence of peptides or in the presence of irrelevant EBV peptide as well as when cocultured with control T cells. IR9 specific T cells were cocultured with peptide-pulsed HLA-A*02:06 aAPCs while IR7 specific T cells were cocultured with peptide-pulsed HLA-A*02:01 aAPCs. **(C)** Representative FACS plots of CD137 expression in IR9 specific T cells (Top row) and in IR7 specific T cells (bottom row) in the presence (leftmost panel) or absence of peptides or in the presence of irrelevant EBV peptide as well as when cocultured with control T cells. IR9 specific T cells were cocultured with peptide-pulsed HLA-A*02:06 aAPCs while IR7 specific T cells were cocultured with peptide-pulsed HLA-A*02:01 aAPCs **(D)** Quantification of the plots in **Fig A-C. (E)** IR7 tetramer and CD8 staining of the IR7 specific T cells. **(F)** IFNγ release assay of IR9 specific T cells when stimulated with HLA-A*02:06 expressing aAPCs in the presence (1st panel) or absence (2nd panel) of IR9 peptides as well as when cocultured with control T cells (3rd panel). T cell training was performed with one healthy donor. The subsequent validation assays ELISpot, GZMB killing assay, CD137 activation assay and tetramer assay were performed with two technical replicates and Student's t-test was used to determine significance (n=2, * $p < 0.05$, ** $p < 0.01$). **(G)** Mass spectrum of IR9 peptide from PepQuery 2.0. **(H)** Mass spectrum of IR7 peptide from PepQuery 2.0.

Supplementary Figure 1: Splice site strength and transcriptional evidence of introns

(A) Pie-chart showing the tumour-unique and normal-unique IR events. **(B)** Scatter plot of all intron splices strength, with upregulated and downregulated intron retention events differently coloured. **(C)** Design of RT-qPCR primers for detection of transcripts with retained introns. **(D)** Agarose Gel Electrophoresis of qPCR analysis for patient CV25 intron sequence 1 (seq 1) and sequence 2 (seq 2), as well as the housekeeping gene β-Actin with the respective DNA bands. **(E)** PCR product 3' intron-exon junction sequence of CV25 intronic sequence 1 with the corresponding Sanger sequencing result for the forward primer (F1) and reverse primer (F2) at the 3' intron-exon junction as well as the 5' intron-exon junction.

Supplementary Figure 2: Tetramer staining

(A) Tetramer staining showing the percentage of the control T cells specific to IR3 peptides (first column) or the percentage of IR3 specific T cells to the irrelevant EBV peptide (second column) or the percentage of IR9 specific T cells to IR9 peptides (third column) using T cells from two different healthy donors (n=2). **(E)** Bar plots showing the HLA allele distribution across all the tested patients.

FIGURES

Figure 1

FIGURE 1

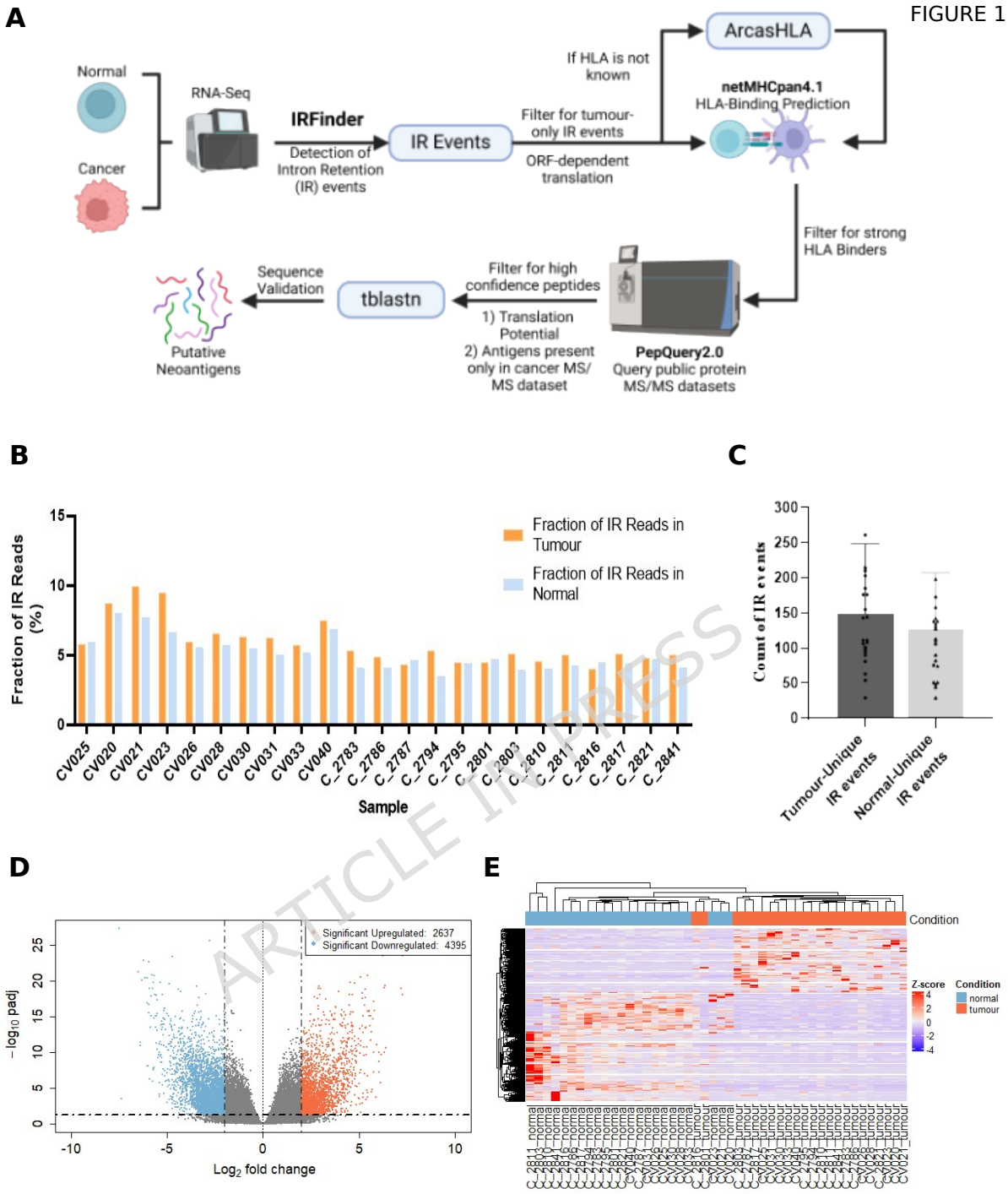


Figure 2

FIGURE 2

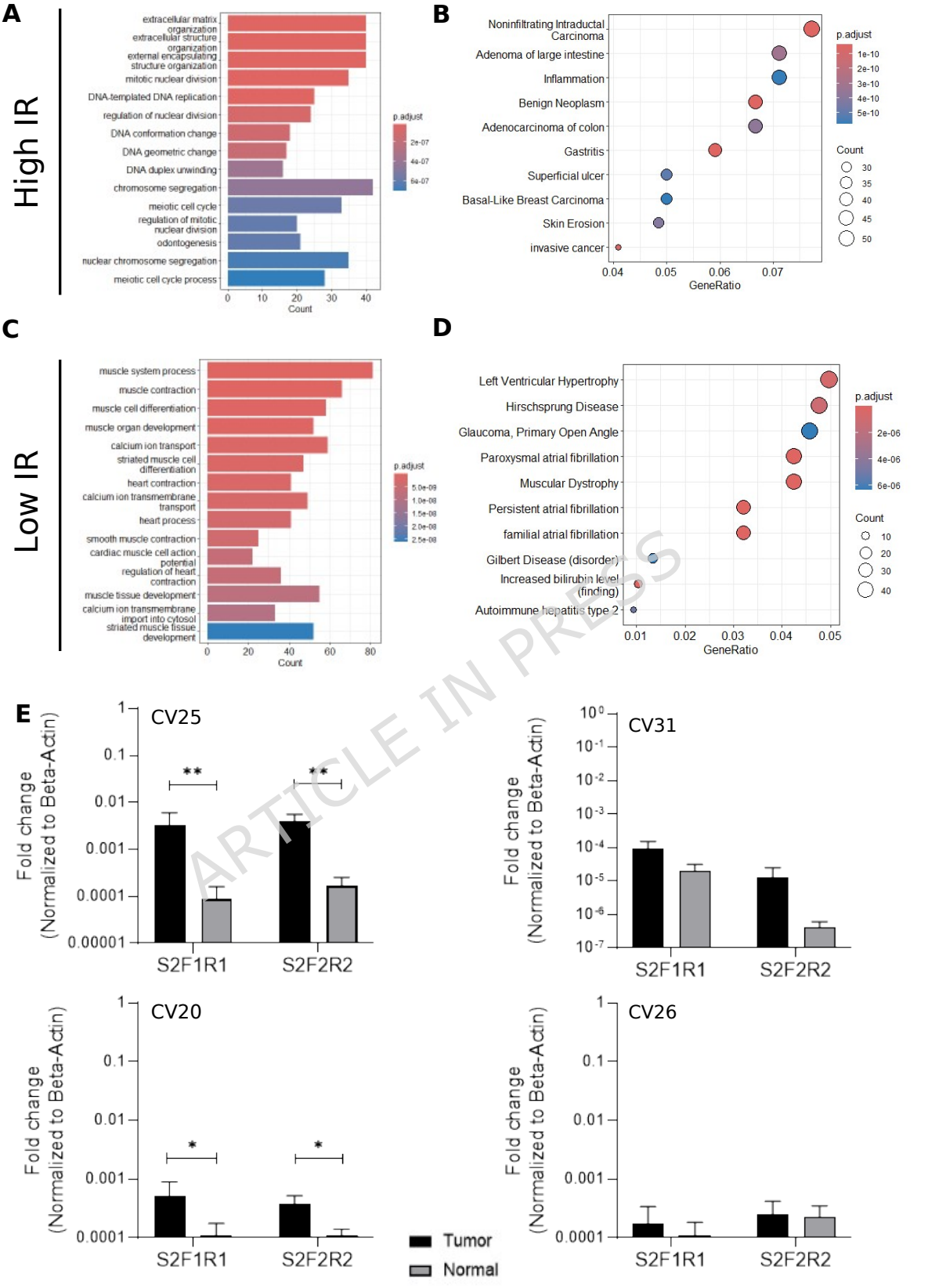
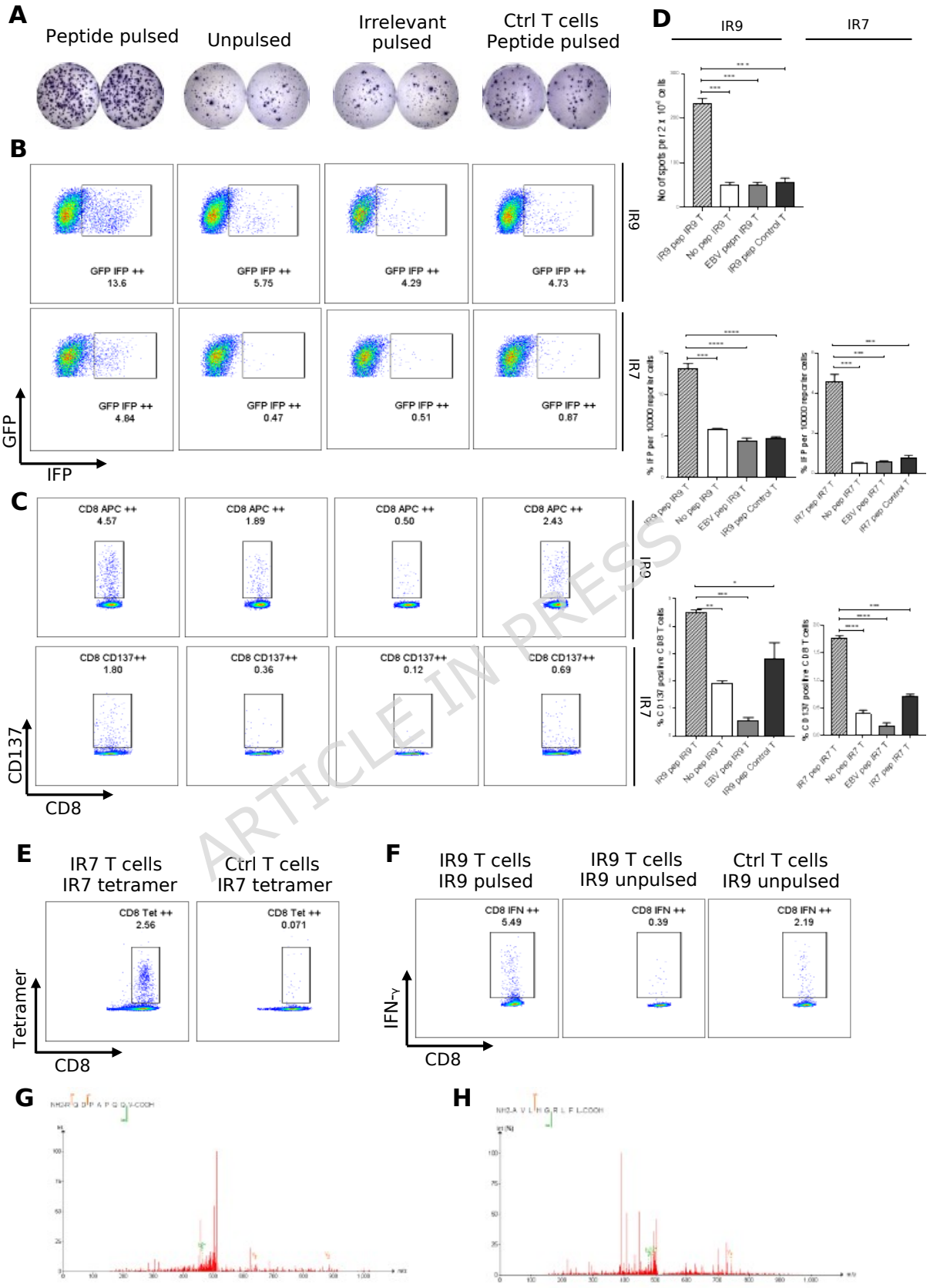
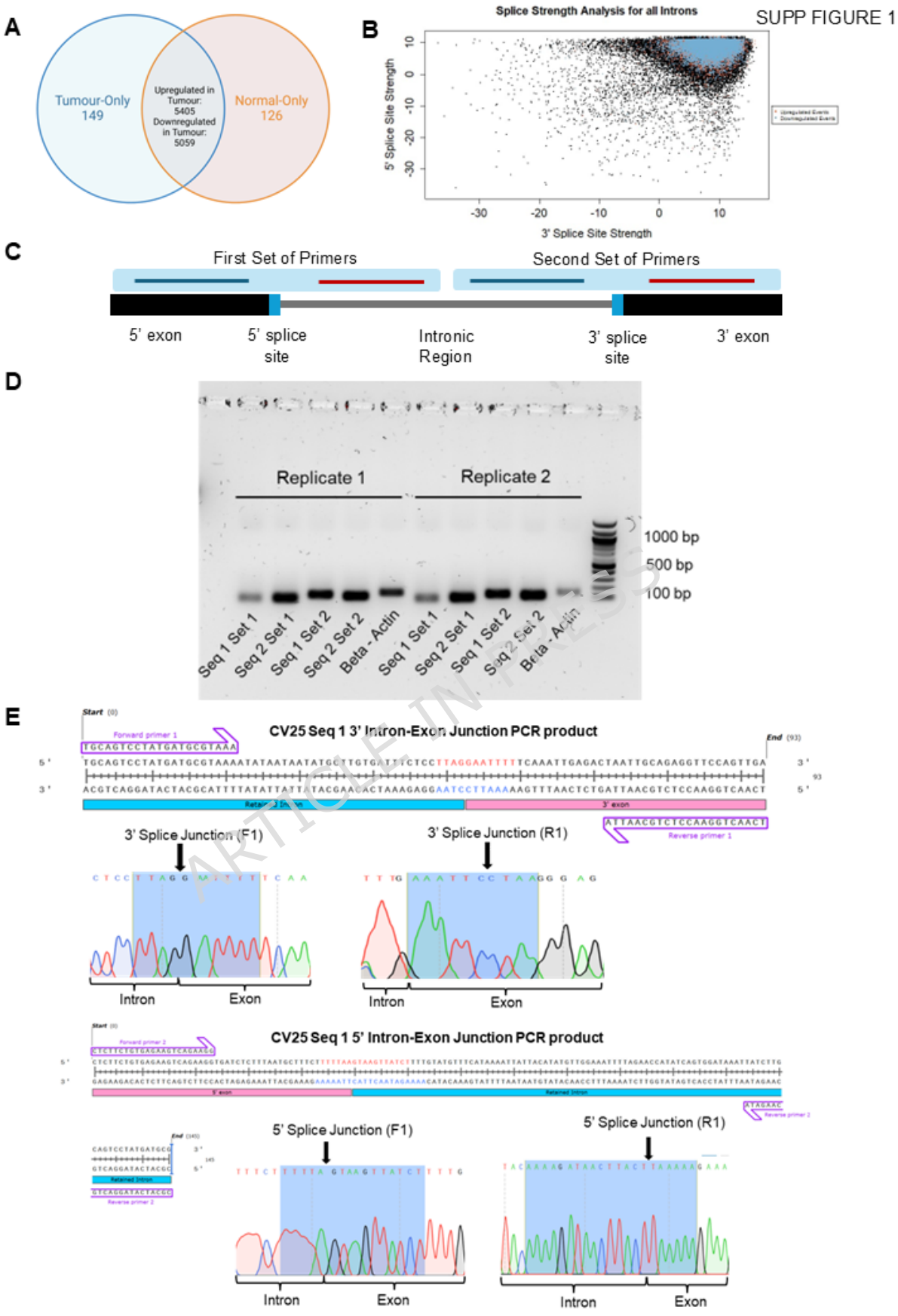


Figure 3

FIGURE 3



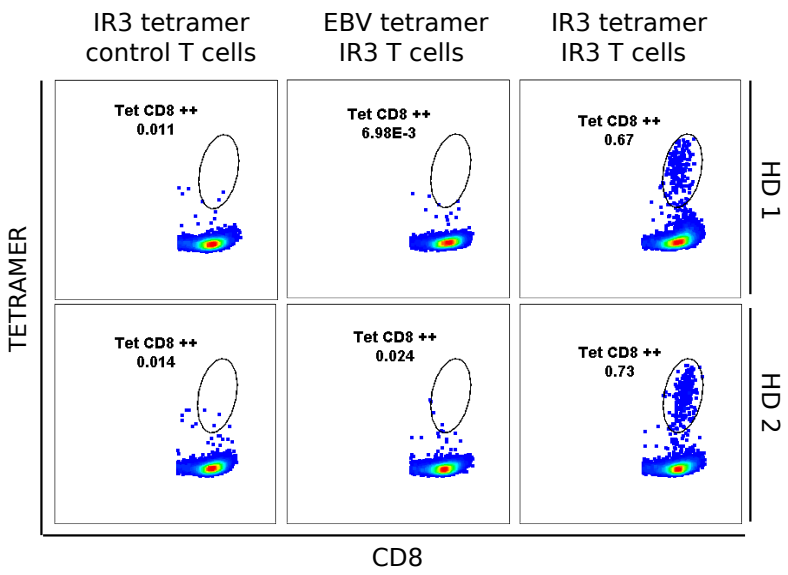
Supplementary Figure 1



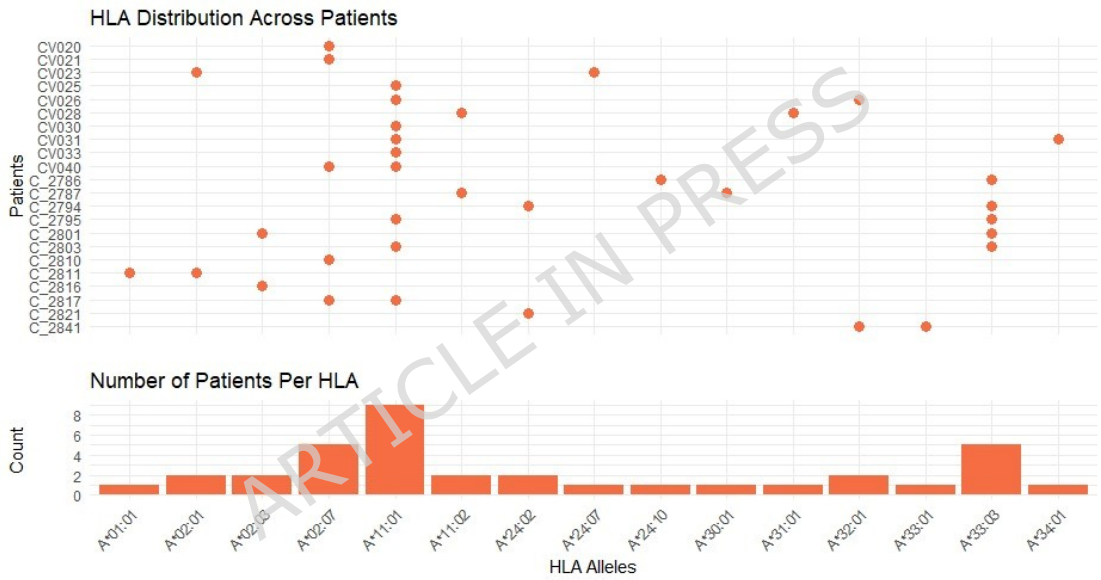
Supplementary Figure 2

SUPP FIGURE 2

A



B



REFERENCES

1. Xie N, Shen G, Gao W, Huang Z, Huang C, Fu L. Neoantigens: promising targets for cancer therapy. *Signal Transduct Target Ther*. 2023;8(1):1-38. doi:10.1038/s41392-022-01270-x

2. Bobisse S, Foukas PG, Coukos G, Harari A. Neoantigen-based cancer immunotherapy. *Ann Transl Med*. 2016;4(14):262. doi:10.21037/atm.2016.06.17

3. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337. doi:10.1038/nature11252

4. Curty G, Marston JL, de Mulder Rougvie M, Leal FE, Nixon DF, Soares MA. Human Endogenous Retrovirus K in Cancer: A Potential Biomarker and Immunotherapeutic Target. *Viruses*. 2020;12(7):726. doi:10.3390/v12070726

5. Capietto AH, Hoshyar R, Delamarre L. Sources of Cancer Neoantigens beyond Single-Nucleotide Variants. *Int J Mol Sci*. 2022;23(17):10131. doi:10.3390/ijms231710131

6. Smith CC, Selitsky SR, Chai S, Armistead PM, Vincent BG, Serody JS. Alternative tumour-specific antigens. *Nat Rev Cancer*. 2019;19(8):465-478. doi:10.1038/s41568-019-0162-4

7. Clancy S. RNA Splicing: Introns, Exons and Spliceosome. In: ; 2008. Accessed November 13, 2023. https://www.semanticscholar.org/paper/RNA-Splicing%3A-Introns%2C-Exons-and-Spliceosome-Clancy/bfb9cb0025c318975122d04857e24638316babac

8. Furuya M, Kobayashi H, Baba M, Ito T, Tanaka R, Nakatani Y. Splice-site mutation causing partial retention of intron in the FLCN gene in Birt-Hogg-Dubé syndrome: a case report. *BMC Med Genomics*. 2018;11(1):42. doi:10.1186/s12920-018-0359-5

9. Conboy JG. A Deep Exon Cryptic Splice Site Promotes Aberrant Intron Retention in a Von Willebrand Disease Patient. *Int J Mol Sci*. 2021;22(24):13248. doi:10.3390/ijms222413248

10. Shah JS, Milevskiy MJG, Petrova V, et al. Towards resolution of the intron retention paradox in breast cancer. *Breast Cancer Res*. 2022;24(1):100. doi:10.1186/s13058-022-01593-1

11. Smart AC, Margolis CA, Pimentel H, et al. Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol*. 2018;36(11):1056-1058. doi:10.1038/nbt.4239

12. Dong C, Cesarano A, Bombaci G, et al. Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. *Oncogene*. 2021;40(42):6130-6138. doi:10.1038/s41388-021-02005-y

13. Jacob AG, Smith CWJ. Intron retention as a component of regulated gene expression programs. *Hum Genet*. 2017;136(9):1043-1057. doi:10.1007/s00439-017-1791-x

14. Monteuuis G, Wong JJL, Bailey CG, Schmitz U, Rasko JEJ. The changing paradigm of intron retention: regulation, ramifications and recipes. *Nucleic Acids Res*. 2019;47(22):11497-11513. doi:10.1093/nar/gkz1068

15. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol J Comput Mol Cell Biol*. 2004;11(2-3):377-394. doi:10.1089/1066527041410418

16. Laumont, C. M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J. P., Gendron, P., Courcelles, M., Hardy, M. P., Côté, C., Durette, C., St-Pierre, C., Benhammadi, M., Lanoix, J., Vobecky, S., Haddad, E., Lemieux, S., Thibault, P., & Perreault, C. (2018). Noncoding regions are the main source of targetable tumour-specific antigens. *Science translational medicine*, *10*(470), eaau5516. https://doi.org/10.1126/scitranslmed.aau5516

17. Xiang et al., Predominant mutated non-canonical tumour-specific antigens identified by proteogenomics demonstrate immunogenicity and tumour suppression in CRC, (2025), Cell Genomics https://doi.org/10.1016/j.xgen.2025.10106

18. Ausserhofer, M., Rieder, D., Facciolla, M., Gronauer, R., Lamberti, G., Lisandrelli, R., … Finotello, F. (2025). NovumRNA: Accurate prediction of non-canonical tumour antigens from RNA sequencing data. *iScience*, *28*(10), 113448. doi:10.1016/j.isci.2025.113448

19. Kula, T., Dezfulian, M. H., Wang, C. I., Abdelfattah, N. S., Hartman, Z. C., Wucherpfennig, K. W., Lyerly, H. K., & Elledge, S. J. (2019). T-Scan: A Genome-wide Method for the Systematic Discovery of T Cell Epitopes. *Cell*, *178*(4), 1016–1028.e13. https://doi.org/10.1016/j.cell.2019.07.009

20. Ren, Y., Manoharan, T., Liu, B., Cheng, C. Z. M., En Siew, B., Cheong, W. K., Lee, K. Y., Tan, I. J., Lieske, B., Tan, K. K., & Chia, G. (2024). Circular RNA as a source of neoantigens for cancer vaccines. *Journal for immunotherapy of cancer*, *12*(3), e008402. https://doi.org/10.1136/jitc-2023-008402

21. Zhang, Y., Chen, Y., Xu, H., Fang, J., Zhao, Z., Hu, W., Yang, X., Ye, J., Cheng, Y., Wang, J., Sun, W., Wang, J., Yang, H., Yan, J., & Fang, L. (2020). SOAPTyping: an open-source and cross-platform tool for sequence-based typing for HLA class I and II alleles. *BMC bioinformatics*, *21*(1), 295. https://doi.org/10.1186/s12859-020-03624-0

22. Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C. L. P., Kahraman, M., Srinivasan, H., Venkatesh, P. N., Poh, Z. W., Loo, J. M., Chia, S., Gan, A., Guo, Y. A., Yap, C. K., Skanderup, A. J., DasGupta, R., Prabhakar, S., & Tan, I. B. (2022). Single-cell and bulk transcriptome sequencing identifies two epithelial tumour cell states and refines the consensus molecular classification of colorectal cancer. *Nature Genetics*, 54(7), 963–975. https://doi.org/10.1038/s41588-022-01100-4