A pan-cancer atlas of T cell targets

Guangyuan Li[1,2], Omar Uriel Guzman-Bringas[1,2], Aman Sharma[1,2], Zachary Harpaz[1,2], Ethan Tardio[1,2], Aswin Natarajan[1,2], Xinya Liu[1,2], Darren Taylor[1,2], Rachel Yamin[1,2], Maxence Dellacherie[1,2], Michelle Krogsgaard[1,2], C. Russell Y. Cruz[3], Mark Yarmarkovich[1,2]

[1]Perlmutter Cancer Center, New York University Grossman School of Medicine, New York, NY, USA
[2]Department of Pathology, NYU Grossman School of Medicine New York NY 10016 USA
[3]Center for Cancer and Immunology Research, Children's National Hospital, Washington, DC, USA

Corresponding author: Mark Yarmarkovich, PhD, Mark.Yarmarkovich@nyulangone.org

Abstract
T-cell-based immunotherapies have revolutionized cancer treatment, yet only a minority of patients respond to these approaches, significantly constrained by the limited knowledge of tumor-specific antigens. Here we present a comprehensive map of T cell targets across 21 cancer types, revealing actionable tumor-specific targets in 86% of tumors. To define the repertoire of actionable T cell targets, we conducted a comprehensive pan-cancer analysis integrating data from 7,473 RNA-Seq datasets, 1,564 immunopeptidomes and 208 cancer single-cell datasets, comparing these against 17,384 normal samples covering 51 tissues. Our analysis uncovered 88 viable surface protein targets and 15,079 tumor-specific HLA-presented antigens, deriving from 11 distinct molecular events, across 21 tumor types, providing a comprehensive resource for T-cell-based immunotherapy development. We highlight 128 promising new tumor targets and validate 20 targets across five antigen classes. Among the antigens uncovered, we highlight 339 previously uncharacterized neoantigens, a new PMEL splicing peptide that we expect to be a superior antigen to the clinical target, novel self-antigens, peptides derived from ORFs previously unknown to encode proteins, and novel tumor-specific microbial targets. These findings significantly expand the therapeutic landscape of T cell therapies. To catalyze therapeutic development, we have made our pan-cancer target atlas and accompanying toolkit available to the scientific community and expect these resources will pave the way for immunotherapies across multiple cancers.

Significance
We identified 88 viable membrane proteins and 15,079 tumor-specific HLA-presented antigens across 21 cancer types through a pan-cancer analysis of RNA-Seq, immunopeptidomes, and normal tissues. This comprehensive atlas highlights diverse antigen classes and key HLA presentation factors, providing a vital resource to catalyze T-cell based immunotherapy development and enhance understanding of antigen presentation and tumor biology.

Introduction

T-cell-based immunotherapies have revolutionized cancer treatment, offering cures to subsets of patients who have historically had limited effective therapeutic options(1). Treatments such as CAR T cell therapies have shown remarkable efficacy, particularly in hematologic malignancies such as acute lymphoblastic leukemia and non-Hodgkin lymphoma(2). While targeting lineage-specific proteins shared between malignant and healthy B cells has proven clinically manageable in blood cancers as patients can tolerate the loss of healthy B cells, such on-target off-tumor effects are unacceptable in solid tumors, where off-tumor reactivity to essential healthy tissues could be catastrophic. This limitation necessitates the identification of antigens with substantial therapeutic windows. Currently, targeting strategies predominantly focus on membrane proteins, which, while promising, often lack sufficient tumor specificity, contributing to on-target, off-tumor side effects(3,4). Indeed, a recent report suggests we may be approaching the saturation point for identifying optimal membrane targets for CAR-T therapy(5), indicating a critical need for innovative targeting strategies that can distinguish between malignant and healthy tissues more effectively.

Human Leukocyte Antigen (HLA) molecules present a snapshot of the cellular proteome on the membrane of tumor cells, exposing potential tumor-specific antigens on the cell surface(6). Multiple therapeutic strategies that exert their therapeutic effects through peptides presented on HLA have delivered curative responses in the clinic, including immune checkpoint inhibitors (ICIs)(7), adaptive transfer of Tumor Infiltrating Lymphocytes (TILs)(8), TCR therapies(9) as well as recent complete response from neoantigen vaccines(10,11). Immunotherapies such as ICIs and TILs often rely on high mutational burden, therefore the majority of curative responses to these therapies are applicable only to a limited number of highly mutated tumors. Our group and others have recently demonstrated that new approaches, such as peptide-centric chimeric antigen receptors (PC-CARs), that can enable the targeting of any peptide on HLA, expanding the targeting of intracellular antigens to low-to-medium mutational tumors(12). Central to the efficacy of such immunotherapies is the identification of tumor-specific HLA-presented peptides (pHLAs), which can be derived from a broad range of genetic aberrations including both tumor-dependency genes that underpin the biology of various cancers or accompanying passenger mutations. Elucidating the landscape of pHLAs derived from these various cancer processes could reveal additional tumor vulnerabilities.

In the past decade, increasing evidence has revealed actionable HLA-presented epitopes arising from a wide spectrum of tumor-specific genetic aberrations, including canonical protein-coding genes and mutations, alternative splicing(13), ectopically expressed transposable elements (14), cryptic ORFs(15) and post-translational modifications (PTMs)(16). Many of these findings have been corroborated by immunopeptidomics, the empirical characterization of HLA-presented peptides (pHLAs) by LC-MS/MS proteomics. It has been reported that over 50% of spectra remain unannotated when mapped to the database of potential peptides from the canonical proteome, representing the “dark matter” of the immunopeptidome(17). We hypothesize that the molecular instability of cancer will result in a multitude of aberrations that can result in potential immunotherapy targets but are not captured using conventional immunopeptidomics. Hence, uncovering the “dark matter” of the immunopeptidome from other non-canonical molecular sources will reveal a rich set of additional tumor antigens. While a large number of

immunopeptidomic datasets are available, most do not have matched RNA-Seq data, restricting the peptides recovered from these searches to the canonical proteome. Comprehensively identifying all categories of molecular events requires harmonized transcriptomic and immunopeptidomic datasets, multimodal computational workflows along with well-curated large-scale normal reference to exclude non-tumor-specific molecular events, a challenge that hasn't been tackled until now. To overcome these challenges, we developed an exhaustive search engine to comprehensively profile 11 classes of genetic aberrations from RNA-Seq, encompassing all known tumor-specific molecular events. We first generated a cancer-specific molecular catalogue of 11 classes of cancer-specific molecular alterations across 21 histologies. We then utilized this catalogue as a search space to interrogate 1,564 immunopeptidomics datasets, revealing a multitude of actionable immunotherapy targets across cancers.

Result

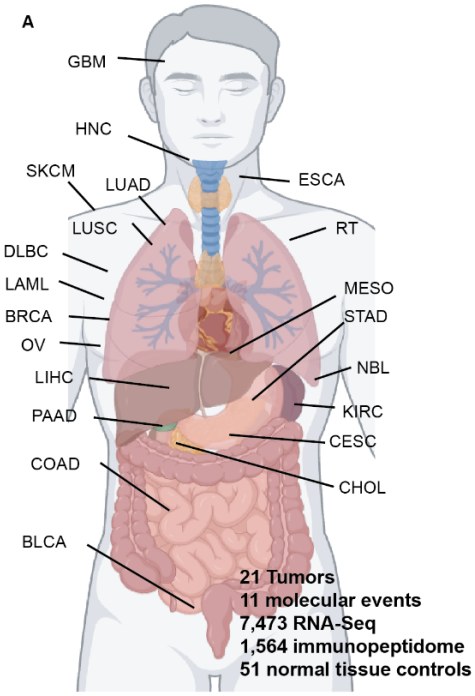
A multimodal approach to establish the comprehensive antigen landscape

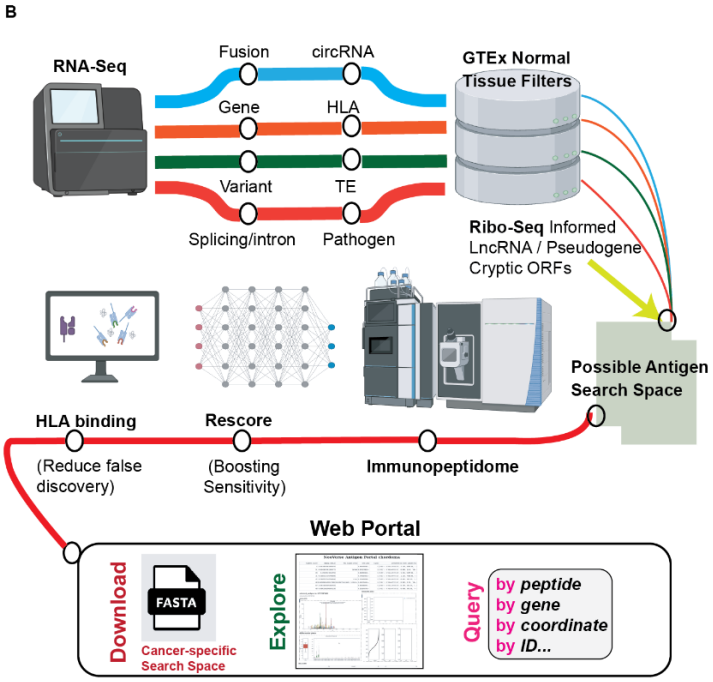
To comprehensively identify pHLAs derived from diverse molecular events, we established a multimodal computational pipeline to profile 11 classes of molecular aberrations from tumor transcriptomic data (**Methods**). We applied stringent filters for tumor specificity, utilizing 17,384 normal tissue samples from Genotype-Tissue Expression (GTEx) to exclude molecular events detected in 51 normal tissues. The retained tumor-specific events were translated into putative HLA peptides to generate search spaces used for searching against curated histology-matched immunopeptidome datasets. Subsequently, we rescored each peptide using a deep-learning model, leveraging high-confidence spectral libraries to refine these results(18) for enhanced sensitivity, followed by incorporating HLA binding prediction at the end to mitigate false positives(19) (**Figure 1A,B**). The identified tumor antigens are accessible through our web portal. We have also made the putative antigen search space available for each tumor type, enabling users to conduct comprehensive immunopeptidomic searches with their own in-house data without the need for specialized pipelines or matched RNA-Seq data.

We sourced RNA-Seq tumor data from The Cancer Genome Atlas Program (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and immunopeptidome data from public repositories including Proteomics Identification Database (PRIDE) and Mass Spectrometry Interactive Virtual Environment (Massive) (**Methods, Table S1,2**). Our dataset encompasses 7,473 RNA-Seq, 1,564 immunopeptidome and 208 single-cell RNA-Seq datasets across 21 tumor types (**Figure 1A**). Breast cancer and melanoma possess the highest number of RNA-Seq (n=1,118) and immunopeptidome samples (n=267), respectively (**Figure S1A,B**). Our tumor immunopeptidomic cohort represents a broad spectrum of 140 HLA allotypes, covering 100% of common HLA-A and HLA-C alleles and 92% of HLA-B alleles present in the U.S. population with a frequency >1% (**Table S3**). Additionally, we utilized GTEx data to compile corresponding normal controls across 51 tissues, stringently filtering out non-tumor-specific events. Overall, we developed our pipeline to identify antigens from 11 molecular sources known to yield bona fide pHLAs, including canonical protein-coding genes, alternative splicing

product, gene fusions, Transposable Element (TE)-chimeric transcripts, autonomous TEs, mutations, intron retentions, tumor-resident microbiome, expressed lncRNAs, pseudogenes and cryptic Open Reading Frames (ORFs). Although our pipeline is also capable of profiling RNA editing and circular RNAs, detecting circular RNAs typically necessitates specific preparation techniques including depleting ribosomal RNA, enriching for poly(A) negative RNA, or selectively removing linear RNA. Consequently, the poly(A) enriched mRNA protocol predominantly used in TCGA studies leads to underrepresented circular RNA transcripts in our data. RNA editing also has the potential to generate tumor antigens, but it cannot be reliably distinguished from DNA mutations without matched WGS and RNA-Seq data. Therefore, we have not considered these two aberrations in the current analysis.

Analysis of tumor-specific events from RNA-Seq data, results in high numbers of somatic mutations in melanoma and lung squamous cell carcinoma, followed by colon and stomach cancers characterized by their microsatellite instability. Conversely, immunologically cold tumors including mesothelioma and neuroblastoma exhibit the fewest mutations. Strikingly, Acute Myeloid Leukemia (AML) shows a high prevalence of tumor-specific transposable elements (n=11,264) and intron retention events (n=3,954). It has been reported that AML are characterized by widespread intron retention events(20) and endogenous retrovirus could serve as genetic enhancers in AML that drive tumorigenesis(21), suggesting a potential interplay between epigenetic changes, endogenous retroviruses, and splicing landscape changes (**Figure 1C**). When calculating the number of tumor-specific pHLAs across cancers, melanoma and AML have the highest mutation-derived and TE-derived antigens, as expected based on their high burden of underlying genetic aberrations. However, discrepancies are observed in other cancers like neuroblastoma, which, despite high number of tumor-specific alternative splicing events detected from RNA-Seq, yields fewer antigens derived from alternative splicing (**Figure 1C**), suggesting potential post-transcriptional regulatory mechanisms in antigen presentation. Notably, the low level of antigen identification in stomach cancer and cholangiocarcinoma can be attributed to minimal HLA coverage and limited sample availability in the immunopeptidome data, underscoring the need for additional immunopeptidomic studies for such cancer types. Taken together, our work provides a comprehensive and actionable landscape for exploring tumor antigens, spanning both a variety of tumors and molecular events.





C

Number of Tumor Specific Molecular Events/ HLA Antigens

		BRCA	KIRC	COAD	STAD	MESO	LIHC	ESCA	CESC	BLCA	RT	AML	DLBC	GBM	NBL	PAAD	HNSC	OV	LUSC	LUAD	CHOL	SKCM
RNA	self-gene	42	27	92	65	39	13	56	88	41	148	146	123	44	139	46	58	57	72	29	56	43
RNA	splicing	2091	1785	544	5780	1310	1816	3730	1925	1395	9925	2383	2716	2109	10234	897	2327	7461	2262	1261	1011	2812
RNA	intron	9	19	2	113	3	1	61	5	7	97	3954	8	17	253	4	9	95	13	12	7	18
RNA	TE	24	85	40	51	51	45	89	36	35	69	11264	61	27	152	22	71	74	31	27	36	48
RNA	variant	1814	502	5846	6377	74	901	602	961	2495	544	142	80	1129	80	353	1235	855	11414	2238	93	13167
RNA	pathogen	1	0	2	4	0	2	3	2	0	0	0	0	1	1	0	1	3	0	0	0	1
RNA	fusion	1893	24	203	2254	82	230	1955	150	575	25	32	47	225	196	258	1269	1987	2100	450	9	128
Peptide	self-gene	57	63	872	11	27	138	116	617	4	16	346	143	233	85	67	178	352	435	198	2	1028
Peptide	splicing	8	5	18	1	0	36	20	14	0	31	14	4	36	18	5	16	86	55	28	0	159
Peptide	TE Autonomous	0	0	5	1	0	9	12	1	0	7	284	0	4	15	0	3	18	3	7	0	22
Peptide	TE Chimera	2	5	25	0	0	28	12	7	0	12	2	2	26	11	0	18	33	46	14	0	89
Peptide	Intron retention	0	0	1	0	0	0	1	0	0	1	36	0	0	3	0	0	6	4	0	0	6
Peptide	pathogen	3	0	459	5	0	5	123	3	0	0	0	0	6	2	0	17	1083	0	0	0	1400
Peptide	fusion	2	0	2	0	0	0	0	0	0	1	0	0	1	0	1	8	3	8	2	0	2
Peptide	variant	1	1	38	0	0	4	0	3	0	0	0	0	2	0	1	6	3	90	15	0	214
Peptide	lncRNA	22	6	93	0	3	34	4	23	0	6	20	3	104	9	4	22	70	86	76	1	303
Peptide	pseudogene	6	4	21	0	0	11	2	10	0	4	7	2	22	2	4	9	22	30	25	0	91
Peptide	cryptic ORF	118	19	436	0	20	177	24	156	1	23	95	11	372	38	21	95	361	368	352	0	1395

Figure 1. Overview of the Analytical Pipeline and Immunopeptidome Atlas (A) Illustration of the cancer types included in this study, cancer abbreviation and full name are provided in method section (B) Computational pipeline detailing the steps used to identify, refine, and enhance the detection of tumor-specific antigens, along with a depiction of the functionalities of associated interactive web portal. (C) The distribution of tumor-specific antigens identified across all analyzed cancers, derived from 11 molecular sources, at both RNA and peptide levels.

Membrane proteins provide limited new therapeutic opportunities

We first evaluated the landscape of surface proteins suitable for conventional CAR-T therapy, applying the identical filtering stringency used for pHLA targets to all annotated human cell membrane proteins with exposed extracellular domains. This unbiased analysis identified 88 membrane protein candidates across tumors with clear therapeutic windows by comparing their gene expression in tumors to all normal human tissues (**Methods**). As expected, this list includes TNFRSF17/BCMA and CD19, the only two FDA-approved CAR-T targets to date, along with eleven additional targets currently being investigated in clinical trials (CLEC12A, FOLR1/PSMA, IL1RAP, CD70, CEACAM5, MSLN, ULBP2, MUC16, GPC2, CDH17, CLDN6). Further literature review shows that 20 additional targets have established CARs and are actively being tested in preclinical models. Moreover, our analysis also highlighted 47 previously unrecognized antigens with favorable therapeutic windows, rendering them as novel candidates for conventional CAR-T therapy (**Figure 2, Table S4**) and warranting further investigation. Notably, several clinical and preclinical CAR targets, including ERBB2/HER2, EPCAM and MUC1, did not pass our filtering, consistent with reported preclinical and clinical toxicities or a lack of a therapeutic window for effective targeting(22–24). CAR targets such as GD2, a glycosphingolipid, are not identifiable through this analysis and several resulting targets are likely due to RNA from infiltrating immune cells in tumors. This analysis both reveals new membrane proteins that can serve as potential CAR targets and underscores the need for the identification of additional tumor antigens.

Row Normalized TPM
BRCA
KIRC
COAD
STAD
MESO
LIHC
ESCA
CESC
BLCA
RT
AML
DLBC
GBM
NBL
PAAD
HNSC
OV
LUSC
LUAD
CHOL
SKCM
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Kidney - Medulla
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Pancreas
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Thyroid
Uterus
Whole Blood
TNFRSF17
CD19
CDH17
CLDN6
GPC2
CLEC12A
FOLR1
IL1RAP
CD70
CEACAM5
MSLN
ULBP2
MUC16
PRSS21
CDH3
LILRB4
NTRK1
CEACAM6
CLDN3
CD79A
CD84
VTCN1
SLC6A3
TNFSF13B
ADAM12
CD48
UPK3B
CDH6
ENPP3
SLC2A5
LRRC15
ESR1
CLDN4
CYSLTR1
SLC10A4
SMPDL3B
MCOLN2
RNF43
SCTR
STRA6
SLC22A16
EPHB2
CLECL1
OR2I1P
TRABD2A
GAPT
CNGB1
LPAR3
NOX1
CLDN16
ITLN1
ST8SIA2
BDKRB1
LRRN4
CLDN2
CX3CR1
CXCR6
C1orf186
HAVCR1
TSPAN10
ADGRG5
ATP8B4
SLC17A3
CHRNB4
GABRB3
ENPEP
SLC6A2
CHRNA3
LY6K
GPR35
TMPRSS3
CXCR3
HAS3
PRTN3
GPR160
ALK
P2RY8
GAL
SLC52A3
CD244
LAG3
TRBV28
TRGV5
TRGC1
SIT1
TM4SF5
CD3D
TIGIT
FDA
Clinical Trial
Preclinical Test
Additional Viable Target
Bystander T cell

Figure 2. Suitable Tumor-specific surface proteins (n=88) for conventional CAR-T therapy. Categorized and colored by whether the targets are FDA approved (BCMA and CD19), currently in clinical trials, currently in preclinical test, additional viable targets, or merely bystander T cells markers due to bulk RNA-Seq expression measurements.

Self-antigens represent a major source of PC-CAR targets

In our recent work, we established the use of PC-CARs to target low-to-medium mutational tumors by enabling the robust and specific targeting of peptides abundantly presented on HLA molecules(12). Our initial efforts focused on neuroblastoma, where we targeted the developmental transcription factor PHOX2B. This factor is silenced in healthy differentiated tissues but is co-opted by tumor cells to drive tumorigenicity. Encouraged by these results, we sought to expand this therapeutic strategy to a broader range of cancer patients.

To further expand the pool of actionable targets for T-cell based therapies, we applied our analysis to pHLAs by comprehensively characterizing tumor-specific self-antigens that are well-suited for PC-CAR targeting, scoring canonical proteins in our immunopeptidomics datasets by (1) abundance, (2) recurrence, (3) absence in normal tissue RNA and immunopeptidomes, (4) HLA binding affinity, (5) importance in maintaining tumor progression and (6) homogeneity as determined by single cell coverage. We identified 3,635 tumor specific antigens from 284 genes that are highly-enriched in tumors without appreciable level of expression in normal tissues (**Table S4**). In contrast to the mere 88 viable membrane proteins that meet the same criteria, pHLAs can drastically expand the current antigen repertoire. Expression profiling revealed distinct clusters of genes exclusive to specific cancers (**Table S4**). Notably, cancers such as melanoma, colon cancer, and ovarian cancer possess uniquely expressed genes, including well-characterized immunotherapy targets like PRAME, MUC16 and CDH17. (**Figure 3A**). This pattern suggests that tumors expose vulnerabilities by upregulating specific genes, unique to each cancer type. Additionally, we identified a cluster of peptides derived from genes involved in cell cycle and mitosis regulation like CDC6 and CDC45, broadly expressed across multiple tumors (**Figure 3B**). Although cell therapies targeting cell cycle genes alone may affect normal cycling cells, given the role of these genes and broad upregulation in cancers, incorporating them into AND-gated strategies(25,26) with other antigens could represent a viable therapeutic strategy, exploiting a common vulnerability in cancer cells while sparing the proliferative cells residing in normal tissues.

To explore how many cancer patients could potentially benefit from pHLA-centric therapies, including PC-CARs and engineered TCRs, we analyzed gene expression profiles across all cancer patients. Our findings reveal that all the cancers in our study harbor at least one tumor-specific gene capable of generating tumor self-antigens. Even when taking into account HLA restrictions, 12 out of 21 cancer types still achieved more than 99% coverage, demonstrating the broad potential of these therapies. However, some cancers, like mesothelioma and rhabdoid tumors, showed lower patient coverage (**Figure 3C**). This analysis is promising, indicating that a wide population could benefit from PC-CAR therapies, albeit with careful consideration of HLA restrictions to maximize therapeutic application of each antigen.

Based on the criteria we set above, we highlight a prime example in clear cell renal cell carcinoma, the most common form of kidney cancers. We identified a self-antigen DLSRRDVSL originating from tumor-specific gene HAVCR1 and present by a common allele HLA-B*08:01, covering over 15% of the U.S. population. While HAVCR1 is a potential surface protein target, it has been reported to frequently shed its ectodomain from the cell surface, providing a potential evasion mechanism to conventional CAR(27). Thus, we suggest that targeting HAVCR1 presented on HLA using PC-CARs can be a viable therapeutic strategy. The HAVCR1 peptide is absent in normal immunopeptidome data with favorable therapeutic window in both RNA and peptide levels (**Figure 3D**). The peptide's MS/MS spectrum is of high confidence with 10 fragment ions and consecutive y ions to support the evidence on peptide level (**Figure S2A**). Furthermore, AlphaFold2 modeling predicts the binding to the cognate HLA, with key biophysical features that enhance targetability due to two exposed charged residues accessible for salt bridge formation (**Figure 3E**). We validated peptide/HLA binding by refolding the HAVCR1 peptide with HLA-B*08:01 and performed size exclusion chromatography, demonstrating the formation of the HAVCR1/HLA-B*08:01/β2M complex (**Figure S2B, Methods**). Additional tumor self-antigens are highlighted in **Figure 3A** and the full catalogue of tumor-specific self-antigens is available through our web portal, which we encourage investigators to explore. Together, our comprehensive pan-cancer analysis and focused investigation identified new tumor-specific surface proteins for conventional CARs, and underscores the substantial potential of targeting HLA-presented self-antigens across a wide range of cancers using PC-CARs and TCRs.

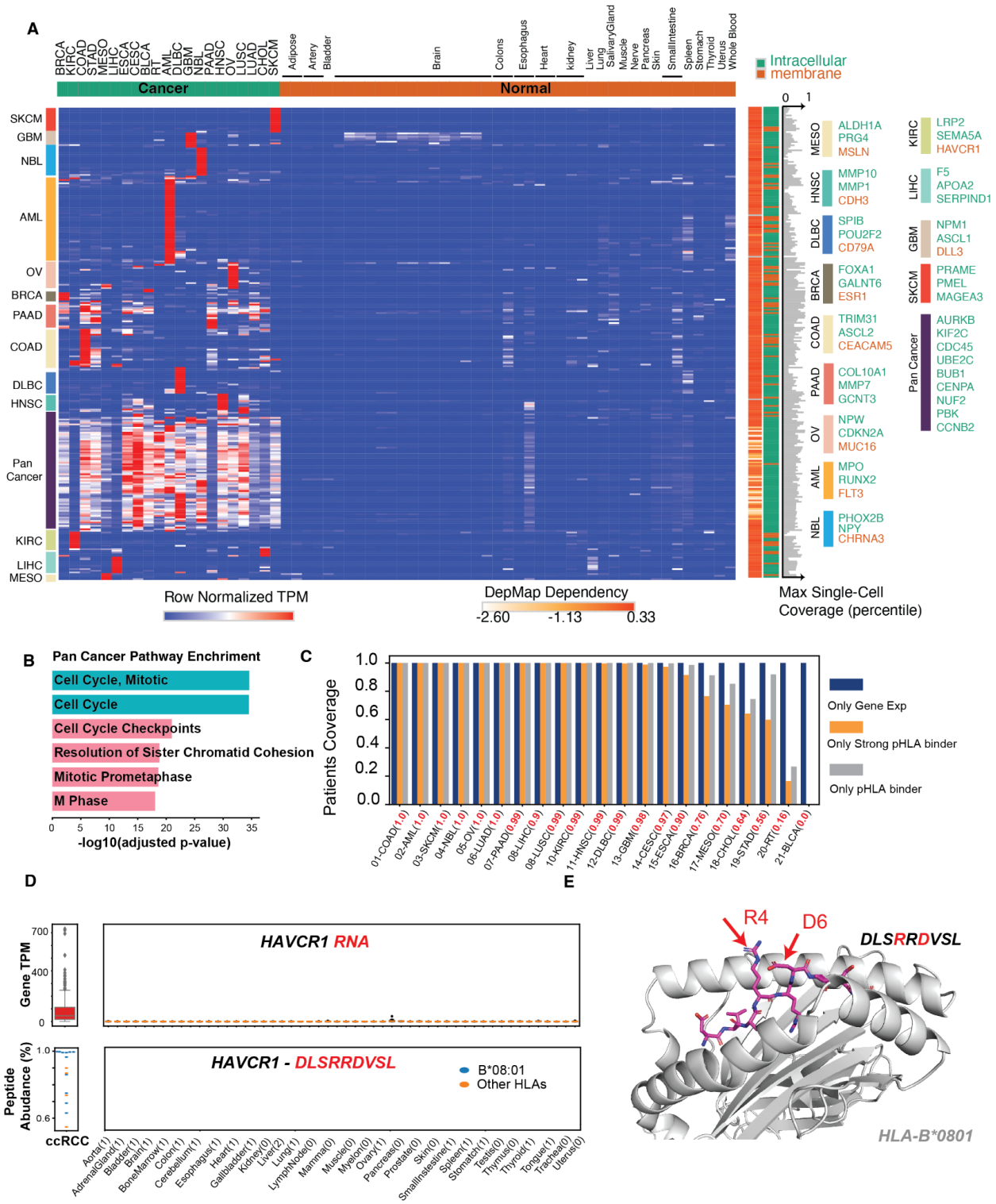


Figure 3. Landscape of tumor specific self-antigen. (A) 284 tumor specific genes that can give rise to tumor-specific antigens. Parent proteins annotated as membrane or intracellular proteins, along with the DepMap reported dependency score (median across all evaluated cell lines) and homogeneity represented as max single cell RNA-Seq coverage (shown in percentile based on ranking all detected genes). Gene expression is depicted by median transcripts per kilobase

million (TPM) in each cancer or tissue. (B) Enriched gene ontology and pathway associated with 68 clustered pan-cancer targets. (C) Patient coverage when evaluated by gene expression alone, and accounting for tumor-cancer specific HLA restrictions, strong binder and weaker binders were evaluated. (D) HAVCR1 shows high differential expression by RNA (top) and peptide abundance (bottom) when comparing between clear cell Renal Cell Carcinoma (ccRCC) and normal tissues (E) Alphafold2 modeling of HAVCR1 peptide and cognate HLA-B*08:01 complex, with exposed two charged residues (R4, D6) highlighted.

Novel mutation-derived neoantigens are revealed from immunopeptidome

While neoantigens arising from somatic mutations and gene fusions have been tested in clinical trials due to their optimal tumor specificity, most antigens are selected based on computational predictions that typically incorporate gene expression and HLA binding affinity. Indeed, only 2.4% of predicted strong pMHC binders are detected within the immunopeptidome of mesothelioma(28). It has also been reported that only 1.6% of mutated antigens demonstrate immunogenicity in gastrointestinal cancer patients(29). Therefore, the vaccine cocktail typically given in cancer vaccines contains a high proportion of peptides that may not be presented by tumors. Moreover, neoantigens may undergo immunoediting at early stages of immune infiltration and can be present only in low amounts(30). For instance, although the KRAS G12D mutation can be detected when overexpressed in engineered cell lines(31) or using high-sensitivity acquisition methods(32), it is not readily identifiable in primary tumors with conventional immunopeptidome acquisition techniques(33) and has largely evaded therapeutic targeting on HLA despite decades of efforts, aside from a case report describing the targeting of this mutation through an infrequent HLA-C allele(34). Furthermore, most mutations and fusions are unique to individual patients, making the development of off-the-shelf therapies challenging.

Consistent with their rare detection in the immunopeptidome, our pan-cancer analysis revealed that SNVs constitute only 0.009% of non-canonical pHLAs on average across tumors (**Figure 4A**). Specifically, lung squamous cell carcinoma (LUSC), skin cutaneous melanoma (SKCM), and colon cancer exhibited the highest numbers of mutation-derived neoantigens, with counts of 223, 92, and 38, respectively, consistent with reported mutational frequencies at genetic level(35) (**Figure 4B**). We note that none of these neoantigens have been documented in the Immune Epitope Database (IEDB). While the IEDB does not catalog all known neoantigens, and it is possible that some have been previously reported elsewhere, this indeed highlights the power of leveraging large-scale genomics and transcriptomics data to discover rare antigens in the immunopeptidome and warrants further investigation of these new uncovered neoantigens. Our analysis confirmed a well-known shared AML neoantigens (AVEEVSLRK) deriving from the NPM1 frameshift mutation. Further, we showed ALK, MUC16, and BRCA2 harbor the highest number of neoantigens evidenced from the immunopeptidome data and that 21% of evidenced neoantigens are derived from known cancer driver genes (**Figure 4C, Table S5**). Furthermore, although gene fusions represent a smaller fraction of neoantigens and are exclusive to individual patients, promising candidates such as UBXN1-BSCL2 and NCOA7-TPD52L1 in breast cancer, and SNX6-CD8B in glioblastoma were identified. (**Table S6**). We show that our pan-cancer molecular catalogues can be used to identify previously unreported neoantigens and suggest that

the empirical detection of these antigens in the immunopeptidome bolsters the rationale for their inclusion in cancer vaccine cocktails or expansion of TILs.

Splicing and Cryptic ORFs yield majority of non-canonical antigens

Recent studies have indicated that splicing alterations and non-coding open reading frames (ORFs) can generate actionable targets across various cancers, potentially yielding targets that are more widely shared across patients(13,15). To investigate this possibility, we evaluated the relative contribution of detected antigens from mutations, fusions, alternative splicing, intron retention, and cryptic ORFs. Our analysis revealed that cryptic ORFs and alternative splicing dominate the antigenic landscape, accounting for approximately 91.1% and 7.0% of the repertoire on average, respectively. Particularly, the cryptic ORFs give rise to a rich source of tumor-specific antigens that are often recurrent across cancers (**Figure 4D, Table S6**). To validate the translational activity of these cryptic ORFs, we utilized ribosomal profiling (Ribo-seq) on eight matched neuroblastoma patient-derived xenograft (PDX) models (**Figure S3A**). This analysis confirms the presence of ribosomes on 37% of these ORF-derived pHLAs, indicating that they are actively translated and are able to generate bona fide peptides on HLA (**Figure S3B**). One notable example is the peptide STIPVLSGY from the downstream untranslated region (UTR) of the TMEM203 gene, which we confirmed by LC-MS/MS of the synthetic peptides and refolding experiments (**Figure S3C-D, Figure S4**). We also identified this tumor-specific antigen in cervical cancer, lung cancer, glioblastoma and melanoma, underscoring its potential as a pan-cancer target.

In examining splicing-derived antigens, we identified a rich source of tumor-specific splicing junctions that were distinct from those in normal tissues(**Figure S5A,C, Table S6**). One particularly interesting finding was the recurrent detection of a splicing variant from PMEL exon 5 skipping, which results in a distinct antigen that retains both HLA-A2 anchor sites shared with two different known PMEL epitopes derived from the canonical proteins (**Figure 4E**). To confirm the validity of our identification, we validated the synthetic peptide by LC-MS/MS and conducted refolding experiment to verify that it is indeed a binder of HLA-A*02:01 (**Figure S5B,E**). We propose that this novel epitope is uniquely recognized by TCRs as compared to the canonical PEML epitopes, highlighting a distinct geometry using structural modeling and showing that two TCR clones against canonical PMEL do not cross-react against the non-canonical peptide (**Figure 4F, Figure S5D**). Despite lower HLA binding affinity, this splicing variant is more abundantly presented on HLA, possibly due to enhanced degradation of short-lived unstable splicing products (**Figure 4G**). We suggest that this non-canonical PMEL antigen should be further investigated and may have superior therapeutic properties than the canonical PMEL antigen currently targeted in the clinic based on its increased abundance and tumor specificity (**Figure 4E**). Additionally, as one subset of aberrant splicing variant, we identified pHLAs derived from retained HERC1 intron 19, MAP3K5 intron 10, and TXNDC15 intron 5, all specific in AML (**Table S7**). The high prevalence of intron retention-derived antigens in AML aligns with previous reports of abundant intron retention aberrations in AML, which have been correlated with mutations in epigenetic modifiers such as DNMT3A(20). Collectively, our findings reveal that tumor-specific antigens derived from alternative splicing events and cryptic open reading frames

constitute the majority of actionable targets across various cancers, with significant potential in expanding immunotherapies.

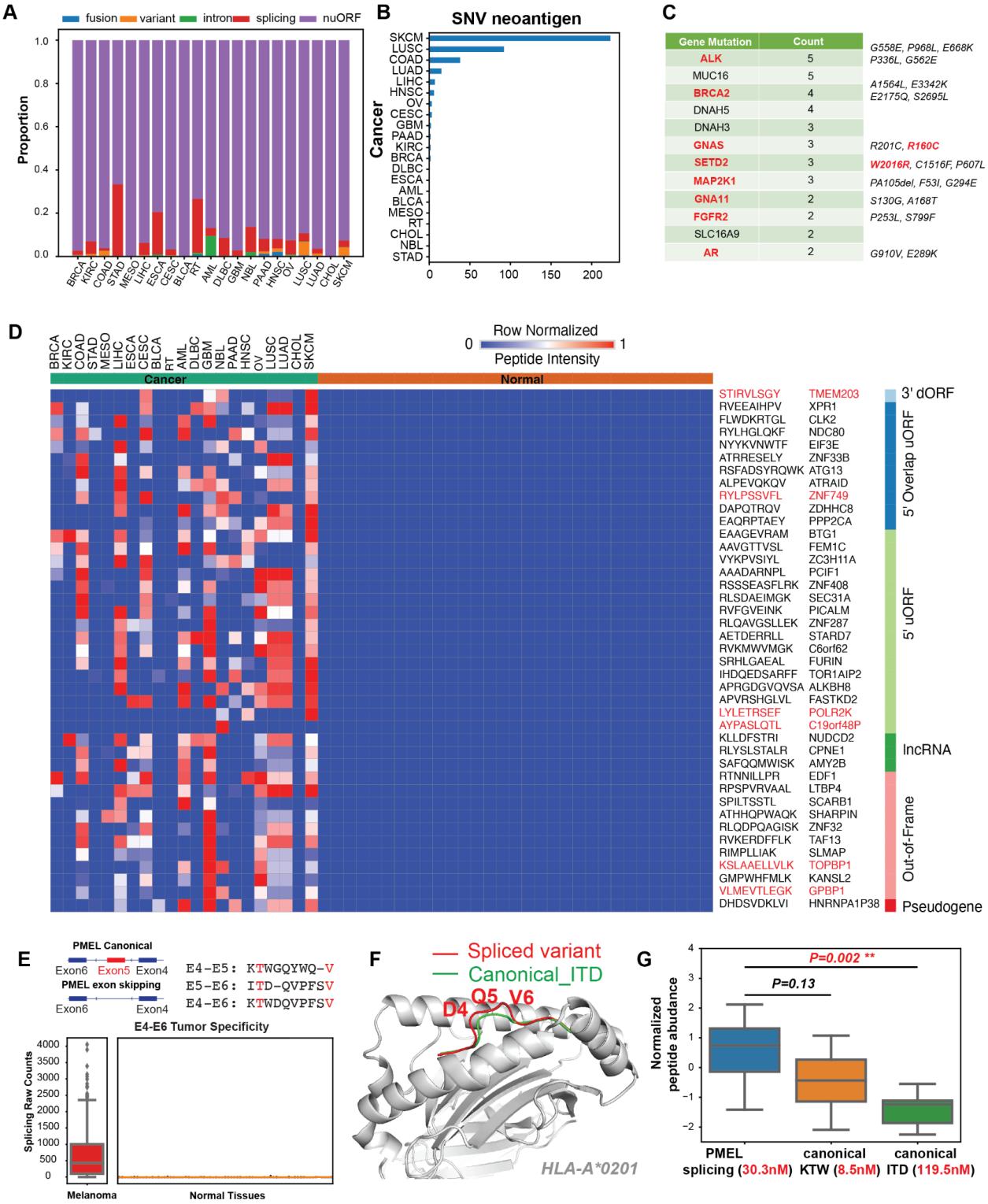


Figure 4. Cryptic ORF and alternative splicing contribute to the majority of antigens. (A) Proportion of non-canonical antigens arising from gene fusion, somatic mutation, splicing, intron

retention and cryptic open reading frames (nuORFs) across tumors (B) Number of immunopeptidome-evidenced neoantigens for each cancer (C) Most frequently detected neoantigens and the number of peptides identified, annotated cancer driver genes are colored coded with the derived mutation locations (positions detected in more than one patient were highlighted) (D) Recurrently detected cryptic ORF derived tumor-specific antigens, the ones colored in red were validated through LC-MS/MS and refolding experiments (E) exon skipping of exon 5 of PMEL results in a splice variant compared to two canonical antigens (top). Supporting RNA-Seq reads of the exon skipping in melanoma versus normal tissues (bottom). (F) Alphafold2 structural modeling of the PMEL splice variant backbone with its cognate HLA-A*02:01 allele, compared to the canonical antigen. Distinct TCR-recognizable residues are highlighted. (G) Normalized peptide abundance of the PMEL spliced variant, and two canonical counterparts, the predicted binding affinity of A0201 were annotated for each peptide.

Tumor resident pathogens provide an unexpectedly rich set of potential therapeutic targets

Viral infections including Hepatitis B Virus (HBV) and Human Papillomavirus (HPV) are well-documented oncogenic viruses directly linked to liver and cervical cancer tumorigenesis. In HBV and HPV, the integration of viral DNA into the host genome and the direct encoding of transforming proteins have been shown to disrupt cellular regulatory networks, inducing oncogenesis through the activation of oncogenes and inhibition of tumor suppressor genes(36). Prophylactic vaccinations against these viruses have proven effective in reducing cancer incidence, suggesting that the immune system is able to clear infected precancerous cells through viral antigens(37,38). We hypothesize that tumor-resident viral pathogens not only contribute to oncogenesis but may also alter the immunopeptidomic landscape by presenting novel antigens, which could reveal new therapeutic targets. To confirm our ability to identify such candidates, our initial studies focused on the immunopeptidome of liver and cervical cancers known to harbor HBV and HPV infections, respectively (**Figure 5A**). Notably, we identified several known viral peptides also residing in primary tumor tissues, such as FLLTRILTI from the HBV large surface antigen, and YMLDLQPET from the HPV protein E7. A comparative analysis with the Immune Epitope Database (IEDB) revealed that 45% of the HBV peptides detected in primary liver tumors were previously documented as immunogenic, highlighting their potential utility as therapeutic vaccine targets (**Figure 5B**). Building on these observations, we further explored other common oncogenic viruses that are linked to cancer. While cytomegalovirus (CMV) infection has been implicated in multiple cancers, and its DNA and protein have been detected in cancers such as neuroblastoma, colon cancer and glioblastoma, it is currently unclear whether CMV infection can generate additional pHLAs suitable for therapeutic development. Strikingly, our pan-cancer analysis identified 158 CMV-derived antigens across 6 different tumor types (**Table S8**). Cross-referencing the CMV peptides identified in primary tumors with the IEDB database revealed seven peptides that have been previously documented in the setting of viral infection. (**Figure 5C**). Particularly, we highlighted the CMV-derived peptide LLDGVTVSL, which was recurrently detected in five cancers (ovarian cancer, breast cancer, neuroblastoma, glioblastoma, cervical cancer) and presented by the common HLA-A*02:01 allele. To the best of our knowledge this is the first time these CMV antigens have been reported in tumors, offering a viable target candidate for various immunotherapeutic modalities (**Figure 5D**).

In addition to oncogenic viruses, we sought to investigate the presence of additional microbiomes, including bacteria and fungi, that could result in novel antigens. Although it is possible to predict pathogens from bulk transcriptome data, the existence of a tumor-specific microbiome remains contentious due to the complexity of distinguishing genuine microbial sequences from host-derived or environmental contaminants(39). This ambiguity is exemplified by the fact that many unmapped reads can be reclassified into the latest reference human genome assembly or may simply be artifacts of contamination, leading to potential false discoveries(40). To address these discrepancies, we established a stringent framework to classify unmapped reads into various species including fungi and bacteria, assigning a read to a specific species only if 90% of its kmers could be unambiguously mapped to that species. Additionally, we excluded species detected in normal tissue to ensure tumor specificity and corroborated our findings with previously published studies(41). Our analysis successfully identified several recurrent strains at the RNA level that also generate peptides in the primary tumor immunopeptidomes, including *H. pylori*, a pathogen associated with gastric cancer(42) and *F. nucleatum*, a well-described pathogen implicated in colon cancer and melanoma(43,44). Aside from these previously described strains, we detected *N. circulans*, a bacteria that has been implicated in sepsis among immunocompromised patients(45), in over 90% of ovarian RNA-Seq samples (**Figure 5A**). Remarkably, we identified 652 unique peptides derived from *N. circulans* originating from 594 unique pathogen genes in ovarian cancers, with 18 of these also consistently detected in esophageal cancer. As immunopeptidomic searches are known to inherently yield false positives, we searched the *N. circulans* proteome against 8 neuroblastoma cell lines as a negative control. None of the 18 peptides were found, suggesting that *N.circulans* was correctly identified at both the RNA and immunopeptidomic level across ovarian tumors (**Figure 5E**). We further validated these peptides by performing LC/MS-MS on the synthetic peptides, confirming the presence of these bacteria peptides in tumors. These findings warrant further investigation into targeting these pathogen-derived peptides as well as the role of bacterial strains in tumorigenesis. Together, we have systematically explored the landscape of pathogen-derived antigens in primary tumors and validated several novel antigens originating from oncogenic viruses or bacteria, highlighting the novel therapeutic potential of these pathogens when presented as tumor HLA peptides.

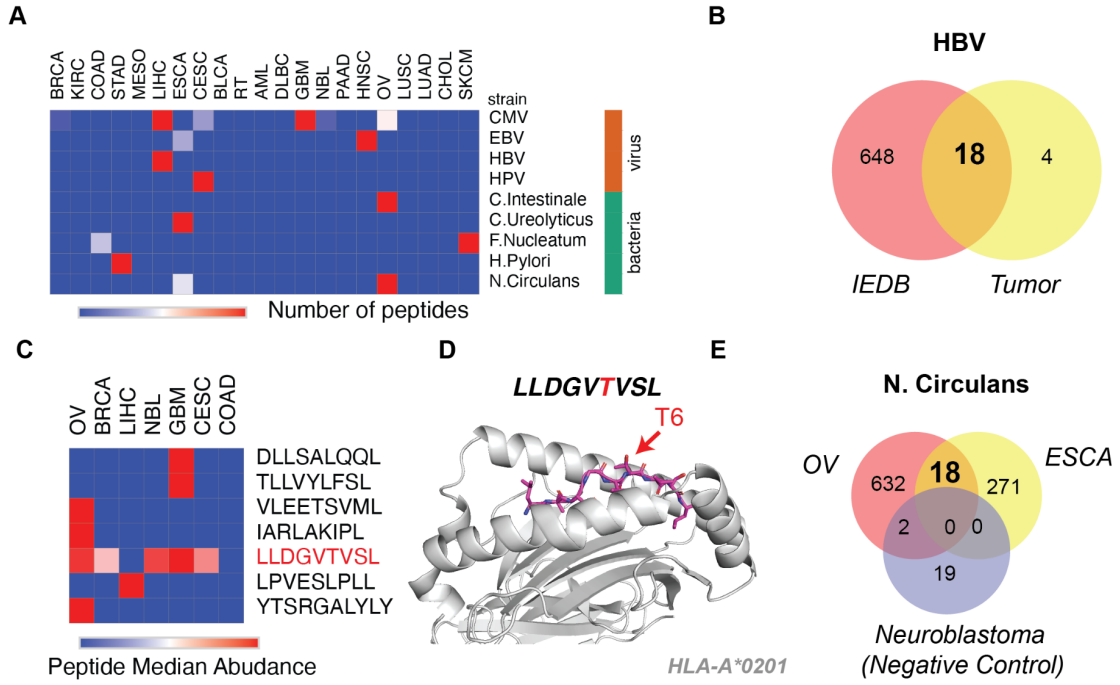


Figure 5. Tumor resident pathogens yield actionable pHLAs. (A) Number of peptides detected in immunopeptidome in each species across cancers. (B) comparison between IEDB documented HBV peptides in viral infection and the ones detected in liver cancer primary tumor immunopeptidome (C) Seven previously documented CMV peptides in IEBD and also detected in primary tumor immunopeptidome, median peptide abundances are depicted across cancers. (D) Alphafold2 modeling of HLA-A*02:01/β2M/LLDGVTVSL complex. (E) Number of detected N. Circulans derived peptides across ovarian and esophageal cancers and neuroblastoma cell line (negative control).

<u>Long interspersed nuclear element-1 (LINE1) ORF2-derived peptides are detected in immunopeptidome</u>

Transposable elements (TEs) can autonomously undergo transcription and subsequent translation, and recent studies have identified them as a novel class of pHLAs(46,47). In healthy somatic tissues, TEs are tightly regulated epigenetically, tending to be silenced. However, this regulation is often lost in cancer, leading to aberrant activation and positioning them as a novel source of tumor-specific antigens. TE expression and retrotransposition is generally associated with poor prognosis, contributing to genomic instability and tumor evolution(48). More recently, immunopeptidomic research has focused on TE-derived chimeric transcripts, such as exonization and their ability to act as alternative promoter(49), splicing donors and acceptors(47,50) in order to expand their therapeutic potential. Notably, a recent pan-cancer analysis of TE-promoter events using RNA-Seq data identified 1,068 widespread tumor-specific aberrations, however, it remains unclear whether such events can generate detectable pHLAs in the primary tumor sample(46). Moreover, certain TEs can encode proteins themselves, including Long Interspersed Nuclear Element (LINE), Long Terminal Repeat (LTR) elements(51), and sporadic reports of Alu, DNA transposons, and SINE-VNTR-Alus (SVA) retrotransposons(52) (**Figure 6A**). We

investigated how various TE categories contribute to tumor-specific pHLAs and characterized their prevalence across multiple cancer types. pHLAs derived from TE chimeras are most abundantly detected in melanoma and lung squamous cell carcinoma, while autonomous-TE-derived pHLAs, which correlates with the highest number of tumor-specific TE elements, are particularly prevalent in AML (**Figure 6B, Table S9**).

LINE-1 ORF2p protein has evaded detection through conventional methods such as western blotting and whole cell proteomics(53). Surprisingly, we detected canonical ORF2 peptides recurrently presented on HLA across nine cancer types, with the highest frequency of detection in colon, head and neck, and lung cancers (**Figure 6C**). We confirmed its presence by synthesizing and validating three peptides recently identified from the immunopeptidome. To our knowledge, this is the first time evidence of ORF2 protein has been detected from endogenous expression. ORF2 is a key player in tumor evolution, encoding an endonuclease and reverse transcriptase to facilitate the retrotransposition of L1 and Alu elements. Through its endonuclease activity it creates genomics instability, induces DNA double strand breaks and leads to large-scale genomic rearrangements and copy number alterations(54). This evidence suggests that ORF2 may be a short-lived protein that is rapidly degraded by the proteasome and presented on HLA before the full-length protein is detectable. Collectively, we assessed the potential of transposable elements to generate tumor-specific pHLAs through various mechanisms, and highlighted a previously elusive ORF2 peptide, suggesting that immunopeptidomic characterization may represent an alternative approach for identifying difficult-to-detect proteins. These findings warrant further investigation into ORF2 biology and its utility as therapeutic targets.

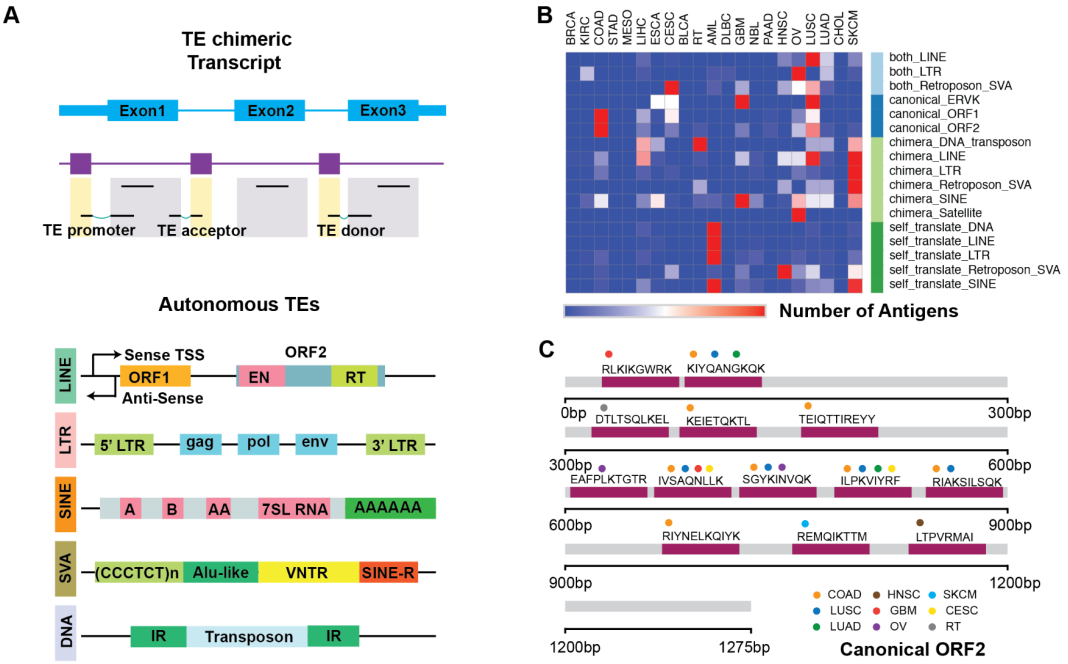


Figure 6 Landscape of Transposable Element (TE) derived antigens (A) schematic overview of diverse potential contributions of TEs to the immunopeptidome, including TE chimeric transcript and autonomous TEs. (B) Tumor-specific pHLAs detected in the immunopeptidome across all

major transposable element (TE) categories in 21 cancer types. This includes antigens from TE chimeras, self-translated TEs, or can be mapped to both a chimeric and autonomous TE element, as well as those mapped to the canonical ERVK and ORF1/2 protein sequences in UniProt (C) Detected peptide fragments from canonical ORF2 in immunopeptidome across cancers.

HLA presentation pathway exhibits distinct expression patterns across cancers

The HLA presentation pathway is a highly orchestrated process involving multiple components. Proteins and peptides are ubiquitinated and subjected to proteasomal degradation, producing shorter peptides with specific C-terminal cleavage sites. Once degraded, the peptides are transported into the endoplasmic reticulum (ER), further trimmed at the N-terminus by ERAP, and loaded onto the corresponding HLA/β2-microglobulin (β2M) with the help of the chaperone proteins. The assembled pHLA complexes are subsequently transported to the cell surface for immune surveillance(55). Given the significant variations in antigen landscape across different cancers, we sought to explore potential intrinsic differences within the HLA presentation pathway itself. Starting with the proteasomal degradation, our analysis identified that DLBC exhibits the highest expression of immunoproteasome signature genes PSMB8/9/10 among all cancers (**Figure 7A**). The immunoproteasome is characterized by its unique beta subunits and altered proteolytic preferences to peptides with hydrophobic and basic C-terminus, which results in residues with more favorable binding to the F-pocket of HLA molecules(56). This observation aligns with the primary role of B cells as antigen-presenting cells in this context. We find only 1.2% of eluted peptides in DLBC eluted peptides end with acidic residues (aspartic acid and glutamic acid), compared to 8.1% of acidic C-terminus peptides in Rhabdoid Tumor (RT), in which both has similar number of eluted total peptides, suggesting reduced propensity for acidic residues at the C-terminus in DLBC (p=0.0006, chi-square independent test). Further, unbiased clustering analysis revealed the lack of acidic C-terminus motifs in DLBC compared to RT, supporting the role of immunoproteasome in altering the DLBC antigen landscapes (**Figure 7B-C, Methods**).

Further investigation into the TAP transporter genes revealed that TAP1 is upregulated in AML, HNSC, and CESC, possibly indicating an enhanced immune response for presenting antigens on the cell surface, which is corroborated by pro-inflammatory genes (CXCL8, CXCL10, NFKB1) and IFN-γ signaling pathway (IFNGR). Conversely, TAP2 expression remains largely unchanged across cancers. Additionally, the peptidases ERAP1 and ERAP2, which are responsible for trimming the N-terminus of peptides, generally do not show high expression in tumors, aligning with previous findings(57) (**Figure 7A, Table S10**). This suggests that cancer cells typically maintain standard trimming activity. In AML, we observed elevated expression levels of both ERAP1 and ERAP2. However, the downstream effects are highly contingent on the specific ERAP haplotypes, which can exhibit hyperfunctional, normal, or hypofunctional trimming activity(58). Amongst the genes encoding components of the peptide loading complex, including the chaperones TAPBP, TAPBPL, CALR, CANX, and PDIA3, we observed generally higher expression levels in cancers compared to normal tissues, likely also reflecting the upregulation of the HLA presentation pathway in response to immune activation and inflammatory signaling in the tumor microenvironment. Interestingly, thyroid exhibits the highest expression levels of chaperone proteins in normal tissues, potentially attributable to the increased abundance of ERs given its intensive hormone production and secretion activities. Lastly, our findings indicate

elevated expression of HLA-A and β2M in DLBC and KIRC, contrasted with generally low expression of HLA-B and HLA-C across cancers. While HLA expressions are generally upregulated upon IFN-γ stimulation, the slight discrepancies for HLA-B and HLA-C may indicate allele-specific regulations and responses(59).

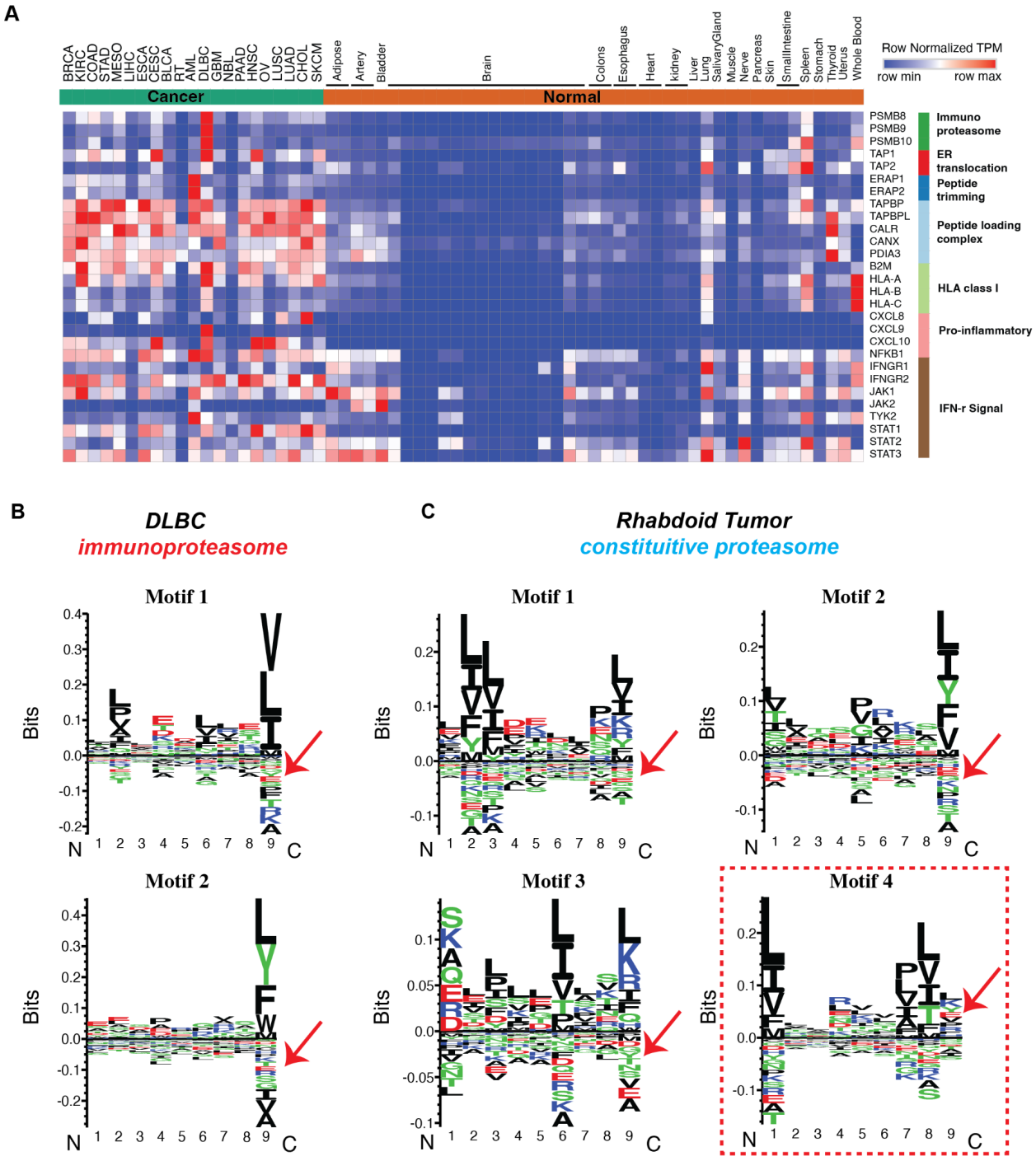


Figure 7 HLA presentation pathway and preferences across cancers. (A) Key genes related in HLA presentation pathway and their gene expression across cancers and normal. (B-C) Motifs of eluted peptides from DLBC immunopeptidome and rhabdoid tumor (RT) immunopeptidome by

Gibbs Clustering, optimal number of clusters are determined by maximum Kullback–Leibler divergence (n=2 in DLBC, n=4 in RT), acidic residues at last positions were highlighted in each motif, the acidic motifs uniquely revealed in RT was highlighted in dashed red box.

Discussion

Here we present the first comprehensive map of tumor-specific T cell antigens across human cancers, revealing an unexpectedly diverse landscape of targetable epitopes derived from both canonical and non-canonical sources. Although some individual molecular classes investigated here have been independently reported to generate pHLAs(60), to our knowledge, this is the first study to simultaneously analyze up to 11 classes of aberrations, providing a full depiction of the T-cell antigen landscape. We provide a deep characterization of the "dark matter" of the immunopeptidome, revealing rich sources of T-cell-based immunotherapy target candidates. We developed a multimodal approach to catalog non-canonical tumor-specific alterations by leveraging large-scale, histology-matched cancer RNA-Seq data and immunopeptidomics. This enabled us to construct a detailed and exhaustive search space for potential pHLAs. Furthermore, our study offered a framework to the unresolved challenge of setting rigorous standards for defining actionable antigens in a therapeutic context. In this respect, we established a robust framework for prioritizing high-confidence pHLAs using six overarching criteria: (1) peptide abundance, (2) peptide recurrence, (3) absence in normal tissue RNA and immunopeptidomes, (4) HLA binding affinity, (5) importance of parent genes in tumor progression, and (6) homogeneity of target expression as measured by single-cell coverage. Although current technologies and data availability limit our ability to apply all six criteria to certain non-canonical antigen classes, our schema provides a roadmap for future investigations, helping tailor analyses to specific antigen types. Additionally, the search space we designed facilitates the integration of users' own immunopeptidome data, and is able to search for all classes of antigens even when matched RNA-Seq data is unavailable or lacks statistical power due to small sample sizes, enabling users to comprehensively characterize their immunopeptidome datasets. Our computational pipeline can also be applied to any in-house short-read RNA-Seq or immunopeptidome data to expand its usage beyond the tumor types considered in the study.

Our analysis identified numerous promising therapeutic targets. We uncovered 47 additional tumor-specific surface proteins with suitable therapeutic windows for conventional CAR targeting. Amongst the 15,079 tumor-specific pHLAs, we highlighted a highly recurrent, abundant, and tumor-specific HAVCR1-derived peptide in clear cell renal cell carcinoma, which can be harnessed to develop PC-CAR as an alternative strategy to avoid the ectodomain shedding for conventional CAR. We also described a spliced variant of the well-characterized melanoma-specific protein PMEL with superior abundance, suitable for exploitation via both TCR-T and PC-CAR therapies. In addition, we discovered dozens of pan-cancer cryptic-ORF-derived tumor-specific antigens that could serve as new targets for PC-CAR therapy and may be particularly important targets in the immunologically cold tumors where conventional neoantigens have been insufficient. Our immunopeptidome-guided search for SNV-derived neoantigens revealed numerous novel neoantigens in melanoma, lung, and head and neck cancers. The empirical evidence from immunopeptidomics for these neoantigens supports their inclusion in the current vaccine cocktail, augmenting current epitope selection methods that solely rely on predictive

modeling. Moreover, we revealed a rich set of tumor-resident pathogen-derived antigens spanning both oncogenic viruses and bacteria, yielding new targets from both known and newly identified pathogens. We highlighted recurrent CMV-derived peptides presented across tumors and new peptides derived from known tumorigenic pathogens such as H. pylori that can provide potential therapeutic avenues against tumors. A previously underexplored species, *N. circulans*, is unexpectedly detected in over 90% of ovarian cancer patients, warranting investigation into the biological role of this species in ovarian tumors and representing numerous high-confidence new candidates for therapeutic targeting.

Our analysis also reveals novel biological insights into the HLA presentation pathway. We demonstrate preferential presentation of defective and short-lived peptides such as splicing products and cryptic ORFs. These findings, along with our identification of previously elusive ORF2 peptides from the LINE1 element, are consistent with the hypothesis that Defective Ribosomal Products (DRiPs) may be overrepresented in HLA-presented peptide repertoire(61), highlighting the diverse and complex nature of antigen presentation and suggest that immunopeptidomics could be used as an alternative strategy to identify difficult-to-detect proteins. We further observed variability in the intrinsic HLA presentation pathways across tumors, modulated by inflammatory signaling, which may influence peptide repertoire properties. While we highlighted several biological insights that are gleaned from these resources, we expect that these datasets can provide numerous new avenues of investigation.

While our toolkit is specifically designed for identifying tumor antigens by HLA class I molecule, our tools can also be adapted into other settings. Our pipeline can be applied to identify HLA-II antigens on tumor surface for immune surveillance, which are expressed in many cancer types(62). Furthermore, HLA-I antigens presented by antigen-presenting cells (APCs) play a critical role in priming T cells within lymph nodes and in tertiary lymphoid structures. When RNA-Seq data from sorted APC populations is available, our tool has the potential to identify antigens that are more likely to elicit a reactive immune response and can be used to study T cell priming and antigen presentation in other disease contexts. This approach can also be used to study the contribution of various antigen classes to central tolerance by studying the repertoires of non-canonical HLA-presented peptides in thymic cells. Moreover, while we utilized our molecular catalogue of tumor-specific events specifically for immunopeptidomic searches, this search space can also be used to comprehensively characterize bulk proteomic datasets, and the catalog of tumor-specific genetic events that we curated to perform this study can also be utilized to study the interplay of various biological processes in tumor. Lastly, our pipeline is highly adaptable, allowing the integration of additional modalities such as whole-genome sequencing (WGS) and Ribo-Seq, as well as new molecular classes, including circular RNAs and RNA editing.

We acknowledge several limitations in the current study. We highlight that the availability of immunopeptidome data varies across cancer types; for instance, in bladder cancer, certain tumor-specific genes like VGLL1 are not identified in the current immunopeptidomic data, potentially due to limited HLA alleles in the public study. We anticipate that expanding the immunopeptidomic data pool in future studies will help bridge existing knowledge gaps. Moreover, while our analysis has delineated a cancer-specific search space, a more granular examination of individual tumor

subtypes could uncover additional vulnerabilities that broader cancer-specific analyses might overlook. Additionally, as our atlas is primarily based on bulk sequencing data, the results presented in the portal should be interpreted in the biological context, considering the potential presence of contaminating stromal or immune cells in both normal and cancerous tissues. Besides the 11 molecular classes, other mechanisms such as antigens arising from BCR hypermutation and VDJ recombination, particularly pertinent to blood cancers like B cell lymphoma, may represent another therapeutic avenue given advancements in high-accuracy BCR sequencing the tumor specificity of BCRs. Additionally, alterations deriving from proteasomal splicing(63), post-translational modifications (PTMs)(16) and alternative codon(64) can to generate pHLAs. The challenges for targeting these are two-fold. First, since these events can arise in any place and in random combinations, enumerating all possibilities has been computationally prohibitive. Second, defining tumor specificity of post-translational events is non-trivial, as these events are poorly cataloged in normal tissues and may also occur under various stress conditions. Future research dedicated to rigorously evaluating the therapeutic value of these novel classes of antigens could further significantly expand the repertoire of tumor antigens.

Recent findings indicate that, while tumor mutation burden (TMB) is widely recognized as a predictor of cancer immunotherapy outcomes, it is often inconsistent as a sole predictor of responses to immune checkpoint inhibitors(65). Given the vast number of non-canonical antigens beyond SNV-derived neoantigens, our study emphasizes that a more comprehensive tumor antigen burden might provide a more effective predictor for the prognosis and correlative analysis. Overall, our work addresses the critical bottleneck in immunotherapy development, the identification of tumor-specific targets. We find that at least one tumor-specific pHLA antigen is present in 86% of the tumors we characterized, a substantial increase in the cancer patients that could benefit from immunotherapies compared to the roughly one-third who are good candidates for currently available immunotherapies. We have released the tumor antigen atlas as a web portal and the immunopeptidomics search algorithm as a cloud-based pipeline, providing the scientific community with a foundational resource that can transform future immunopeptidome analyses. From this dataset, we highlight several examples of promising therapeutic avenues for targeting tumors through membrane proteins, differentially expressed pHLAs, tumor-specific splicing events, TEs and through peptides derived from tumor-specific pathogens. We anticipate that these findings and resources will provide a roadmap that addresses the critical barrier of identifying tumor-specific T cell targets, catalyzing the development of new immunotherapies that offer safe and effective new options to a broad range of cancer patients.

Acknowledgement
We would like to thank Drs. Alvin Farrel and John M. Maris for providing the neuroblastoma PDX Ribo-Seq data. We also thank Drs. Rowan Beck and Cera Fisher from Velsera to provide technical support for running the pipeline on the Cancer Genome Cloud (CGC) platform. We thank Drs. Ioana Clotea and Beatrix M. Ueberheide from NYU proteome core for the antigen spike-in validation. We thank Drs. Shohei Koide and Akiko Koide for sharing the HLA and β2M construct. We thank Dr. Ritchlynn Aranha for the valuable inputs on the HLA presentation pathway. This work was delivered as part of the NexTGen team supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (CGCATF-2021/100038 to M.Y) and the National

Cancer Institute (CA290738-01 to M.Y.), NIH Director's New Innovator Award (1DP2CA301080 to M.Y.), Damon Runyon-Rachleff Innovation Award (86-25 to M.Y.), and Cancer Genome Cloud (CGC) collaborative fund (G.L and M.Y.).

Data Availability

The scripts used to generate the figures are available at https://github.com/frankligy/pan_cancer_intracellular_antigen_atlas, and the pipeline for generating the search space is hosted on the Cancer Genome Cloud at https://cgc.sbgenomics.com/u/li2g2uc/deploy-pipelines.

Contributions

G.L. and M.Y. conceived and designed the study. G.L. developed the computational pipeline, constructed the atlas, and conducted all bioinformatics analyses. O.U.G. carried out peptide spike-in validation and refolding experiments. A.S. assisted in analyzing the normal immunopeptidome data and refolding experiments. G.L, M.D., R.C.C and M.Y. designed the experiments. E.T., A.N., and M.K. performed the PMEL TCR cross-reactivity assays. X.L. contributed to modeling the protein structure. Z.H. performed single-cell analysis for self-antigens. D.T. provided critical insights into the analysis of Transposable Elements. R.Y. and M.D., R.C.C. executed the PMEL validation experiments. G.L. and M.Y. wrote the manuscript.

Conflict of Interests

M.Y. is a co-founder and holds equity in HuLA Therapeutics. G.L and M.Y. have filed provisional patents for the bioinformatics pipeline and the targets discovered in the paper.

Method

Collection of public bulk RNA-Seq

The tumor bulk RNA-Seq data was collected from TCGA (phs000178.v11) and TARGET (phs000218.v26) upon dbGaP approval through Cancer Genome Cloud (CGC) platform which deposited uniformly processed BAM files aligned to Hg38 human genome assembly using STAR 2-pass mode. 21 adult and pediatric tumor types were selected based on the availability of matched immunopeptidome public datasets, particularly, they are Breast Invasive Carcinoma (BRCA), Kidney renal clear cell carcinoma (KIRC), Colon adenocarcinoma (COAD), Stomach adenocarcinoma (STAD), Mesothelioma (MESO), Liver hepatocellular carcinoma (LIHC), Esophageal carcinoma (ESCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Bladder Urothelial Carcinoma (BLCA), Rhabdoid Tumor (RT), Acute Myeloid Leukemia (AML), Diffuse Large B-Cell lymphoma (DLBC), Glioblastoma multiforme (GBM), Neuroblastoma (NBL), Pancreatic adenocarcinoma (PAAD), Head and Neck squamous cell carcinoma (HNSC), Ovarian serous cystadenocarcinoma (OV), Lung squamous cell carcinoma (LUSC), Lung adenocarcinoma (LUAD), Cholangiocarcinoma (CHOL), Skin Cutaneous Melanoma (SKCM). The normal tissue bulk RNA-Seq raw data was downloaded from NCBI SRA upon dbGap approval (phs000424.v6.p1). The normal tissue gene expression data (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz) was retrieved from the

GTEx portal (https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression) as pre-calculated Transcripts Per Millions (TPMs) values.

Collection of public immunopeptidome data

Publicly available immunopeptidome data sets were manually curated and downloaded from both PRIDE (https://www.ebi.ac.uk/pride/) and Massive platform (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp). The biological context associated with each immunopeptidome experiment was retrieved from the original publications, along with the reported HLA alleles wherever possible. Normal tissue immunopeptidome raw data, sample annotations and HLA alleles were downloaded from HLA ligand atlas with the accession number PXD019643.

Collection of public tumor single-cell RNA-Seq data

Publicly available tumor single-cell RNA-Seq for 14 cancer types (thyroid carcinoma, colon cancer, gastric cancer, bladder cancer, head and neck cancer, pancreatic adenocarcinoma, ovarian cancer, lung squamous cell carcinoma, lung adenocarcinoma, glioblastoma, sarcoma, rhabdoid tumor, Uterine corpus endometrial carcinoma, breast cancer) were downloaded from Cancer Single-Cell Expression map (CancerSCEM v1.0)(66) (https://ngdc.cncb.ac.cn/cancerscem/). Uniformly processed h5ad files were downloaded and the malignant cells were selected based on marker genes expression (EPCAM, FOLH1, KLK3, KRT8, KRT18, KRT19) as documented by CancerSCEM official tutorial. The average gene expression across malignant cells was calculated, ranked (in ascending order) and the percentile of each gene across all detected genes were reported as the single-cell coverage (homogeneity) in the analysis.

Computational Pipeline to identify diverse tumor-specific molecular events

A comprehensive computational pipeline was developed to profile 11 classes of molecular aberrations in the tumor, filtered by normal tissue presence, and assembled into sample-specific or tumor-specific peptide search space for downstream immunopeptidome analysis. The detailed descriptions of each module were as following:

(1) Self-gene: For any user-supplied data, the raw RNA-Seq fastq files were subject to Kallisto (version 0.44.0) pipeline(67) using transcriptome index generated from the human Gencode V36. Transcripts Per Millions (TPMs) for each gene were calculated by summing up all transcript-level TPMs associated with each gene. The normal tissue safety profiles for each gene were retrieved from BayesTS database (https://github.com/frankligy/BayesTS/blob/main/database/full_results_XY_essential_tissues.txt) based on multiple evidence including normal tissue gene expression, protein staining and essential tissue distributions(3). The median TPM in each normal tissue is calculated, and the maximum of them serves as a proxy for normal tissue expression for filtering purposes. Next, A gene is defined as tumor specific only if met following criteria: (a) median TPM in tumor is greater than 20, (b) median TPM in tumor should be greater than the normal expression calculated as above, (c) BayesTS score is less than 0.3 to ensure safety of targeting it therapeutically, (d) significantly upregulated in the tumor

versus histology-matched normal tissue when the differential gene analysis data is available. For the TCGA and TARGET data, we retrieved the gene expression TPMs from Xena Browser (https://xenabrowser.net/datapages) for each cancer, the differentially gene expression between tumor versus histology-matched controls were retrieved from GEPIA(68) API (http://gepia.cancer-pku.cn/detail.php), and the genes with adjusted p-value less than 0.05 and log fold change (LFC) greater than 0.58 were defined as significantly upregulated genes in each tumor. The protein sequence of canonical self-genes and isoforms were downloaded from Ensembl and Uniprot, respectively. We note slight differences between ensembl-annotated main transcript and Uniprot-annotated main transcript for a subset of genes, thus utilizing the references from both sources can cover all the reviewed protein coding isoforms. Using the tumor gene expression, we filtered out proteins and isoforms that are not expressed in the tumor sample., and only the expressed proteins and isoforms were assembled as either sample or tumor-specific search space in fasta format.

(2) Alternative Splicing: For user-supplied RNA-Seq raw data, the fastq files were first aligned to human hg38 reference using STAR 2-pass mode to generate the BAM file(69). pysam package (https://github.com/pysam-developers/pysam) was utilized to identify all the junction reads (spanning the introns based on the N in the CIGAR string) and quantify the read counts that support each splicing junction (find_introns function). The comprehensive GTEx normal tissue splicing junctions profiles were downloaded from SNAF database (http://altanalyze.org/SNAF/download.tar.gz)(13). A splicing junction was defined as tumor-specific only if met following criteria: (a) average read count in tumor is greater than 10, (b) average read count in all normal tissues is less than 1, (c) the log fold change between tumor and normal tissues is greater than 4. To further remove potential alignment error from genuine splicing events, splicing junctions with no annotated splicing sites will be discarded. For TCGA and TARGET data, pysam packages were applied on the CGC uniformed processed BAM files from STAR 2-pass mode. For the pan-cancer analysis, since the RNA-Seq and immunopeptidome are not matched by exact sample but only by histology, frequently detected splicing junctions (frequency > 20%) were retained as they should be considered as shared events in each tumor type instead of specific sample. The tumor-specific splicing junctions and the flanking sequences (n=33nt) were further subject to *in silico* translations, potential splicing peptides were defined as ones without stop codon and must contain at least one residue deriving from the splicing junction. When the tumor-specific splicing junction coincides with annotated isoforms in Uniprot, the whole documented isoform sequences were used instead. All the tumor-specific isoforms and tumor-specific local junction derided peptides were assembled into sample-specific or tumor-specific search space.

(3) Intron Retention: The intron regions were defined based on previously described exon-intron model(13) by first collapsing all the annotated isoforms in Ensembl associated with each gene, only regions that are consistently annotated as intron across all isoforms will be considered. Introns that are overlapping with exons from another gene were also discarded because the reads coverage likely come from the exons not the introns. Next,

Stringtie2 (version 2.1.3) was run in de novo mode(70) to predict all the potential novel transcripts including the ones with intron retentions. When a defined intron was supported by a predicted read-through Stringtie transcript, it was considered as an intron-retention candidate, and the expression (TPM) of the exon covering the intron regions was utilized as a proxy for the expression of that intron retention event. Exact same pipeline was applied to the GTEx normal samples across 31 histologies. We randomly subsample 5 normal samples from each histology as the representative intron retention profiles in each tissue. Only intron retentions that are absolutely absent in normal tissue were retained as tumor-specific intron intentions. We further utilized the previously reported SNAF database for intron retention detected from GTEx to make sure all the intron retention considered are not present in the normal tissues. For pan-cancer analysis, same as the splicing event, we defined frequently identified intron retention events (frequency > 15%) in the tumor RNA-Seq cohort to include in the immunopeptidome search. Next, we utilized ORFfinder program downloaded from NCBI (https://www.ncbi.nlm.nih.gov/orffinder/) in default setting to predict all the potential ORFs that can be generated from the intron-retained transcript, and the most probable ORF was ranked primarily by the length of the ORF. When the length difference between the best-scored ORF and the second-best scored ORFs is less than 8nt, we further ranked them based on the GC content and the codon potential in humans, as described before(13). The most probable ORF was identified and translated into peptide. Lastly, only peptides that overlap with the intron region of interest were considered as the causal peptides originating from the intron and were reported in the sample-specific or tumor-specific search space.

(4) Tumor-resident microbiome: The unmapped reads in the tumor BAM files were collected using samtools(71) and reverted back to the fastq format using bedtools(72). The fastq files were then subject to the kraken2 program(73) to be classified into hierarchical taxonomies, using a comprehensive indexed database (PlusPF) downloaded from (https://benlangmead.github.io/aws-indexes/k2), which was constructed using archaea, bacteria, virus, plasmid, human univec_core, protozoa and fungi reference sequences. While there is no consensus cutoff for kranken2 program and is heavily dependent on the specific use case, to mitigate the false discoveries rate, we adopted a stringent cutoff (confidence=0.9) such that a read would only be assigned to taxonomy if over 90% of the consecutive kmers can be unambiguously mapped to it. Similar to the intron retention, a representative normal cohort was compiled and the exact same pipeline was applied to remove pathogens detected in normal tissuel. Tumor-specific microbiomes were defined as species that are not detected at species level in normal control. For the TCGA data, the frequently detected microbiomes (frequency > 20%) were selected, and were further cross-referenced with the published TCMbio atlas(41) that utilized the same datasets with dedicated quality control steps to remove potential contaminating reads and a bayesian inference step to distribute ambiguous reads mapped to parental taxonomies, potentially leading to more accurate quantifications(74). Microbiomes that are recurrently detected in both two procedures were selected for histology-matched immunopeptidome search. The reference proteomes for the selected microbiome were retrieved from Uniprot, when multiple reference proteomes were available, we compared the completeness of the

proteomes and selected the most complete one. The microbiome search space was further appended by the literature-supported tumor-resident virus, or when the immunopeptidome datasets were known to harbor additional microbiomes.

(5) Transposable element (TE)-chimeric transcript: The exonization of TE into a spliced transcripts can manifest through either TE-promoter, or TE splicing site donor/acceptor. Firstly, the 8 prebuilt TE locus index for human hg39 reference were downloaded from TElocus(75) database (https://labshare.cshl.edu/shares/mhammelllab/www-data/TElocal/prebuilt_indices), along with the known genes from hg38 downloaded from UCSC (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/genes/) and the chromosome coordinate annotation table (https://labshare.cshl.edu/shares/mhammelllab/www-data/TElocal/annotation_tables/). Following the previously reported procedures(46,47), the tumor-specific splicing junctions defined above were further annotated based on whether the splicing sites fall into annotated TE regions, and the TE regions following the splicing site till the end were included in the spliced transcript. Both sense and antisense strands of each TE region were considered for TE-chimeric transcripts. Similar to the *in silico* translation protocol described for local splicing junctions, TE-chimeric regions including both the canonical exons and the TE regions were translated and assembled into sample-specific or tumor-specific search space.

(6) Autonomous transposable element (TE): Similar to TE-chimeric transcript pipeline, annotated TE regions were quantified by TElocus program. Due to the nature of highly repetitive sequences associated with TE, TElocus efficiently utilized both uniquely mapped and multi-mapped reads to re-distribute the RNA-Seq reads previously mapped to the gene-only reference. The TE-locus reported raw counts were further normalized by sample-specific sequencing depth into Counts Per Millions (CPMs), as described before(76). Similar to intron retention and tumor-resident microbiome analysis, normal tissue cohorts were assembled to help filter out non-tumor-specific TEs. An autonomous TE was defined as tumor specific when the median CPM in tumor is greater than 1 and the max CPM in normal tissue is less than 0.5, together with the log fold change greater than 5. These cutoffs were selected based on empirical observation between tumor and normal TE expressions, and the resultant TE events following this protocol often possess desirable therapeutic windows. For TCGA and TARGET analysis, detection over 20% of tumor patients from RNA-Seq were added for the immunopeptidome search. Amongst all the TE classes and TE families, the Long Terminal Repeats (LTRs) including Human Endogenous Retrovirus (HERVs), Long Interspersed Nuclear Element (LINE), Short Interspersed Nuclear Element (SINE), DNA transposon and SVA family from retroposon class were selected as potential translatable TEs. Both sense and antisense directions were considered for *in silico* translation.

(7) Gene fusion: The user-supplied RNA-Seq fastq files were analyzed by STARfusion(77) (version 1.13.0) packaged as a docker container downloadable from (https://data.broadinstitute.org/Trinity/CTAT_SINGULARITY/STAR-Fusion/). The

STARfusion database was downloaded from (https://data.broadinstitute.org/Trinity/CTAT_RESOURCE_LIB/GRCh38_gencode_v44_CTAT_lib_Oct292023.plug-n-play.tar.gz), and was executed in the default setting documented on the official tutorials. The reported in-frame fusion transcripts that are not labelled as (1) GTEx_recurrent_StarF2019, (2) BodyMap, (3) DGD_PARALOGS, (4) HGNC_GENEFAM, (5) Greger_Normal, (6) Babiceanu_Normal, (7) ConjoinG were considered tumor-specific fusion events, and the fusion peptides were used as the sample or tumor-specific search space. For TCGA data, the previously curated gene fusions were retrieved from ChimerDB (https://www.kobic.re.kr/chimerdb/) with both detected fusions from different softwares(78). The chromosomal coordinates of each gene fusion were extracted from the downloaded fusion result files and were further translated to fusion peptides. Fusions detected in TCGA matched paratumor samples were filtered out due to their suspicious tumor-specificities. For TARGET data, the CGC uniformly processed gene fusion files from STARfusion were downloaded. Following the same procedures, chromosomal coordinates of the fusion breakpoints were utilized to retrieve the fusion flanking nucleotides sequence and to derive the fusion junction peptides.

(8) Variants and somatic mutation: For user-supplied RNA-Seq raw data, one of the recommended RNA variant detection protocols that achieves superior performance in the benchmark were used to identify variants from RNA-Seq reads(79). Specifically, Opossum(79) was applied to the raw BAM files to conduct a series of quality controls, rendering the resultant BAM files suitable for variant caller. It includes (a) removing low quality and improperly-aligned reads, (b) deduplication, (c) collapsing overlapping reads, (d) separating intron-spanning reads into multiple segments. All the parameters were left as default except the MinFlankStart and MinFlankEnd were set as 10 to remove spurious bases at the start or end of aligning reads, a common artifact in calling variants from genomics and transcriptomics data(80). Subsequently, the modified BAM files were analyzed by Platypus(81), a haplotype-based variant caller to report high-confidence variants in the tumor samples. Variant Effect Predictor (VEP) was applied to characterize the impact of each variant on protein coding regions(82). The missense variant, inframe deletion and insertion, along with frameshift variants were retained. While definitely calling somatic mutations from tumor samples alone without the sample-matched normal control is impossible, certain predictors have been proposed that can achieve reasonable prediction power to distinguish somatic mutation from germline polymorphisms(83). A variant was predicted as likely somatic mutation if met all following criteria: (a) absence or extremely low population frequency (<0.0001) in dbSNP, (b) absence in normal tissue RNA editing database REDIportal(84), (c) maximum variant allele frequencies in tumor samples less than 0.95. When a variant is confirmed as somatic mutation from the external COSMIC database, it was retained as well(85). While these variants were annotated as likely somatic mutations, all the variants will be utilized to construct the search space as germline polymorphism also directly affects the protein sequence and should be considered. Additionally, all the detected A to G variants were annotated as likely RNA editing events in the final results. For TCGA data, somatic mutations called from Whole Exome/Genome Sequencing with matched normal samples were downloaded from Xena

Browser (https://xenabrowser.net/datapages/). For TARGET data, the uniformly processed somatic mutation vcf files were downloaded from CGC platforms. Due to lack of mutation data in TARGET for rhabdoid tumor, a recently published rhabdoid tumor WGS study(86) was used instead to impute the somatic mutation burden for this tumor type.

(9) lncRNA, (10) Pseudogene and (11) cryptic open-reading frames (ORFs): Since RNA-Seq usually lack the capacity to comprehensively quantify the underlying lncRNA and pseudogene expression levels, due to suboptimal capture protocol. A Ribo-Seq informed database nuORF(15) (version 1.0) was used as a more comprehensive catalogue for likely translatable lncRNA, pseudogene and unannotated open-reading frames either residing in the 3' /5' untranslated regions or overlapping with annotated ORF but adopting a different reading frames. Using the provided hg19 transcripts and chromosomal coordinates information in the original nuORF database, we first rederived all the ORF peptide sequences and assured the correct length compared to the original paper. These nuORF peptide databases were appended to all the immunopeptidome searches to enable the identification of these three molecular sources.

Immunopeptidome data Analysis

The exhaustive tumor-specific search space was generated by incorporating all tumor-specific molecular aberrations to search the histology-matched large-scale immunopeptidome datasets using MaxQuant(87) (version 2.4.9.0). MaxQuant was run in nonspecific enzyme mode with peptide length ranging from 8-11 mer, a typical setting for MaxQuant HLA-I peptide discovery described in numerous previous studies(88,89). Match between run and de novo sequencing were enabled for each search by default. Particularly, no peptide FDR was enforced in the initial run to obtain the full list of peptide-spectrum matches (PSMs). The full PSM list was then subject to MS2Rescore progrem(18), which showed to increase the immunopeptidome identification rate by 46% using additional features and deep-learning predicted spectral libraries, to rescore the original p-value and re-rank the PSM lists. Subsequently, a 5% FDR, typically used in immunopeptidome search and has been shown to rescue bona fide antigens which would otherwise be missed at 1% FDR(60,90) was applied to both MaxQuant reported Posterior Error Probability (PEP) and MS2Rescore reported q-value to get the FDR-controlled peptide lists. The union of both were summarized for each tumor immunopeptidome dataset as the initial list of evidenced peptides. Next, netMHCpan 4.1(91) was employed to predict the binding affinity between each identified peptide and the cognate HLA alleles from the sample it was identified, both strong (rank < 0.5%) and weak binders (rank < 2%) were considered, and only the peptides predicted to bind with cognate HLAs were retained to further remove possible false discovery. The maxquant-reported precursor intensity associated with each PSM was reported to estimate the peptide abundance in each immunopeptidome experiment, when multiple scans mapped to the same peptide, maximum precursor intensity was selected as the peptide abundance estimate. When comparing peptide abundance across different peptides, quantile normalization was conducted on the raw intensity values. Briefly, all the eluted peptide in each immunopeptidome experiments were ranked in descending order based on raw intensity values, the upper quantile (75%) value as selected as the base, and the raw intensity of each peptide was divided by the

base value and then been logarithmized with the base 2. The absence of Cryptic-ORF-derived peptides was cross-referenced to IEAtlas database(92).

Alphafold2 modeling

The non-docker version of Alphafold2(93) (version 2.3.0) was downloaded and installed from (https://github.com/kalininalab/alphafold_non_docker). The HLA alleles sequences were retrieved from IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla/). Signal peptides predicted by SignalP(94) (version 5.0) were removed from the N-terminus and the remaining HLA peptide sequence was served as input, together with β2M and peptide sequence. Alphafold2 was run in multimeter mode without relaxation, and the -t argument was wet as 2023-04-29. The highest ranked PDB files (rank_0.pdb) were used for visualization in Pymol3 software (http://www.pymol.org/pymol).

Ribo-Seq Analysis

Ribo-seq for 8 neuroblastoma PDX models was conducted by the Maris lab Children's Hospital of Philadelphia (CHOP). The Ribo-Seq raw fastq files were analyzed by the RibORF (version 2.0) pipeline(95). Briefly, the adapter sequence (TGGAATTCTCGGGTGCCAAGG) was removed from the pair-end Ribo-Seq files. Bowtie2(96) (version 2.3.1) was used to first build the ribosomal RNA index and to align the raw reads to the rRNA reference, removing the noninformative reads. Reads that cannot be mapped to the rRNA reference were subsequently aligned to the human hg19 genome by TopHat(97) to get the SAM file. For each Ribo-Seq read length, a diagnostic plot was automatically generated by the RibORF pipeline for users to specify the optimal offset length to correct the ribosome length, which is normally 12 in the human. The specified offsets were further used as input to convert each aligned ribosomal read to its A site to obtain the simplified SAM file. Both the simplified and original SAM files were directly visualized in Integrative Genomics Viewer (IGV) to evaluate the ribosomal occupancy of selected high confidence cryptic ORF regions. Further, the translatable probability of each cryptic ORFs was reported by RibORF. Both the visual inspection and the quantitative translatable likelihood were considered to decide whether a region is considered translatable in neuroblastoma PDX models.

Peptide validation

Twenty peptide candidates across different molecular classes were synthesized (GenScript, Piscataway, NJ) at the minimum of 95% purity with an average yield of 5-9 mg. Peptides were reconstituted with either 100% water, 5% DMSO/95% water or 0.1% HCOOH/99.9% water to a working concentration of 2mM and further diluted using HPLC-grade water to a final stock concentration of 10pmol/uL. Peptides were pooled and subject to LC-MS/MS run using the similar settings as the original study where each peptide wes identified. Synthetic peptides' raw files were searched against the 20 peptides using MaxQuant and the matched spectra were visualized using spectrum_utils(98) (https://spectrum-utils.readthedocs.io/en/latest/plotting.html) and pyteomics packages (https://github.com/levitsky/pyteomics), and subsequently assessed by MS experts to confirm their identity.

Peptide MHC (pMHC) refolding

HLA expression

The HLA heavy chain was expressed and purified using the bacterial expression vector pET28. BL21/pBirA bacterial strain was transformed with the pET28 vector containing the HLA-inserted DNA and incubated overnight at 37 °C, 170 rpm in 5 mL LB medium supplemented with chloramphenicol (Cm) at 15 µg/mL and Kanamycin (Kan) at 30 µg/mL. 100 mL culture of 2xYT medium supplemented with Cm and Kan was inoculated with 2 mL of preculture; The culture was grown at 37°C with shaking at 170 rpm until an optical density at 600 nm (O.D.600) of approximately 0.5 was reached. At this point, biotin was added to the culture at a final concentration of 50 µM, and incubation was continued for 10 minutes. HLA expression was induced by adding 1 mM IPTG, followed by incubation for 4 hours at 37°C with shaking at 170 rpm. After induction, the culture was centrifuged to pellet the cells. The pellet was lysed by sonication, washed, and resuspended in 8 M Urea buffer to extract inclusion bodies. The Urea-soluble fraction was subjected to affinity chromatography using Econo-Pac Chromatography Columns (Bio-Rad, cat. no. 7321010) packed with Ni Sepharose 6 Fast Flow histidine-tagged protein purification resin (Cytiva, cat. no. 17531801). HLA bound to the resin was eluted with 10 mL of urea buffer containing 0.5 M imidazole. Protein-containing fractions were pooled and concentrated using an Amicon Ultra Centrifugal Filter (10 kDa MWCO, Millipore, cat. no. UFC9010). A diafiltration step with Urea buffer was performed to remove the Imidazole, and the protein was further concentrated to a final concentration of >100 µM.

β2m expression

The BL21 (DE3) pHFT2_β2M bacterial stock was inoculated into 5 mL of LB medium supplemented with kanamycin (30 µg/mL) and incubated overnight at 37°C with shaking. 100 mL culture of 2xYT medium with Kanamycin was inoculated with 2 mL of the overnight preculture and grown at 37°C, shaking at 170 rpm, until the O.D.600 reached approximately 0.5. Expression of β2-microglobulin (β2m) was induced by adding 1 mM IPTG, and the culture was incubated overnight under the same conditions. The following day, the culture was centrifuged to pellet the cells. β2m purification was performed using a similar protocol to that of HLA, with a key modification to include an on-column refolding step. This was achieved by passing a gradient of Urea buffer (8 M urea, 0.1 M Tris-HCl, pH 8) and Native buffer (50 mM Tris-HCl, 250 mM NaCl, pH 8) through the column. The gradient was applied to gradually reduce the urea concentration until only native buffer was used. β2m was eluted with 10 mL of Native buffer containing 0.5 M Imidazole, and protein-containing fractions were pooled and concentrated. To remove Imidazole and separate persistent aggregates from the monomeric fraction, the concentrated β2m solution was loaded onto the AKTA Pure system and subjected to size exclusion chromatography (SEC) using a Superdex 75 Increase 10/300 GL column (Cytiva, cat. no. 29148721). Monomeric β2m fractions were collected for further processing. Since β2m is a small protein, the His-tag used during purification could interfere with subsequent refolding and alignment of the MHC complex. To address this, the His-tag was cleaved using TEV protease (Sigma, cat. no. T4455), as the β2m sequence included a TEV protease cleavage site. Following cleavage, the protein mixture was run through the AKTA system again to isolate the cleaved β2m monomer. The purified monomeric β2m, free of His-tag, was then used for peptide-MHC refolding.

pMHC refolding

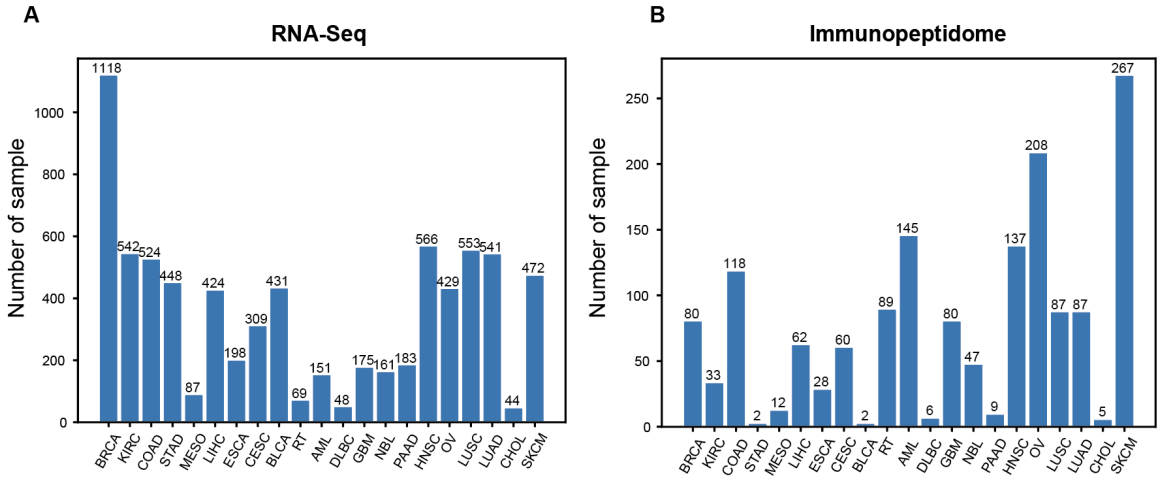
Synthetic peptides were dissolved to a final concentration of 2 mM in the appropriate buffer, as determined by the solubility test. Refolding of the peptide-MHC (pMHC) complex was performed in the PBS buffer using 3 μM cleaved β2m, 30 μM peptide, and 3 μM HLA, added sequentially in a total volume of 2 mL PBS. The solution was incubated overnight at 4°C with gentle rotation to ensure proper mixing and refolding. The following day, the refolding mixture was centrifuged to remove any insoluble material. The clarified supernatant containing the pMHC complex was loaded onto the AKTA Pure system for purification. The fractions containing the correctly refolded pMHC complex were collected and further analyzed.

Cross-reactivity TCR experiment

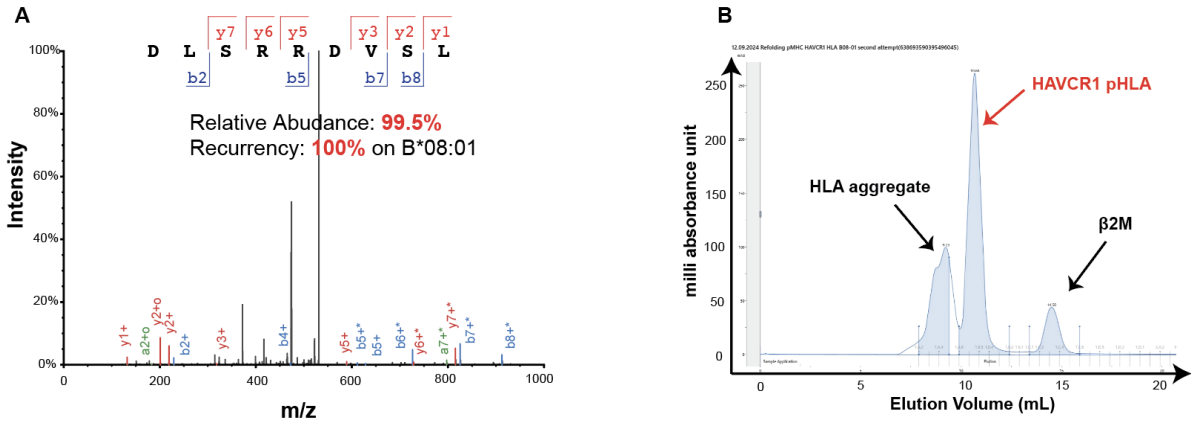
Mouse 58-/- T cell hybridoma cells(99), which expresses mouse CD3 but not TCRab (from David Kranz, University of Illinois) and Chinese hamster ovary (CHO) cells expressing HLA-A2 (generous gift from Dr. James L. Riley, University of Pennsylvania) were cultured in RPMI 1640 medium and DMEM, respectively, supplemented with 10% FBS, sodium pyruvate, non-essential amino acids, glutaMAX-1, penicillin-streptomycin and β-mercaptoethanol. Human variable-mouse constant chimeric gp100(2M)-specific TCR constructs-19LF6 or 16LD6(100) and human CD8 were retrovirally transduced into the hybridoma cells(101). Transduced cells were sorted, expanded for 6 days, quantified for TCR and CD3e expression, and prepared for the coculture cytokine assay. 10,000 CHO cells expressing HLA-A2 were loaded with different concentrations of gp100(2M) peptide (IMDQVPFSV), in which anchor position 2 was mutated from original peptide T to M to increase binding affinity and PMEL spliced variant peptide (KTWDQVPFSV) and incubated with 10,000 T cell hybridoma clones expressing human CD8 and 19LF6 or 16LD6 TCR in triplicates for 16 h at 37 °C, 5% CO2. A standard ELISA sandwich was used to quantify cytokine mouse IL-2 production(102).

Clustering analysis of eluted peptide motifs

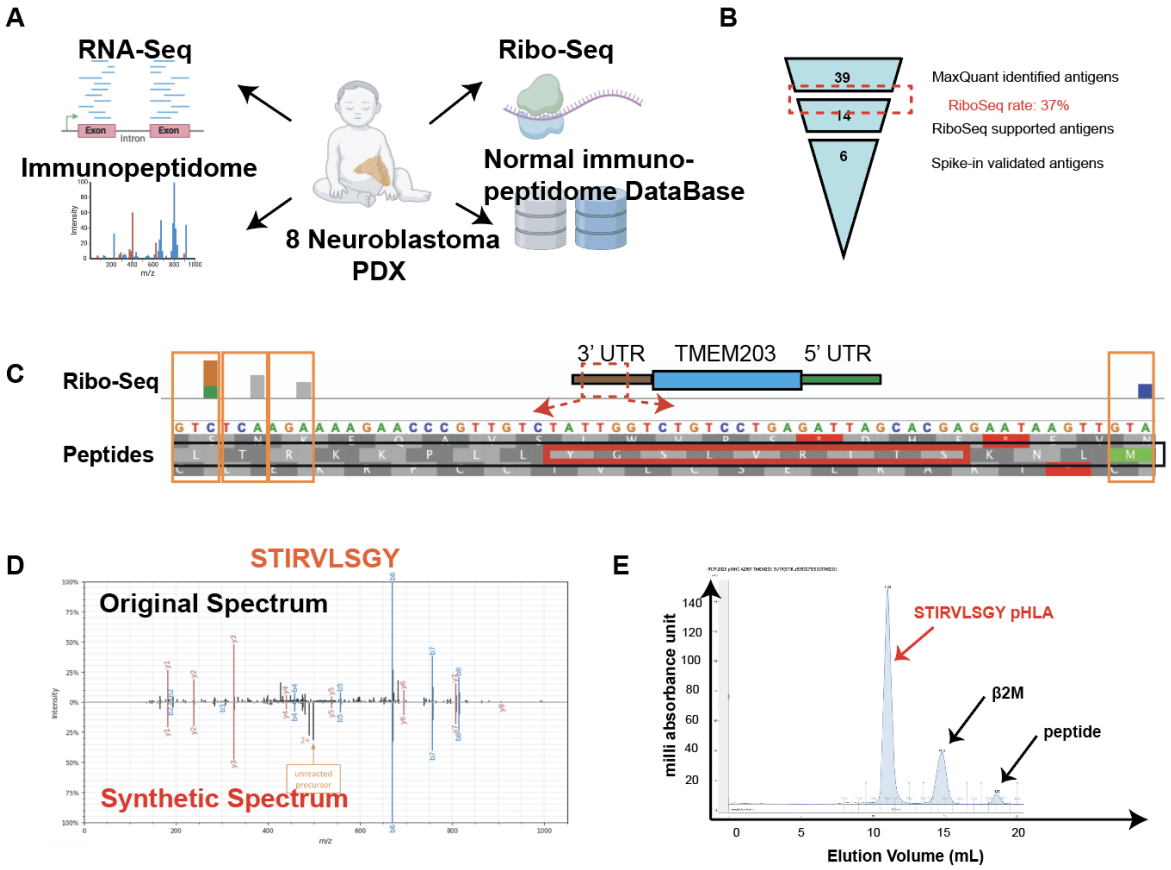
Clustering analysis was conducted using GibbsCluster(103) version 2.0 through the web server (https://services.healthtech.dtu.dk/services/GibbsCluster-2.0). All eluted HLA-presented antigens from DLBC and Rhabdoid tumor (RT) were included as input and subject to Gibbs clustering using default setting. The optimal number of clusters were determined by the maximum Kullback–Leibler divergence.



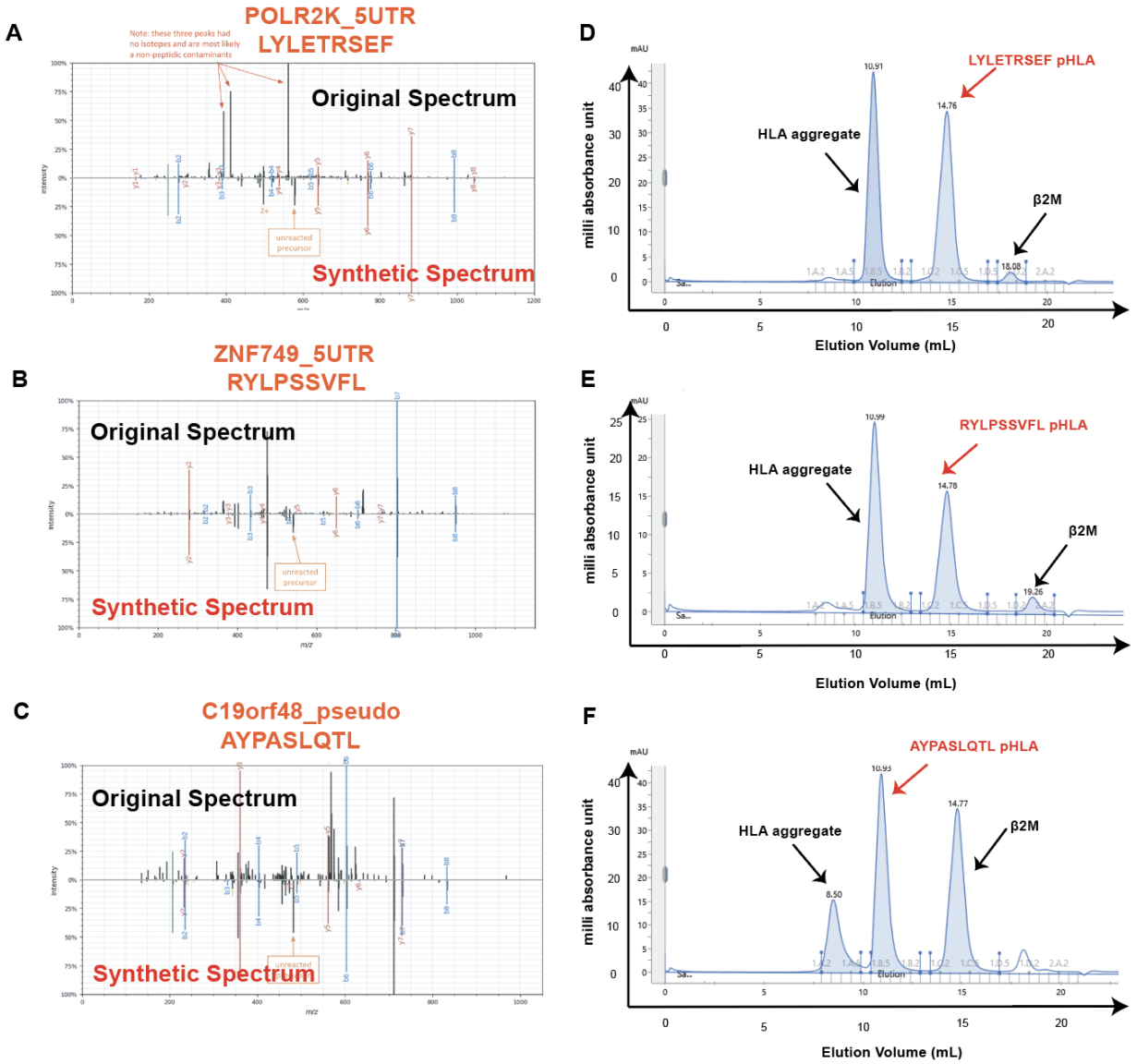
Supplementary Figure 1. Key statistics associated with public RNA-Seq and immunopeptidome dataset. (A) Number of RNA-Seq datasets for each tumor across TCGA and TARGET. (B) Number of immunopeptidomic datasets for each tumor collected throughout public repositories.



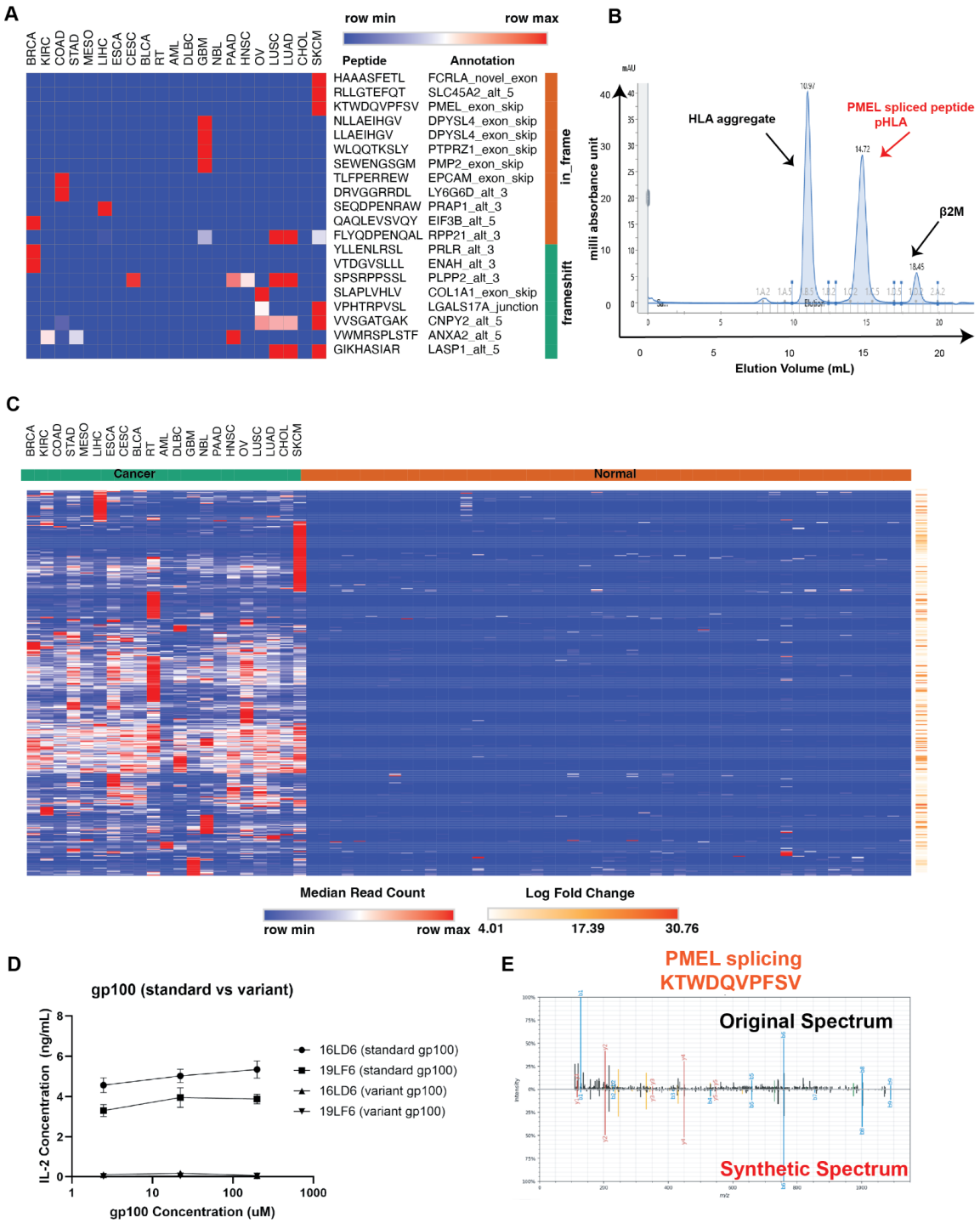
Supplementary Figure 2. HAVCR1 derived tumor specific self-antigen in clear cell renal cell carcinoma. (A) high confidence MS spectrum for the HAVCR1 antigen, matched b- and y-ions are annotated with the relative abundance and recurrency in the evaluated datasets
(B) Size Exclusion Chromatography of the refolded HAVCR1 peptide-HLA-B*08:01 complex.



Supplementary Figure 3. Cryptic ORF antigen validation workflow and TMEM203 3' UTR derived STIRVLSGY antigen. (A) workflow of assessing translational activity of cryptic ORFs using Ribo-Seq in matched PDX neuroblastoma models (B) funnel plot of validated cryptic ORF antigen in matched Neuroblastoma PDX model based on Ribo-Seq evidence (C) Ribosome occupancies on 3' UTR of TMEM203 from Ribo-Seq, the STIRVLSGY peptide was highlighted (D) Mirror plot of synthetic peptide and original spectra to confirm the MS identification (E) Size exclusion chromatography for the refolded STIRVLSGY-HLA-A*26:01 complex.



Supplementary Figure 4. Validation of additional Cryptic ORF antigens. (A-C) Mirror plots of original spectrum from the tumor samples and the synthetic peptide spectra to validate the MS identification for POLR2K_5UTR, ZNF749_5UTR, C19orf48_pseudogene (top to bottom). (D-E) Size Exclusion Chromatography of the refolded pHLA for LYLETRSEF-A2402 complex, RYLPSSVFL-A2402 complex, AYPASLQTL-A2402 complex (top to bottom).



Supplementary Figure 5. Splicing antigen landscapes across tumor types. (A) frequently detected splicing antigen across cancers, annotated by splicing types and whether the spliced product is in-frame or out-of-frame compared to canonical transcript. (B) Size exclusion chromatography of the refolded PMEL spliced peptide and HLA-A*02:01 allele (C) tumor specific splicing junctions across cancer types. (D) Cross-reactivity assay of a TCR-T specific to the canonical ITD epitope, assessing its activity against both the canonical target and a newly identified spliced variant (E) mirror plot of synthetic PMEL spliced peptide and original spectrum.

Reference

1. Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. Nature Reviews Immunology. Nature Publishing Group; 2020;20:651–68.

2. Labanieh L, Mackall CL. CAR immune cells: design principles, resistance and the next generation. Nature. Nature Publishing Group; 2023;614:635–48.

3. Li G, Schnell D, Bhattacharjee A, Yarmarkovich M, Salomonis N. Quantifying tumor specificity using Bayesian probabilistic modeling for drug and immunotherapeutic target discovery. Cell reports methods [Internet]. Cell Rep Methods; 2024 [cited 2025 Jan 8];4. Available from: https://pubmed.ncbi.nlm.nih.gov/39515334/

4. MacKay M, Afshinnekoo E, Rub J, Hassan C, Khunte M, Baskaran N, et al. The therapeutic landscape for cells engineered with chimeric antigen receptors. Nature Biotechnology. Nature Publishing Group; 2020;38:233–44.

5. Madan S, Sinha S, Chang T, Gutkind JS, Cohen EEW, Schäffer AA, et al. Pan-Cancer Analysis of Patient Tumor Single-Cell Transcriptomes Identifies Promising Selective and Safe Chimeric Antigen Receptor Targets in Head and Neck Cancer. Cancers [Internet]. Cancers (Basel); 2023 [cited 2025 Jan 8];15. Available from: https://pubmed.ncbi.nlm.nih.gov/37835579/

6. Leone P, Shin E-C, Perosa F, Vacca A, Dammacco F, Racanelli V. MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. J Natl Cancer Inst. 2013;105:1172–87.

7. Sun Q, Hong Z, Zhang C, Wang L, Han Z, Ma D. Immune checkpoint therapy for solid tumours: clinical dilemmas and future trends. Signal Transduct Target Ther. 2023;8:320.

8. Rosenberg SA, Dudley ME. Adoptive cell therapy for the treatment of patients with metastatic melanoma. Curr Opin Immunol. 2009;21:233–40.

9. Tran E, Urba WJ, Leidner R. Neoantigen T-Cell Receptor Gene Therapy in Pancreatic Cancer. Reply. N Engl J Med. 2022;387:574.

10. Rojas LA, Sethna Z, Soares KC, Olcese C, Pang N, Patterson E, et al. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. Nature. 2023;618:144–50.

11. Weber JS, Carlino MS, Khattak A, Meniawy T, Ansstas G, Taylor MH, et al. Individualised neoantigen therapy mRNA-4157 (V940) plus pembrolizumab versus pembrolizumab monotherapy in resected melanoma (KEYNOTE-942): a randomised, phase 2b study. Lancet. 2024;403:632–44.

12. Yarmarkovich M, Marshall QF, Warrington JM, Premaratne R, Farrel A, Groff D, et al. Targeting of intracellular oncoproteins with peptide-centric CARs. Nature. 2023;623:820–7.

13. Li G, Mahajan S, Ma S, Jeffery ED, Zhang X, Bhattacharjee A, et al. Splicing neoantigen discovery with SNAF reveals shared targets for cancer immunotherapy. Sci Transl Med. 2024;16:eade2886.

14. Nadal R, Barisic S, Scurti GM, Cherkasova E, Chen L, Wood K, et al. Final results of a phase I trial of HERV-E TCR transduced T cells for the treatment of HLA-A*11 patients with metastatic clear cell renal cell carcinoma (mccRCC). J Clin Oncol. American Society of Clinical Oncology (ASCO); 2024;42:435–435.

15. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, Aguet F, et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat Biotechnol. 2022;40:209–17.

16. Kacen A, Javitt A, Kramer MP, Morgenstern D, Tsaban T, Shmueli MD, et al. Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. Nat Biotechnol. 2023;41:239–51.

17. Griss J, Perez-Riverol Y, Lewis S, Tabb DL, Dianes JA, del-Toro N, et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nature methods. 2016;13:651.

18. MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. Molecular & Cellular Proteomics. Elsevier; 2022;21:100266.

19. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. Journal of immunology (Baltimore, Md : 1950) [Internet]. J Immunol; 2017 [cited 2025 Jan 8];199. Available from: https://pubmed.ncbi.nlm.nih.gov/28978689/

20. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. Immunity. Cell Press; 2021;54:737–52.e10.

21. Deniz Ö, Ahmed M, Todd CD, Rio-Machin A, Dawson MA, Branco MR. Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. Nature communications [Internet]. Nat Commun; 2020 [cited 2025 Jan 8];11. Available from: https://pubmed.ncbi.nlm.nih.gov/32665538/

22. Morgan RA, Yang JC, Kitano M, Dudley ME, Laurencot CM, Rosenberg SA. Case report of a serious adverse event following the administration of T cells transduced with a chimeric antigen receptor recognizing ERBB2. Mol Ther. 2010;18:843–51.

23. Qin D, Li D, Zhang B, Chen Y, Liao X, Li X, et al. Potential lung attack and lethality generated by EpCAM-specific CAR-T cells in immunocompetent mouse models. Oncoimmunology. 2020;9:1806009.

24. Erler P, Kurcon T, Cho H, Skinner J, Dixon C, Grudman S, et al. Multi-armored allogeneic MUC1 CAR T cells enhance efficacy and safety in triple-negative breast cancer. Science Advances [Internet]. American Association for the Advancement of Science; 2024 [cited 2025 Jan 21]; Available from: https://www.science.org/doi/10.1126/sciadv.adn9857

25. Tousley AM, Rotiroti MC, Labanieh L, Rysavy LW, Kim W-J, Lareau C, et al. Co-opting signalling molecules enables logic-gated control of CAR T cells. Nature. 2023;615:507.

26. Roybal KT, Williams JZ, Morsut L, Rupp LJ, Kolinko I, Choe JH, et al. Engineering T Cells with Customized Therapeutic Response Programs Using Synthetic Notch Receptors. Cell. 2016;167:419–32.e16.

27. Cuadros T, Trilla E, Vilà MR, de Torres I, Vilardell J, Messaoud NB, et al. Hepatitis A virus cellular receptor 1/kidney injury molecule-1 is a susceptibility gene for clear cell renal cell carcinoma and hepatitis A virus cellular receptor/kidney injury molecule-1 ectodomain shedding a predictive biomarker of tumour progression. European journal of cancer (Oxford, England : 1990) [Internet]. Eur J Cancer; 2013 [cited 2025 Jan 17];49. Available from: https://pubmed.ncbi.nlm.nih.gov/23352434/

28. Chiaro J, Antignani G, Feola S, Feodoroff M, Martins B, Cojoc H, et al. Development of mesothelioma-specific oncolytic immunotherapy enabled by immunopeptidomics of murine and human mesothelioma tumors. Nature Communications. Nature Publishing Group; 2023;14:1–16.

29. Parkhurst MR, Robbins PF, Tran E, Prickett TD, Gartner JJ, Jia L, et al. Unique Neoantigens Arise from Somatic Mutations in Patients with Gastrointestinal Cancers. Cancer Discov. 2019;9:1022–35.

30. Yarmarkovich M, Farrel A, Sison A, di Marco M, Raman P, Parris JL, et al. Immunogenicity and Immune Silence in Human Cancer. Front Immunol. Frontiers; 2020;11:471967.

31. Choi J, Goulding SP, Conn BP, McGann CD, Dietze JL, Kohler J, et al. Systematic discovery and validation of T cell targets directed against oncogenic KRAS mutations. Cell Reports Methods. 2021;1:100084.

32. Minegishi Y, Kiyotani K, Nemoto K, Inoue Y, Haga Y, Fujii R, et al. Comprehensive-immunopeptidomics analysis reveals presentation of potential neoantigens carrying cancer driver mutations [Internet]. bioRxiv. 2021 [cited 2025 Jan 8]. page 2021.04.09.439118. Available from: https://www.biorxiv.org/content/10.1101/2021.04.09.439118v3.abstract

33. Kraemer AI, Chong C, Huber F, Pak H, Stevenson BJ, Müller M, et al. The immunopeptidome landscape associated with T cell infiltration, inflammation and immune editing in lung cancer. Nature Cancer. Nature Publishing Group; 2023;4:608–28.

34. Tran E, Robbins PF, Lu Y-C, Prickett TD, Gartner JJ, Jia L, et al. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. The New England journal of medicine. 2016;375:2255.

35. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. Nature Publishing Group; 2020;578:94–101.

36. Chen Y, Williams V, Filippova M, Filippov V, Duerksen-Hughes P. Viral Carcinogenesis: Factors Inducing DNA Damage and Virus Integration. Cancers. 2014;6:2155.

37. Mahmood F, Xu R, Awan MUN, Song Y, Han Q, Xia X, et al. HBV Vaccines: Advances and Development. Vaccines. 2023;11:1862.

38. Cheng L, Wang Y, Du J. Human Papillomavirus Vaccines: An Updated Review. Vaccines. 2020;8:391.

39. Battaglia TW, Mimpen IL, Traets JJH, van Hoeck A, Zeverijn LJ, Geurts BS, et al. A pan-cancer analysis of the microbiome in metastatic cancer. Cell. 2024;187:2324–35.e19.

40. Gihawi A, Ge Y, Lu J, Puiu D, Xu A, Cooper CS, et al. Major data analysis errors invalidate

cancer microbiome findings. mBio [Internet]. American Society for Microbiology1752 N St., N.W., Washington, DC; 2023 [cited 2024 Dec 27]; Available from: https://journals.asm.org/doi/10.1128/mbio.01607-23

41. Pan-cancer atlas of tumor-resident microbiome, immunity and prognosis. Cancer Letters. Elsevier; 2024;598:217077.

42. Parsonnet J, Friedman GD, Vandersteen DP, Chang Y, Vogelman JH, Orentreich N, et al. Helicobacter pylori infection and the risk of gastric carcinoma. N Engl J Med. 1991;325:1127–31.

43. Zepeda-Rivera M, Minot SS, Bouzek H, Wu H, Blanco-Míguez A, Manghi P, et al. A distinct Fusobacterium nucleatum clade dominates the colorectal cancer niche. Nature. Nature Publishing Group; 2024;628:424–32.

44. Kalaora S, Nagler A, Nejman D, Alon M, Barbolin C, Barnea E, et al. Identification of bacteria-derived HLA-bound peptides in melanoma. Nature. 2021;592:138.

45. First report of spondylodiscitis caused by Bacillus circulans in an immunocompetent patient: Clinical case and review of the literature. IDCases. Elsevier; 2021;23:e01058.

46. Shah NM, Jang HJ, Liang Y, Maeng JH, Tzeng S-C, Wu A, et al. Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. Nature Genetics. Nature Publishing Group; 2023;55:631–9.

47. Merlotti A, Sadacca B, Arribas YA, Ngoma M, Burbage M, Goudot C, et al. Noncanonical splicing junctions between exons and transposable elements represent a source of immunogenic recurrent neo-antigens in patients with lung cancer. Science Immunology [Internet]. American Association for the Advancement of Science; 2023 [cited 2025 Jan 9]; Available from: https://www.science.org/doi/10.1126/sciimmunol.abm6359

48. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nature genetics [Internet]. Nat Genet; 2020 [cited 2025 Jan 14];52. Available from: https://pubmed.ncbi.nlm.nih.gov/32024998/

49. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. Mobile DNA [Internet]. Mob DNA; 2016 [cited 2025 Jan 14];7. Available from: https://pubmed.ncbi.nlm.nih.gov/27980689/

50. Burbage M, Rocañín-Arjó A, Baudon B, Arribas YA, Merlotti A, Rookhuizen DC, et al. Epigenetically controlled tumor antigens derived from splice junctions between exons and transposable elements. Science Immunology [Internet]. American Association for the Advancement of Science; 2023 [cited 2025 Jan 9]; Available from: https://www.science.org/doi/10.1126/sciimmunol.abm6360

51. Takahashi Y, Harashima N, Kajigaya S, Yokoyama H, Cherkasova E, McCoy JP, et al. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. The Journal of clinical investigation [Internet]. J Clin Invest; 2008 [cited 2025 Jan 9];118. Available from: https://pubmed.ncbi.nlm.nih.gov/18292810/

52. [No title] [Internet]. [cited 2025 Jan 9]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC39958/pdf/pnas01508-0090.pdf

53. Ardeljan D, Wang X, Oghbaie M, Taylor MS, Husband D, Deshpande V, et al. LINE-1 ORF2p expression is nearly imperceptible in human cancers. Mobile DNA. 2019;11:1.

54. Baldwin ET, van Eeuwen T, Hoyos D, Zalevsky A, Tchesnokov EP, Sánchez R, et al. Structures, functions and adaptations of the human LINE-1 ORF2 protein. Nature. Nature Publishing Group; 2023;626:194–206.

55. Cruz-Tapias P, Castiblanco J, Anaya J-M. Major histocompatibility complex: Antigen processing and presentation. Autoimmunity: From Bench to Bedside [Internet]. El Rosario University Press; 2013.

56. Ferrington DA, Gregerson DS. Immunoproteasomes: structure, function, and antigen presentation. Progress in molecular biology and translational science [Internet]. Prog Mol Biol Transl Sci; 2012 [cited 2025 Jan 9];109. Available from: https://pubmed.ncbi.nlm.nih.gov/22727420/

57. The roles of ERAP1 and ERAP2 in autoimmunity and cancer immunity: New insights and perspective. Molecular Immunology. Pergamon; 2020;121:7–19.

58. de Castro L, José A. How ERAP1 and ERAP2 Shape the Peptidomes of Disease-Associated MHC-I Proteins. Front Immunol. Frontiers; 2018;9:415782.

59. Yarzabek B, Zaitouna AJ, Olson E, Silva GN, Geng J, Geretz A, et al. Variations in HLA-B cell surface expression, half-life and extracellular antigen receptivity. eLife Sciences Publications Limited; 2018 [cited 2025 Jan 20]; Available from: https://elifesciences.org/articles/34961

60. caAtlas: An immunopeptidome atlas of human cancer. iScience. Elsevier; 2021;24:103107.

61. Yewdell JW, Antón LC, Bennink JR. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? Journal of immunology (Baltimore, Md : 1950) [Internet]. J Immunol; 1996 [cited 2025 Jan 21];157. Available from: https://pubmed.ncbi.nlm.nih.gov/8757297/

62. Seliger B, Kloor M, Ferrone S. HLA class II antigen-processing pathway in tumors: Molecular defects and clinical relevance. Oncoimmunology [Internet]. Oncoimmunology; 2017 [cited 2025 Jan 9];6. Available from: https://pubmed.ncbi.nlm.nih.gov/28344859/

63. Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, Mifsud NA, et al. A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. Science immunology [Internet]. Sci Immunol; 2018 [cited 2025 Jan 17];3. Available from: https://pubmed.ncbi.nlm.nih.gov/30315122/

64. Tsour S, Machne R, Leduc A, Widmer S, Slavov N. Alternate RNA decoding results in stable and abundant proteins in mammals [Internet]. bioRxiv. 2024 [cited 2024 Dec 18]. page 2024.08.26.609665. Available from: https://www.biorxiv.org/content/10.1101/2024.08.26.609665v1.abstract

65. Hsiehchen D, Elliott A, Xiu J, Seeber A, El-Deiry W, Antonarakis ES, et al. Mutation burden

and anti-PD-1 outcomes are not universally associated with immune cell infiltration or lymphoid activation. Cancer cell [Internet]. Cancer Cell; 2024 [cited 2024 Dec 30];42. Available from: https://pubmed.ncbi.nlm.nih.gov/39577419/

66. Zeng J, Zhang Y, Shang Y, Mai J, Shi S, Lu M, et al. CancerSCEM: a database of single-cell expression map across various human cancers. Nucleic acids research [Internet]. Nucleic Acids Res; 2022 [cited 2025 Jan 19];50. Available from: https://pubmed.ncbi.nlm.nih.gov/34643725/

67. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. Nature Publishing Group; 2016;34:525–7.

68. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45:W98–102.

69. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29:15.

70. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biology. BioMed Central; 2019;20:1–13.

71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078.

72. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841.

73. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology. BioMed Central; 2019;20:1–13.

74. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. PeerJ Inc.; 2017;3:e104.

75. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics (Oxford, England) [Internet]. Bioinformatics; 2015 [cited 2025 Jan 19];31. Available from: https://pubmed.ncbi.nlm.nih.gov/26206304/

76. Smith CC, Beckermann KE, Bortone DS, De Cubas AA, Bixby LM, Lee SJ, et al. Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. The Journal of Clinical Investigation. 2018;128:4804.

77. Haas BJ, Dobin A, Stransky N, Li B, Yang X, Tickle T, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq [Internet]. bioRxiv. 2017 [cited 2025 Jan 19]. page 120295. Available from: https://www.biorxiv.org/content/10.1101/120295v1.abstract

78. Jang YE, Jang I, Kim S, Cho S, Kim D, Kim K, et al. ChimerDB 4.0: an updated and expanded database of fusion genes. Nucleic acids research [Internet]. Nucleic Acids Res; 2020 [cited 2025 Jan 19];48. Available from: https://pubmed.ncbi.nlm.nih.gov/31680157/

79. Oikkonen L, Lise S. Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. Wellcome open research [Internet]. Wellcome

Open Res; 2017 [cited 2025 Jan 19];2. Available from: https://pubmed.ncbi.nlm.nih.gov/28239666/

80. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Medicine. BioMed Central; 2020;12:1–13.

81. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature genetics. 2014;46:912.

82. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. BioMed Central; 2016;17:1–14.

83. Del Corvo M, Mazzara S, Pileri SA. TOSCA: an automated Tumor Only Somatic CAlling workflow for somatic mutation detection without matched normal samples. Bioinform Adv. 2022;2:vbac070.

84. Mansi L, Tangaro MA, Lo Giudice C, Flati T, Kopel E, Schaffer AA, et al. REDIportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. Nucleic Acids Res. 2021;49:D1012–9.

85. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47:D941–7.

86. Marcu A, Schlosser A, Keupp A, Trautwein N, Johann P, Wölfl M, et al. Natural and cryptic peptides dominate the immunopeptidome of atypical teratoid rhabdoid tumors. J Immunother Cancer. BMJ Specialist Journals; 2021;9:e003404.

87. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology. Nature Publishing Group; 2008;26:1367–72.

88. Zhang T, Celiker B, Shao Y, Gai J, Hill CM, Wang C, et al. Comparison of Shared Class I HLA-bound Non-canonical Neoepitopes between Normal and Neoplastic Tissues of Pancreatic Adenocarcinoma. Clinical cancer research : an official journal of the American Association for Cancer Research [Internet]. Clin Cancer Res; 2024 [cited 2025 Jan 19]; Available from: https://pubmed.ncbi.nlm.nih.gov/39699517/

89. Weingarten-Gabbay S, Klaeger S, Sarkizova S, Pearlman LR, Chen D-Y, Gallagher KME, et al. Profiling SARS-CoV-2 HLA-I peptidome reveals T cell epitopes from out-of-frame ORFs. Cell. 2021;184:3962.

90. Narayan R, Olsson N, Wagar LE, Medeiros BC, Meyer E, Czerwinski D, et al. Acute myeloid leukemia immunopeptidome reveals HLA presentation of mutated nucleophosmin. PLOS ONE. Public Library of Science; 2019;14:e0219547.

91. 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 Res. 2020;48:W449–54.

92. Cai Y, Lv D, Li D, Yin J, Ma Y, Luo Y, et al. IEAtlas: an atlas of HLA-presented immune epitopes derived from non-coding regions. Nucleic Acids Res. 2023;51:D409–17.

93. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. Nature Publishing Group; 2021;596:583–9.

94. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology. Nature Publishing Group; 2019;37:420–3.

95. Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. 2019;47:D230–4.

96. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. Nature Publishing Group; 2012;9:357–9.

97. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. BioMed Central; 2013;14:1–13.

98. Bittremieux W, Levitsky L, Pilz M, Sachsenberg T, Huber F, Wang M, et al. Unified and Standardized Mass Spectrometry Data Processing in Python Using spectrum_utils. J Proteome Res. 2023;22:625–31.

99. Letourneur F, Malissen B. Derivation of a T cell hybridoma variant deprived of functional T cell receptor alpha and beta chain transcripts reveals a nonfunctional alpha-mRNA of BW5147 origin. Eur J Immunol. 1989;19:2269–74.

100. Zhong S, Malecek K, Johnson LA, Yu Z, Vega-Saenz de Miera E, Darvishian F, et al. T-cell receptor affinity and avidity defines antitumor response and autoimmunity in T-cell immunotherapy. Proceedings of the National Academy of Sciences. National Academy of Sciences; 2013;110:6973–8.

101. Zhong S, Malecek K, Perez-Garcia A, Krogsgaard M. Retroviral transduction of T-cell receptors in mouse T-cells. Journal of visualized experiments : JoVE [Internet]. J Vis Exp; 2010 [cited 2025 Jan 19]; Available from: https://pubmed.ncbi.nlm.nih.gov/21048669/

102. Natarajan A, Velmurugu Y, Becerra Flores M, Dibba F, Beesam S, Kikvadze S, et al. In situ cell-surface conformation of the TCR-CD3 signaling complex. EMBO Rep. 2024;25:5719–42.

103. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. Nucleic acids research [Internet]. Nucleic Acids Res; 2017 [cited 2025 Jan 19];45. Available from: https://pubmed.ncbi.nlm.nih.gov/28407089/