eLife

Reviewed Preprint
v2 • September 11, 2024
Revised by authors

Reviewed Preprint
v1 • April 12, 2024

Cancer Biology

The T Cell Receptor β Chain Repertoire of Tumor Infiltrating Lymphocytes Improves Neoantigen Prediction and Prioritization

Thi Mong Quynh Pham, Thanh Nhan Nguyen, Bui Que Tran Nguyen, Thi Phuong Diem Tran, Nguyen My Diem Pham, Hoang Thien Phuc Nguyen, Thi Kim Cuong Ho, Dinh Viet Linh Nguyen, Huu Thinh Nguyen, Duc Huy Tran, Thanh Sang Tran, Truong-Vinh Ngoc Pham, Minh-Triet Le, Thi Tuong Vy Nguyen, Minh-Duy Phan, Hoa Giang, Hoai-Nghia Nguyen, Le Son Tran

Medical Genetics Institute, Ho Chi Minh City, Vietnam • University Medical Center Ho Chi Minh City, Ho Chi Minh City, Vietnam

Abstract

In the realm of cancer immunotherapy, the meticulous selection of neoantigens plays a fundamental role in enhancing personalized treatments. Traditionally, this selection process has heavily relied on predicting the binding of peptides to human leukocyte antigens (pHLA). Nevertheless, this approach often overlooks the dynamic interaction between tumor cells and the immune system. In response to this limitation, we have developed an innovative prediction algorithm rooted in machine learning, integrating T cell receptor β chain (TCRβ) profiling data from colorectal cancer (CRC) patients for a more precise neoantigen prioritization. TCRβ sequencing was conducted to profile the TCR repertoire of tumor-infiltrating lymphocytes (TILs) from 28 CRC patients. The data unveiled both intra-tumor and inter-patient heterogeneity in the TCRβ repertoires of CRC patients, likely resulting from the stochastic utilization of V and J segments in response to neoantigens. Our novel combined model integrates pHLA binding information with pHLA-TCR binding to prioritize neoantigens, resulting in heightened specificity and sensitivity compared to models using individual features alone. The efficacy of our proposed model was corroborated through ELISpot assays on long peptides, performed on four CRC patients. These assays demonstrated that neoantigen candidates prioritized by our combined model outperformed predictions made by the established tool NetMHCpan. This comprehensive assessment underscores the significance of integrating pHLA binding with pHLA-TCR binding analysis for more effective immunotherapeutic strategies.

eLife assessment

The study presents a potentially **valuable** approach by combining two measurements (pHLA binding and pHLA-TCR binding) to improve predictions of which mutations in colorectal cancer are likely to be presented to and recognised by the immune system. While this approach is promising, the evidence supporting the primary claim remains somewhat **incomplete**. The experimental validation of the computational predictions with actual immune responses is still limited, despite the increase in sample size from 4 to 8 in this revision.

https://doi.org/10.7554/eLife.94658.2.sa2

Introduction

In metastatic CRC patients, immunotherapy has fulfilled promise in improving survival rate (1). Immune checkpoint inhibitors (ICIs), which block negative regulatory pathways in T-cell activation, have been approved by the US Food and Drug Administration (FDA) for the treatment of deficient mismatch repair (dMMR) or high microsatellite instability (MSI-H) CRC patients (2–4). However, there is an urgent need for alternative immunotherapy strategies for metastatic CRC patients, as patients with proficient mismatch repair (pMMR) or microsatellite stability (MSS) have not shown significant responses to immune checkpoint inhibitors (2, 5).

Neoantigens (neopeptides) have emerged as potential targets for personalized cancer immunotherapy, including CRC (6, 7). Neoantigens are peptides that result from somatic mutations and can be displayed by class I human leukocyte antigen (HLA-I) molecules on the surface of cancer cells, thereby activating immune-mediated tumor killing (8). Recent studies have demonstrated that the presence of neoantigens is associated with better responses to immune checkpoint inhibitor (ICI) therapy in CRC patients (9, 10). A high neoantigen burden has been linked to improved overall survival and progression-free survival in patients with various solid tumors, including CRC (9, 10). Therefore, neoantigen-based immunotherapies are considered to have significant potential for improving treatment outcomes for CRC patients.

The successful development of neoantigen-based therapies hinges upon the identification of neoantigens that exhibit a strong binding affinity to their respective HLA-I molecules and demonstrate high immunogenicity. Initially, DNA sequencing of tumor tissues and paired Peripheral Blood Mononuclear Cells identifies cancer-associated genomic mutations. RNA sequencing then determines the patient's HLA-I allele profile and the gene expression levels of mutated genes. *In silico* tools are then applied to analyze the tumor somatic variant, HLA-I allele, and gene expression data, predicting the binding affinity of neoantigens to the patient's HLA-I alleles and their potential to activate T cell responses (11–13). This standard workflow has been successful in identifying clinically relevant neoantigens in various malignancies (12, 14).

However, despite its achievements, the current approach's impact on patient outcomes remains limited due to the scarcity of mutations in cancer patients that lead to effective immunogenic neoantigens. This limitation arises primarily from the fact that the selection of neoantigen candidates by these workflows relies solely on neoantigen-HLA-I binding affinity as the criteria for immunogenicity prediction (15). While HLA-I binding is indeed a crucial factor for neoantigen presentation, it does not fully account for T cell receptor (TCR) recognition and interaction. The recognition of the neoantigen-HLA complex through TCR is of paramount importance for T cell

activation and eliciting an immune response (16, 17). Various factors, such as the specific TCR repertoire, TCR clonality, and the structural characteristics of the TCR-peptide-HLA complex, profoundly influence TCR recognition. Unfortunately, these critical aspects are not entirely captured by HLA-I binding prediction alone.

The TCR is a critical component of the adaptive immune system, responsible for recognizing specific antigens presented by antigen-presenting cells (APCs). This membrane-bound heterodimer protein complex is expressed on the surface of T cells and exists in two distinct forms: TCRα/TCRβ for αβ T cells and TCRγ/TCRδ for γδ T cells, both intricately associatedwith invariant CD3 chain molecules (18, 19). TCR diversity arises from the recombination of variable (V), diversity (D), and joining (J) genes at the TCRβ and TCRδ loci, along with VJ recombination at the TCRα and TCRγ loci (20), giving rise to a broad array of unique TCRs collectively known as the TCR repertoire. The complementarity-determining region 3 (CDR3), situated at the junction of V, D, and J gene segments (21), plays a pivotal role in antigen recognition, with the unique combination of CDR3 sequences contributing significantly to the specificity and diversity of the TCR repertoire. With advancements in technology, TCR sequencing has become a powerful technique used to characterize the diversity and composition of TCR repertoires. Valuable information directly obtained through TCR sequencing, including TCR clonotype diversity, V(D)J gene usage, repertoire size, clonal expansion, and repertoire changes, has provided invaluable insights into immune responses, antigen recognition, and the development of targeted immunotherapies.

Sequencing the T cell receptors (TCRs) of TILs or lymphocytes found in peripheral blood provides crucial insights into the T-cell repertoire and their responses against neoantigens associated with tumors (22–24). This information holds paramount importance in identifying TCRs that specifically target these neoantigens. Additionally, it proves to be invaluable in assessing the immunogenic potential of predicted neoantigens. Unfortunately, the current bioinformatic workflows used in neoantigen selection and prioritization do not incorporate TCR sequencing data. Therefore, we speculate that integrating TCR sequencing data into the assessment of immunogenicity of predicted neoantigens holds promise for unveiling effective neoantigens which can induce immune response, consequently, advancing the development of personalized immunotherapies for cancer.

To achieve this goal, we first performed TCR sequencing on frozen samples collected from CRC patients to profile TCRβ repertoire of tumor filtrating T cells. We next exploited the peptide-HLA (pHLA) and pHLA-TCR binding affinity of peptides with known immunogenicity status from six public databases including 10X Genomics (25–28), McPAS (29), PRIME (30), VDJdb (31), IEDB (32) and TBAdb (33) to develop a predictive algorithm. Subsequently, we employed this algorithm to predict and rank neoantigens by using the TCRβ HLA types and mutation profiles identified from patients' frozen tumor tissues. Finally, we experimentally verified the efficacy of our model by measuring the immunogenicity of the ranked peptides by an ELISpot assay.

Materials and methods

Tumor biopsy and peripheral blood collection

Twenty-eight patients diagnosed with CRC were enrolled in this study at the University Medical Center in Ho Chi Minh City between June 2022 and May 2023. CRC confirmation was based on abnormal colonoscopies and histopathological analysis. The stages of CRC were determined according to guidelines provided by the American Joint Committee on Cancer and the International Union for Cancer Control (the eighth version) (34, 35). Prior to participation, all patients provided written informed consent for tumor and whole blood sample collection. Relevant clinical data, including demographics, cancer stages, and pathology information, were

extracted from the medical records of the University Medical Center. Detailed information regarding the clinical factors of the patients can be found in **Supplementary Table 1**. This study was approved by the Ethics Committee of the University of Medicine and Pharmacy in Ho Chi Minh City, Vietnam. For four patients, ten mL of peripheral blood was collected before surgery and stored in Heparin tubes prior isolation of PBMCs.

Targeted DNA, RNA and TCRβ sequencing

The DNA/RNA samples were isolated using either the AllPrep DNA/RNA Mini Kit or the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen, Germany) as per the manufacturer's protocol. In addition, matched genomic DNA from the white blood cells (WBC) of individuals was also extracted from the buffy coat using the GeneJET Whole Blood Genomic DNA Purification Mini kit (ThermoFisher, MA, USA), following the manufacturer's instructions. Genomic DNA samples from the patients's paired tumor tissues and WBCs were used to prepare DNA libraries for DNA sequencing with the ThruPLEX Tag-seq Kit (Takara Bio, USA). The libraries were then pooled and hybridized with pre-designed probes for 95 targeted genes (Integrated DNA Technologies, USA). This gene panel encompasses commonly mutated genes in CRC tumors, as reported in the Catalog of Somatic Mutations in Cancer (COSMIC) database. The DNA libraries were then subjected to massive parallel sequencing on the DNBSEQ-G400 sequencer (MGI, Shenzhen, China) for paired-end reads of 2×100 bp with a sequencing depth of 10X.

Isolated total RNA was processed using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, MA, USA) to isolate intact poly(A)+ RNA following the manufacturer's instructions. RNA libraries were prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs). These libraries were subsequently sequenced for paired-end reads of 2×100 bp on an MGI system at a sequencing depth of 50X.

For TCRβ library construction, mRNA was utilized with the SMARTer Human TCR a/b Profiling Kit v2 (Takara, USA). The resulting libraries were sequenced for paired-end reads of 2×250 bp on an Illumina system at a sequencing depth of 7.5X.

Microsatellite instability status

Microsatellite instability (MSI) status was determined in both tumor DNA and corresponding DNA from blood, serving as the control, using the MSI Analysis System, version 1.2 (Promega, US). This analysis identified alterations in 5 mononucleotide repeat markers: BAT-25, BAT-26, NR-21, NR-24, and MONO-27. The results were classified as microsatellite stability (MSS) in cases where either none or only one marker was unstable. Samples showing alterations in two or more markers were categorized as microsatellite instability-high (MSI-H).

Variant calling from DNAseq and RNAseq data

The approach of integrating RNA-seq into calling variants was performed as described in our previous publication (36). Briefly, we used Dragen (Illumina) (37) in tumor-normal mode to detect somatic mutations from DNAseq data. The default filtering thresholds of Dragen were applied for the detection of single nucleotide polymorphisms (SNPs) and insertions/deletions (indels). SNPs were further filtered using the dbSNP and 1000 Genome datasets. Germline mutations in tumor tissues were identified by comparison with matched WBC DNA samples. Mutations within immunoglobulin and HLA genes were excluded due to alignment difficulties in these highly polymorphic regions, necessitating specialized analysis tools. Additionally, synonymous mutations were removed from downstream analysis. For analysis, we included somatic mutations that exceeded a minimum threshold of ≥2% variant allele frequency (VAF) in DNA extracted from fresh frozen tissues.

Sequencing reads underwent trimming using Trimmomatic (38) and were then aligned to the human reference genome using STAR (version 2.6.0c) (39). Prior to alignment, raw sequencing reads were subjected to quality checks using FastQC version 0.11.9 (40). VarScan 2 (41), , which accepts both DNA and RNAseq data, was employed to identify mutations in paired tumor and WBC samples across 95 cancer-associated genes. This analysis was carried out in tumor-normal mode. Four filtering steps were applied: (i) only calls with a PASS status were utilized, (ii) population SNPs overlapping with a panel of normal samples from the 1000 Genome dataset were excluded, (iii) somatic mutations included for analysis met a minimum threshold of ≥10 × read depth and ≥2% variant allele frequency (VAF) in RNA extracted from FF tissue, and (iv) synonymous mutations and those related to HLA were removed from downstream analysis.

The resulting BAM files were sorted, indexed using Samtools version 1.10 (42), and had PCR duplicates removed using Picard tools version 2.25.6 (43). Somatic variants were manually reviewed using Integrative Genomics Viewer (v2.8.2). The VCF files generated by Dragen (for DNAseq) and VarScan 2 (for RNAseq) were subsequently annotated using the Ensembl Variant Effect Predictor (VEP version 105) (44) to extract information about the potential effects of variants on the phenotypic outcome.

CDR3β calling from TCRseq data

To define the CDR3β from TCRseq, we utilized Cogent NGS Immune Profiler Software v1.0 (45). Before calling clonotypes, the raw sequencing reads underwent filtering based on the following criteria: (i) allowing only one mismatch while splitting reads by matching read sequences to different receptor chains, (ii) excluding reads shorter than 30 bp and reads ambiguously matched to multiple receptor chains, (iii) excluding reads that failed correction when linker-based correction was enabled, (iv) excluding reads that failed the abundance check during sequencing error correction, and (v) excluding molecular identifier groups (MIGs) with fewer than 3 unique molecular identifier (UMI) reads. The filtered MIGs were then assembled and aligned to the V(D)J reference to define the TCR clonotype.

Shannon index and clonality

We utilized two indices to characterize T-cell diversity and expansion: the Shannon entropy index and the clonality index.

Shannon entropy:

$$-\sum_{i=1}^{n} p_i \log_e (p_i)$$

Clonality:

$$1 - \frac{\sum_{i=1}^{n} p_i \log_e (p_i)}{\log_e (n)}$$

These indices consider both the number of T-cell clone types 'n' and the frequency 'p_i' of each clone. Here, 'p_i' represents the proportion of the i-th clone within the TCR library containing n clones.

In silico prediction of peptide-HLA binding and peptide-HLA-TCR binding

Class I HLA alleles (HLA-A/B/C) with two-digit resolution were identified from patient tumor RNAseq data using the OptiType tool (**Supplementary Table 2)** (46, 47). The annotated VCF files were analyzed using pVACseq, a tool in pVACtools (v1.5.9) (11, 48, 49). We used the default settings, except for disabling the coverage and MAF filters. We used all peptide-HLA-I binding algorithms that were implemented in pVACseq to predict 8 to 11-mer epitopes binding to HLA-I (A, B, or C) for downstream analysis. Mutants with lower binding affinity than wildtype peptides were prioritized using a two-step ranking process. The minimum binding affinity score was calculated from the distribution of scores among the peptides derived from each mutation, and priority was given to mutations with the lowest binding affinity scores. Moreover, in pVACseq, we collected peptides with binding affinity scores lower than wildtype from multiple tools, calculated minimum binding affinity scores for each unique peptide, and incorporated this data into a combined machine learning model.

We used the peptide-HLA-TCR binding algorithm implemented in pMTNet (50) to predict peptide binding to HLA and TCR with default settings. These scores represented the predicted likelihood of peptides being immunogenic. A ranking value for immunogenicity was assigned to each unique peptide by determining the minimum TCR ranking of its immunogenicity score.

Construction of a combined machine learning model

To identify immunogenic peptides, we conducted a thorough search across multiple databases, including 10X Genomics (25–28), McPAS (29), VDJdb (31), IEDB (32), and TBAdb (33). For the creation of non-immunogenic pHLA-TCR complex, we assembled a negative dataset using the PRIME tool (30). This dataset was generated by associating each pHLA with 10 randomly generated TCRs sourced from the following databases: 10X Genomics, McPAS, VDJdb, IEDB, and TBAdb. To objectively train and evaluate the model, we separated the dataset mentioned above into two data subsets: discovery dataset (70%) and validation dataset (30%). These subsets are mutually exclusive and do not overlap (**Supplementary Figure 1**). pHLA-TCR complex in the discovery dataset, whether labeled as immunogenic or non-immunogenic, were used for model training to classify whether a peptide is positive or not. We examined three machine learning algorithms - Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGB) - for each feature type (pHLA binding, pHLA-TCR binding), as well as for combined features. Feature selection was tested using a k-fold cross-validation approach on the discovery dataset with 'k' set to 10-fold. This process splits the discovery dataset into 10 equal-sized folds, iteratively using 9 folds for training and 1 fold for validation. Model performance was evaluated using the 'roc_auc' (Receiver Operating Characteristic Area Under the Curve) metric, which measures the model's ability to distinguish between positive and negative peptides. The average of these scores provides a robust estimate of the model's performance and generalizability. The model with the highest 'roc_auc' average score, XGB, was chosen for all features. The model cut-off was set based on a threshold specificity of >90% or >95% to achieve high specificity for selecting peptides. This combined model's performance was evaluated on the independent validation dataset of immunogenic and non-immunogenic pHLA-TCR complex to assess its ability to classify positive peptides from negative peptides. In our GitHub repository, we have included links to the GitHub repositories for NetMHCpan and pMTNet (*https://github.com/QuynhPham1220/Combined-model*).

Ranking coverage score calculation

The approach of compare ranking between two algorithms was performed as described in a previous publication (51). Briefly, the rank coverage score was based on the ranking value calculation given by the formula:

$$\text{Rank Coverage Score} = \frac{\sum_{n \epsilon \text{negative}} \text{rank (n)}}{T \times \text{num (n)}} \times \text{coverage(n)} - \frac{\sum_{n \epsilon \text{positive}} \text{rank (p)}}{T \times \text{num (p)}} \times \text{coverage(p)}$$

$$\text{coverage (k)} = \frac{\max \text{rank(k)}}{T} \quad k \epsilon (n, p)$$

Where, *T* presented the total neoepitope number identified and *p* and *n* presented the positive and negative peptides, respectively, that were experimentally validated in vitro.

If positive peptides have smaller rank values than negative peptides, this will result in a high rank coverage score, indicates the better ranking result.

Isolation, culture, and stimulation of PBMCs with long peptides

Peripheral blood samples were collected from eight patients prior to surgery using BD Vacutainer Heparin Tubes (BD Biosciences, NJ, USA). Peripheral blood mononuclear cells (PBMCs) were isolated through gradient centrifugation using Lymphoprep (STEMCELL Technologies) within 4 hours of blood collection. The PBMCs were then resuspended in a solution of FBS (10%) and DMSO (7-10×10^6 cells/mL) for cryopreservation in liquid nitrogen. Frozen PBMCs were thawed in AIM-V media (Gibco, Thermo Scientific, MA, USA) supplemented with 10% FBS (Cytiva, USA) and DNase I (Stemcell Technology, Canada) (1 mg/mL) solution. A total of 10^5 PBMCs were allowed to rest in a 96-round bottom well-plate containing AIM V media supplemented with 10% FBS, 10 mM HEPES, and 50 mM β-mercaptoethanol overnight before stimulation with synthesized long peptides at a concentration of 5 mM in a humidified incubator at 37°C with 5% CO2. PBMCs were further stimulated with GM-CSF (2000 IU/mL, Gibco, MT, USA) and IL-4 (1000 IU/mL, Invitrogen, MA, USA) for 24 hours. Following this initial stimulation, LPS (100 ng/mL, SigmaAldrich, MA, USA) and IFN-γ (10 ng/mL, Gibco, MT, USA) were added to the PBMCs along with the peptides for an additional 12 hours. On the following day, IL-7, IL-15, and IL-21 (each at a concentration of 10 ng/mL) (Peprotech, NJ, USA) were added to the PBMC culture. The restimulation process involved exposing the peptides to fresh media containing IL-7, IL-15, and IL-21 every 3 days, for a total of 3 times. On day 12, PBMCs were restimulated with peptides and cultured in media without cytokines. ELISpot assays were performed on stimulated PBMCs on day 13.

ELISpot assay on PBMCs stimulated with long peptides

Cultured T cells were transferred to an enzyme-linked immunospot (ELISpot) plate (Mabtech, Sweden) and incubated for 20 hours at 37°C. Peripheral blood mononuclear cells (PBMCs) cultured with DMSO were used as a negative control group, while PBMCs stimulated with anti-CD3 were used as a positive control group. The ELISpot assay was performed on treated PBMCs using the ELISpot Pro: Human IFN-g (ALP) kit (Mabtech, Sweden), following the manufacturer's protocol. Developed spots on the ELISpot plate were then counted using an ELISpot reader (Mabtech, Sweden). Reactivity was determined by measuring the fold increase in the number of spots in PBMCs treated with mutant peptides compared to those treated with wild-type peptides. A fold change of two was selected as the cutoff for positivity, indicating a significant increase in reactivity (52).

Statistical analysis

In this study, t-tests were used to compare the TCR clones, clonality, and Shannon-index among two groups of four categories (Microsatellite status, stage, gender, tumor location). Chi-square test was used to compare the proportions of immununogenic and non-immunogenic peptides. All

statistical analyses were performed using R (version 4.3.0) with common data analysis packages, including ggplot2 and pROC. The 95% confidence interval (95% CI) was presented in brackets next to a value as appropriate.

Results

The workflow of identifying neoantigens by combining both HLA and TCR binding characteristics

The identification of neoantigens has traditionally heavily relied on peptide-HLA (pHLA) binding prediction while often neglecting the significance of pHLA-TCR interactions (50). In this study, we introduce a novel workflow that integrates both pHLA and pHLA-TCR interactions to enhance the precision of neoantigen identification and prioritization (**Figure 1**). In the first step (**Figure 1A**), we conducted RNA and DNA sequencing on matched tumor tissues and peripheral blood mononuclear cells (PBMCs) collected from 28 CRC patients to detect cancer-associated nonsynonymous mutations and determine HLA types, as detailed in our previous work (36). Additionally, T-cell receptor sequencing (TCR-Seq) was performed to profile the TCRβ repertoire of TILs. In the second step (**Figure 1B**), we used the pVACseq and pMTNet tools to predict the binding affinities of both pHLA and pHLA-TCR interactions. To exploit the information of both pHLA and pHLA-TCR binding for selecting and ranking neoantigen candidates, we constructed a machine learning model by using immunogenic and non-immunogenic peptide information sourced from five publicly available databases (**Figure 1C**). Finally, we designed long peptide (LP) candidates encompassing the selected neoantigens and experimentally assessed their immunogenicity using peripheral blood mononuclear cells (PBMCs) obtained from the same patients (**Figure 1D**).

Heterogenous tumor infiltrating TCRβ profiles in colorectal cancer patients

Autologous TILs have exhibited varying reactivity levels to neoantigens, suggesting the potential of TIL-based recognition to improve the identification of immunogenic neoantigens (53). Furthermore, characterizing T cell receptors (TCRs) can complement efforts to predict immunogenicity. To explore this, we initiated our study by characterizing the TIL repertoire in a cohort of 28 colorectal cancer patients. This characterization involved sequencing the complementarity-determining region 3 (CDR3) of T-cell receptor beta (TCRβ), renowned for its remarkable diversity within the TCR gene. Our sequencing analysis yielded an average of 2,992,949 productive TCR reads per sample, with a range between 256,035 and 10,888,726 (**Supplementary Table 3**), following correction for duplications, sequencing errors, and exclusive use of uniquely barcoded reads mapped to TCRβ CD3 sequences from the ImMunoGeneTics (IMGT) databases. Across the 28 patients, we observed variable numbers of TCRβ-CDR3 clonotypes, ranging from 433 to 27,749 (**Figure 2A & Supplementary Table 3,** 28 patients were arranged in ascending order), indicating the intra-tumor heterogeneity of TCR clonotypes. Of these clonotypes, 59.5% exhibited a single uniquely barcoded read mapped to the TCRβ-CDR3 sequences (depicted in yellow in **Figure 2A**), while 40.5% had at least two TCRβ-CDR3 reads confidently identified (**Figure 2A**). As observed previously (54), the length distribution of CDR3 was approximately normally distributed (median 14, range 4-43, **Supplementary Figure 2A & Supplementary Table 4**). Subsequently, we explored potential associations between TCR diversity, as quantified by the Shannon index, and patient characteristics. In accordance with prior investigations, we identified an inverse relationship between the number of TCR clones and the Shannon index (**Figure 2B**). Based on our current sequencing depth, we have observed that many of our samples (14 out of 28) have reached sufficient saturation (**Figure 2C**) as their diversity of clonotypes was saturated. Additionally, the TCR clones selected in our studies are unique molecular identifier (UMI)-

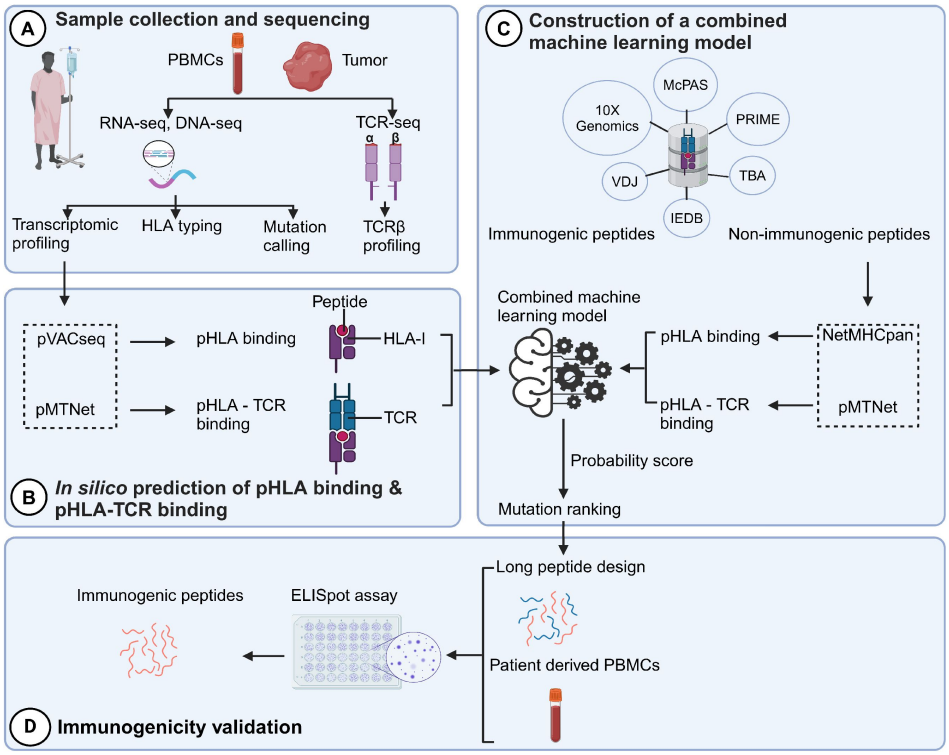


Figure 1.

A novel workflow based on machine learning that integrates TCRβ sequencing data for the identification and ranking of CRC neoantigens.

(A) Tumor biopsies and peripheral blood from CRC patients were subjected to targeted DNA-seq, RNA-seq, and TCR-seq. (B) The prediction of peptide-HLA binding and peptide-HLA-TCR binding by indicated tools using the DNA-seq, RNA-seq, and TCR-seq data was performed. (C) Machine learning models were subsequently constructed based on the analysis of the peptide-HLA binding and peptide-HLA-TCR binding features to distinguish immunogenic antigens from non-immunogenic peptides. The immunogenicity of predicted neoantigen candidates prioritized by the model was validated by ELISpot to evaluate the effectiveness of this approach.

collapsed reads, with each representing at least three raw reads sharing the same UMI. Consequently, despite some samples having low sequencing depth for TCRβ sequencing, likely due to the diversity of TIL infiltration between patients, our TCRβ profiling analysis is robust and reliable. However, we did not detect any significant correlations between TCR clonality or diversity and clinical variables, including microsatellite status, tumor staging, patient gender and tumor location (**Supplementary Figure 4**). In contrast, MSI-H tumors, which are known to be rich in neoantigens (55, 56), displayed a significantly lower number of TCRβ clonotypes compared to MSS tumors (2181 versus 5330, p=0.057, **Supplementary Figure 3 & 4A**), aligning with previous research (57).

After filtering out TCR clones with read counts below 1, we obtained a total of 74,590 TCR clones. Subsequently, we conducted a comprehensive assessment of TCRβ repertoire similarity by calculating the proportion of overlapping TCRβ CDR3 clones across the 28 patients. The preeminent majority, constituting 95.1% of all identified TCRβ-CDR3 clones, comprised unique clonotypes found in one patient, while the remaining 4.9% were recurrently observed in at least two patients (**Supplementary Figure 2B**). This observation underscores the substantial inter-patient heterogeneity in TCR profiles. The TCRβ repertoire is generated by the random rearrangement of variable (V), diversity (D) and joining (J) segments. We conducted a search specifically targeting the D segment by scoring the similarities of sub-sequences within the junction sequences against the reference sequences of D segments. Consistent with a previous study, we were unable to unambiguously assign the D segments to any defined D region usage due to the truncation of this region (58). In contrast to the diversified D segments, V and J sequments displayed high recurrent rates across 28 patients (**Supplementary Figure 2B**). As anticipated, we identified 59 distinct V segments (**Supplementary Figure 2C**) and 13 distinct J segments (**Supplementary Figure 2D**), collectively sharing 185,627 clones across the 28 tumor tissue samples. This underscores the conservation of these segments (**Supplementary Figure 2C** & **D**). Conversely, we observed a varied combination of V and J segments, which significantly contributes to the heterogeneity of the TIL TCRβ repertoire (**Supplementary Figure 2E**). Collectively, our data elucidate the presence of both intra-tumor and inter-patient heterogeneity in the TCRβ repertoires of CRC patients, likely stemming from the stochastic utilization of V and J segments in response to neoantigens.

pHLA and pHLA-TCR interactions are two complement determinants of neoantigen immunogenicity

While *in silico* tools predicting HLA-peptide binding affinity have traditionally played a pivotal role in determining neoantigen immunogenicity (59), the evaluation of TCR-peptide binding for screening immunogenic neoantigens remains understudied. In light of this, our study aimed to assess the contributions of both pHLA and pHLA-TCR binding affinity in predicting immunogenic neoantigens. To accomplish this, we gathered HLA and TCRβ sequences from established datasets containing immunogenic and non-immunogenic pHLA-TCR complexes (**Supplementary Table 5)**. Subsequently, we employed NetMHCpan (13) and pMTNet (50) tools to predict pHLA and pHLA-TCR binding, respectively.

For comparative purposes, we generated plots depicting predicted percentile rank values, with lower percentile ranks signifying stronger binding affinity. As anticipated, our analysis revealed a significantly higher prevalence of peptides with robust HLA binding (percentile rank < 2%) among immunogenic peptides in contrast to their non-immunogenic counterparts (**Figure 3A** & **B**, p< 0.00001). Similarly, immunogenic peptides exhibited a greater proportion of peptides with percentile ranks indicating pHLA-TCR binding of < 2% compared to non-immunogenic peptides (**Figure 3C** & **D**, p=0.086). As recommended by NetMHCpan and pMTNet, we considered peptide candidates with predicted percentile ranks below 2% as binders. Utilizing this predefined threshold, both pHLA and pHLA-TCR binding affinity exhibited comparable positive predictive values (PPV), at 68.5% (**Figure 3B**) and 64.3% (**Figure 3D**), respectively. This substantiates the

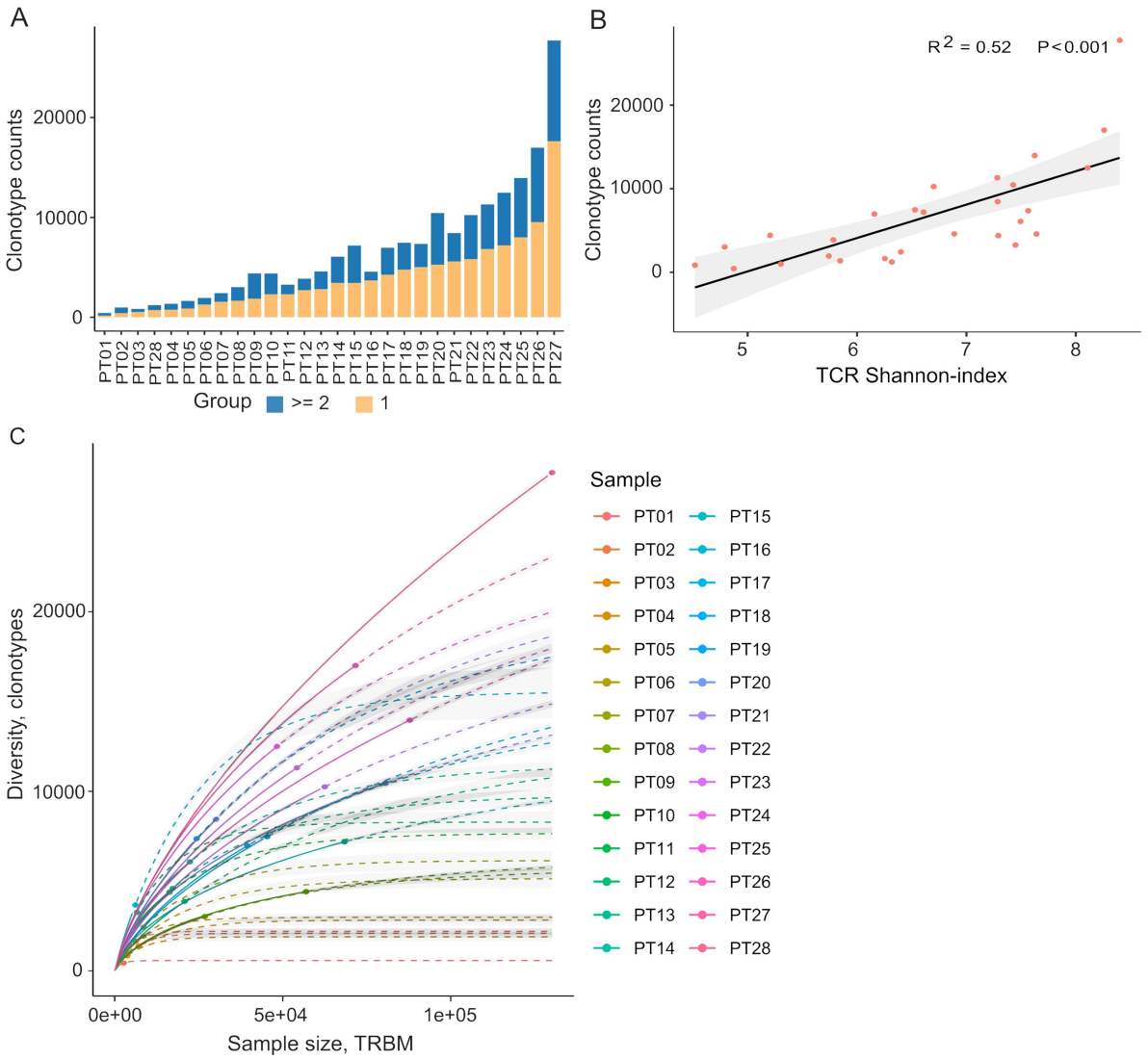


Figure 2.

Tumor-infiltrating TCRβ profiles in 28 colorectal cancer patients.

(A) A bar plot depicting the distribution of TCR clonotypes among 28 CRC patients, categorized into two groups: those with a unique read count and those with read counts greater than or equal to 2 for each TCR clonotype. (B) The scatter plot illustrates the relationship between Shannon-index and the number of TCR clones. (C) The rarefaction plot shows the variable between sample size and diversity among 28 CRC samples.

significance of both pHLA and pHLA-TCR binding in determining the immunogenicity of peptides. When we simultaneously applied the cutoff values for both HLA-peptide and TCR-peptide percentile ranks (Q1 group, **Figure 3E**), the PPV increased to 76.9% from 68.5% and 64.3% for individual binding features (**Figure 3F**). This underscores the rationale for combining these two criteria to enhance the accuracy of neoantigen prediction. Notably, pHLA-TCR binding displayed a remarkedly lower sensitivity but higher specificity in comparison to pHLA binding (**Figure 3G**), implying their potential as a complementary criteria for the selection of immunogenic peptides.

Combination of peptide-HLA and peptide-HLA-TCR interactions improves neoantigen prediction

The combination of peptide-HLA and peptide-HLA-TCR binding ranking with a fixed cutoff values of 2% for each feature resulted in high specificity but low sensitivity (**Figure 3G**). To optimize the precision of immugenic neoantigens, we examined three distinct machine learning classifiers, namely Random Forest (RF), Logistic Regression (LR), and Extreme Gradient Boosting (XGB) classifiers. We first partitioned the pHLA binding and pHLA-TCR binding ranks for immunogenic and non-immunogenic peptides from publicly available databases into separate discovery and validation datasets. These datasets were used to develop and validate machine learning algorithms, as illustrated in **Figure 4A**. We next assessed the performance of the three examined algorithms, employing a 10-fold cross-validation strategy. Among these algorithms, XGB demonstrated the highest performance, achieving an area under the receiver operating characteristic curve (AUC) of 0.82 in the training dataset and 0.84 in the validation dataset (**Supplementary Figure 5**). Consequently, the model combining pHLA and pHLA-TCR binding, referred to as the 'combined model' was chosen for further validation.

Our findings revealed that the combined model outperformed methods relying on pHLA-TCR or pHLA binding feature separately. The combined model yielded sensitivity AUC values of 0.82 (95% CI 0.81-0.84) and 0.84 (95% CI 0.82-0.86) for discovery and validation cohorts, whereas the pHLA-TCR feature alone achieved AUC values of 0.69 (95% CI 0.66-0.71) and 0.74 (95% CI 0.71-0.77). Meanwhile, the pHLA feature resulted in AUC values of 0.76 (95% CI 0.75-0.78) and 0.74 (95% CI 0.72-0.77) (**Figure 4B**). In order to address the elevated false positive rates associated with current prediction tools for neoantigens, we set the specificity at high thresholds of >95% and >99%. We observed that the combined model achieved a greater sensitivity (39.7% versus 5.9% and 19.7% at > 95% specificity; 47.1% versus 3.5% and 18.8% at > 99% specificity, **Figure 4C**), Negative Predictive Value (NPV, 87.7% versus 82% and 84.2% at > 95% specificity; 88.9% versus 81.5% and 83.9% at > 99% specificity, **Figure 4D**), and Positive Predictive Value (PPV, 65.9% versus 22.2% and 47.1% at > 95% specificity; 62.3% versus 11.8% and 39.4% at > 99% specificity, **Figure 4E**) compared to the single feature methods.

The accurate prioritization of neoantigen candidates with high immunogenicity holds the potential to streamline the validation process, reducing both costs and time expenditures. Consequently, we proceeded to evaluate the combined model's capacity to prioritize neoantigens by computing the ranking coverage score, which considers the accuracy in ranking immunogenic peptides versus non-immunogenic peptides (51). In the validation phase, the combined model exhibited superior rank coverage scores in comparison to the individual feature-based methods. The combined model attained a ranking coverage score of 0.37, while the single-feature methods, pHLA-TCR and pHLA, yielded scores of -0.26 and 0.25, respectively (**Figure 4F**). These findings underscore the notion that the incorporation of pHLA and pHLA-TCR binding criteria can enhance the accuracy of prediction and prioritization of immunogenic neoantigens.

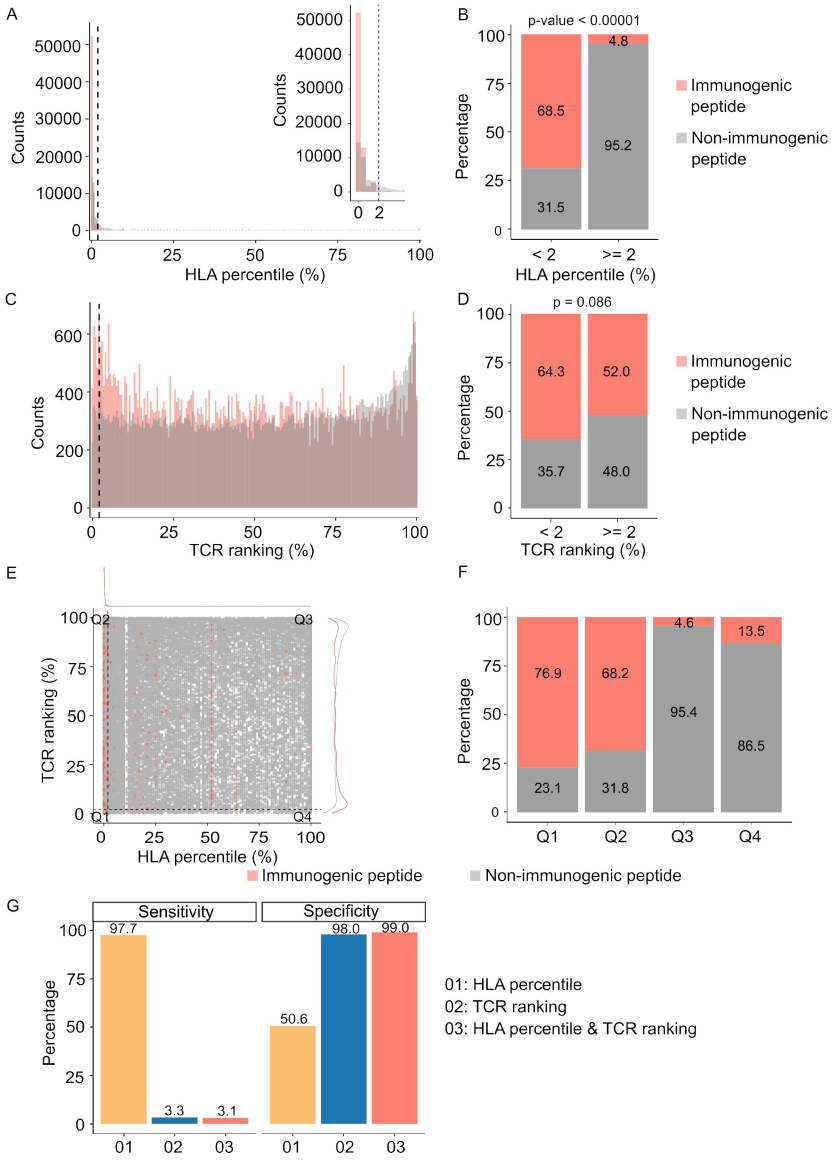


Figure 3.

Peptide-TCR and peptide-HLA interactions are two complementary determinants of neoantigen immunogenicity.

(A) The histogram displays the HLA percentile distribution of immunogenic antigens (red bar) and non-immunogenic peptides (grey bar). (B) The percentage of immunogenic antigens (red bar) and non-immunogenic peptides (grey bar) is compared between two groups based on HLA percentile: <2% and >= 2% (Chi-square test, p< 0.00001). (C) The histogram displays the TCR ranking distribution of immunogenic antigens (red bar) and non-immunogenic peptides (grey bar). (D) The percentage of immunogenic antigens (red bar) and non-immunogenic peptides (grey bar) is compared between two groups based on TCR ranking: <2% and >= 2% (Chi-square test, p=0.086). (E) The scatter plot illustrates the relationship between the HLA percentile distribution and TCR ranking of immunogenic antigens (red bar) and non-immunogenic peptides (grey bar). (F) The percentage of immunogenic antigens (red bar) and non-immunogenic peptides (grey bar) is analyzed in four distinct groups based on cutoffs of HLA percentile and TCR ranking. (G) The bar plot illustrates the sensitivity and specificity of three neoantigen prioritization approaches: based on neoantigen-HLA binding affinity alone (yellow bar), neoantigen-TCR binding ranking alone (blue bar), and the combined method using both features (red bar).

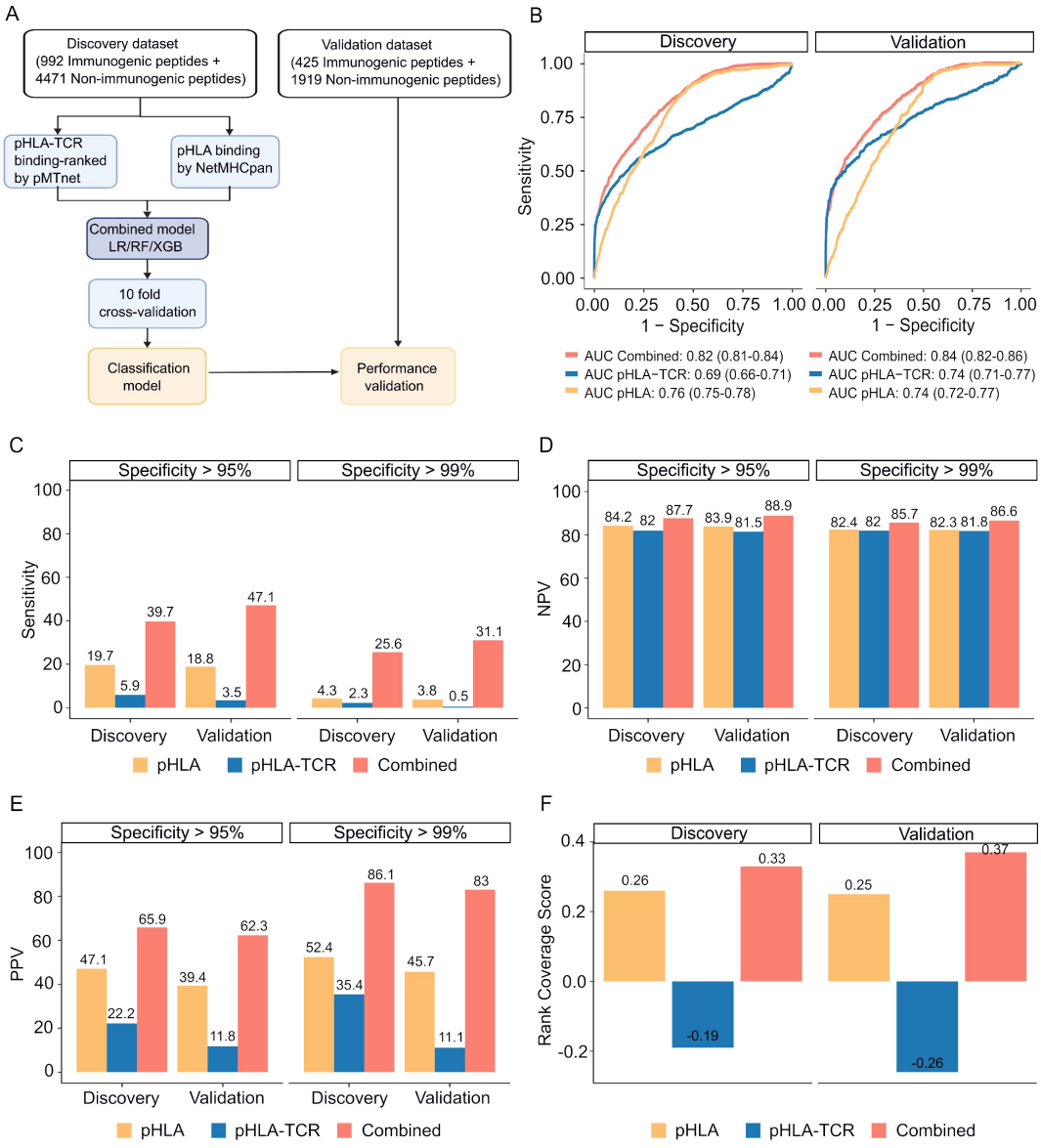


Figure 4.

The combined model demonstrates improved sensitivity and specificity for neoantigen prioritization.

(A) The workflow for constructing the model. (B) The ROC curves demonstrate the performance of both the combined model and individual models in both the discovery and validation cohorts. The bar graphs illustrate the sensitivity (C), negative predictive value (NPV) (D), and positive predictive value (PPV) (E) at specificity levels of at least 95% or 99% for both the combined and individual models in both the discovery and validation cohorts. (F) Ranking coverage scores for the specified models in either the discovery or validation cohorts.

Experimental validation of the pHLA-TCR and pHLA combined approach in selecting neoantigen candidates in patients with CRC

To experimentally validate the efficacy of our combined approach, we conducted a comparative analysis of the percentage of confirmed immunogenic neoantigen candidates and the ranking coverage scores between the conventional NetMHCpan method, which relies solely on pHLA binding percentiles, and our combined approach (**Figure 5A**). Due to the limited availability of patient blood samples, we were only able to perform validation on eight patients who possessed sufficient quantity of PBMCs. For each patient, we chose the top three neoantigen candidates predicted by each method. Among the neoantigen candidates, three were found to be common between the two methods, while twenty neoantigen candidates were uniquely identified by either NetMHCpan or the combined model (**Figure 5B**), bringing the total number of candidates to 44. Subsequently, we synthesized 44 long peptides (LPs) covering these 44 selected neoantigen candidates, along with 44 LPs corresponding to the wild-type sequences. The LPs were utilized in an *ex vivo* ELISpot assay to measure the release of interferon-gamma (IFNγ) from the four patients' PBMCs. If the stimulation with a mutant LP resulted in a > 2-fold increase in IFNγ spots compared to its corresponding wildtype LP, the respective neoantigen candidate was classified as an immunogenic neoantigen.

Out of the 44 selected LPs, we confirmed the immunogenicity of three LPs (RNF213_W719S, MAP3K1_L1202F and TRRAP_T148TEL) identified by the NetMHCpan method and seven LPs (NCOR2_R1963H, TRRAP_F1568S, DICER1_P1592L, KMT2C_K3848T, BRAF_EV275-276-, SMAD4_G230R and PTPN13_D184A) selected by the combined method (**Figure 5C** & **5D**). Notably, six patients exhibited at least one immunogenic peptide identified by the combined method, whereas none of the LPs in five patients, PT07, PT18, PT21, PT24 and PT25, chosen via the NetMHCpan method were validated as immunogenic (**Figure 5C** & **5D**). To further assess the ranking accuracy of these two methods, we calculated the ranking coverage scores. In agreement with our *in silico* analyses, we observed higher rank coverage scores for the combined method in five out of the eight patients, resulting in an overall score of 0.04 (**Figure 5E & Supplementary Figure 6**). In contrast, the NetMHCpan method exhibited a lower rank coverage score of -0.37 (**Figure 5E & Supplementary Figure 6**). To further evaluate our model, we gathered additional public data and assessed its effectiveness in comparison to other models. We utilized immunogenic peptides from databases such as NEPdb (60), NeoPeptide (61), dbPepneo (62), Tantigen (63), and TSNAdb (64), ensuring there was no overlap with the datasets used for training and validation. For non-immunogenic peptides, we used data from 10X Genomics Chromium Single Cell Immune Profiling (25–28). The findings indicate that the combined model from pMTNet and NetMHCpan outperforms NetTCR tool (65) (**Supplementary Table 6**). These outcomes conclusively demonstrate the ability of the combined approach to enhance the prediction and ranking of immunogenic neoantigens in cancer patients.

Discussion

The selection of neoantigens plays a pivotal role in enhancing the efficacy of personalized treatments in cancer immunotherapy. Historically, neoantigen selection has predominantly hinged upon the prediction of peptide-human leukocyte antigen (pHLA) binding. However, the limitations of this approach have become increasingly evident (66), as it often neglects the dynamic interplay between tumor cells and the immune system. The findings of our study shed new light on this critical aspect of neoantigen selection. Our research highlights the potential of integrating pHLA binding prediction with the assessment of pHLA-Tumor-Infiltrating Lymphocyte T Cell Receptor (TIL TCR) binding (**Figure 1**). By encompassing the interaction between neoantigens and the tumor microenvironment, we have demonstrated a substantial enhancement in the

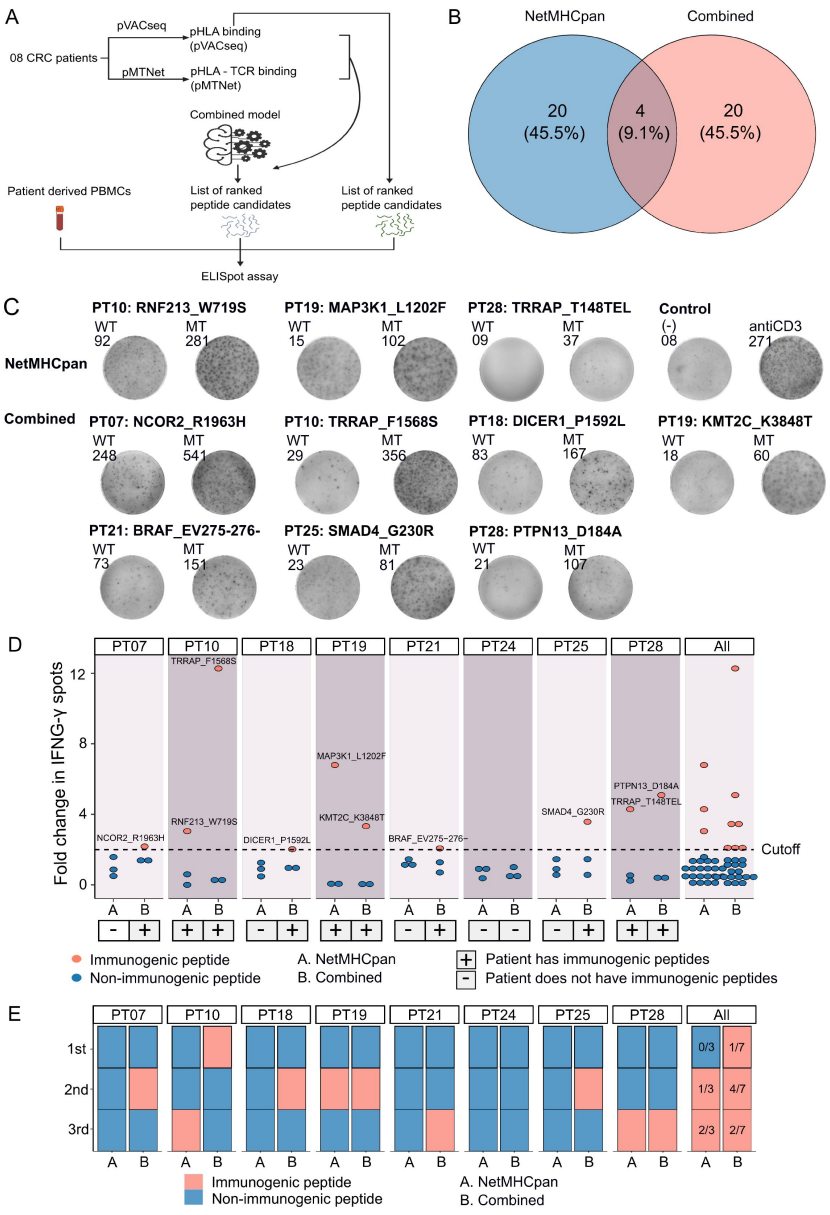


Figure 5.

Validation of neoantigens identified *in silico* from the novel workflow through ELISpot assays conducted on four CRC patients.

(A) A schematic diagram illustrates the procedural steps of neoantigen prioritization and the ELISpot assay. (B) The count of neoantigens identified from each pipeline. (C) The fold change in IFN-γ spots, relative to the wildtype peptides, is shown for 21 long peptides. Note: Only the mutants that result in a positive value in ELISpot are depicted, along with their corresponding amino acid changes and their associated rankings. (D) ELISpot assays on six long peptides resulting in at least a 2-fold change in IFN-γ spots. (E) The bar graphs display ranking of validated long peptides identified from the NetMHCpan tool (blue bar) or the combined method (red bar) for individual patients and all patients.

accuracy of neoantigen selection and prioritization. Our findings underscore the significance of comprehensive neoantigen assessment in harnessing the full potential of immunotherapeutic strategies in the fight against cancer.

Our investigation into the TIL TCRβ repertoire across 28 CRC patients has unveiled a complex tapestry of intra-tumor and inter-patient heterogeneity (**Figure 2A**, **Supplementary Figure 2B**). One underlying cause of this diversity is the random rearrangement of variable (V) and joining (J) segments, which contributes to the distinct TCRβ sequences observed (**Supplementary Figure 2C, D &E**). This remarkable diversity in the TCRβ profile, as elucidated in our study, presents a potential link to the heterogeneity of cancer mutations and neoantigens within CRC. The intricate relationship between the TCRβ repertoire and the genetic alterations within the tumor microenvironment underscores the need for a more comprehensive understanding of this dynamic interplay to develop more precise immunotherapeutic strategies (67). Furthermore, our research revealed a compelling observation regarding patients with MSI-H, who are known to exhibit high mutation and neoantigen burdens (68, 69). Intriguingly, MSI-H patients with strong immune reactivity to neoantigens was shown to display a lower number of TCR clonotypes and reduced diversity, as characterized by the Shannon index, compared to CRC patients with Microsatellite Stable (MSS) (**Supplementary Figure 4A**). This stark contrast in TCRβ diversity hints at the potential enrichment of neoantigen-reactive TCR clonotypes in MSI-H patients. These findings provide valuable insights into the immune responses to different subtypes of CRC and may guide the development of more tailored immunotherapies for patients with differing mutational landscapes.

By utilizing proven immunogenic and non-immunogenic peptides from public databases, we have demonstrated that predicting the strength of pHLA-TCR binding can be a crucial criterion for selecting immunogenic candidates (**Figure 3**). For the prediction of pHLA-TCR binding, we employed the well-established tool pMTnet (50). This choice was driven by its proven efficacy, based on its consistently high performance across different validated datasets (50). Although it exhibited lower sensitivity and positive predictive value (PPV) when compared to pHLA binding alone, pHLA-TCR binding strength exhibited notably superior specificity (**Figure 3G**). This implies that incorporating pHLA-TCR binding strength as a selection criterion would result in a reduced false positive rate, a crucial factor in the context of neoantigen selection. What makes our study particularly compelling is the convergence of these observations. The integration of both criteria, specifically, pHLA binding and pHLA-TCR binding strength, emerged as a strategy that not only enhanced the PPV but also fortified the rationale for combining these two features in the neoantigen selection process.

The integration of both pHLA binding and pHLA-TCR binding strength features in our approach exhibited superior performance in neoantigen selection and prioritization when compared to the single-feature method (**Figure 4**). This observation aligns with the findings of previous studies, which consistently indicate that combining multiple criteria enhances the accuracy and efficacy of neoantigen identification (51, 70). Furthermore, our experimental validation confirmed the robustness of our approach by consistently demonstrating performance consistent with our analysis on publicly available data (**Figure 5 & Supplementary Table 6**). When examining the percentile ranks of positive LPs predicted by netMHCpan from those predicted by our combined model, the results further underscored the strength of our approach (**Figure 5E**). This alignment between experimental validation and computational analysis enhances the reliability and applicability of our neoantigen selection method. Interestingly, the four identified neoantigen candidates with confirmed immunogenicity by the combined approach have not been previously reported in public neoantigen databases, indicating that they could serve as novel targets for neoantigen based therapies.

However, several limitations must be acknowledged. Firstly, the relatively small sample size employed for validation raises the potential for ranking score bias. Additionally, while both TCRα and TCRβ regions play essential roles in engaging peptide-bound HLA complexes, our study focused solely on TCRβ sequences to predict pHLA-TCR binding strength. Future investigations should include TCRα sequences to provide a more comprehensive analysis. Although the PBMC and TILs were previously shown to be congruent in neoantigen reactivity (71), it is important to recognize that differences in the contribution of TIL and PBMC TCR repertoires in neoantigen selection may introduce variability in the selection and validation of neoantigen candidates, and future studies are warranted to address the consistency in neoantigen reactivity between these two sources of T cells. Moreover, to improve the accuracy and effectiveness of the machine learning model in predicting and ranking neoantigens, we have developed an in-house tool called EpiTCR. This tool will utilize immunogenic assays, such as ELISpot and single-cell sequencing, for validation.

In summary, our study delves into the diversity and variation within the TCRβ repertoire of TILs in the tumor tissues of colorectal cancer patients. Through our research, we have introduced a novel approach for the identification and prioritization of neoantigen candidates. This approach combines the assessment of peptide-human leukocyte antigen (pHLA) binding and peptide-human leukocyte antigen-T cell receptor (pHLA-TCR) binding. Our findings underscore the significance of considering pHLA-TCR binding interactions as part of the process for selecting neoantigen candidates. This is a crucial step in the development of personalized immunotherapy strategies. By combining these two factors, we can more accurately identify the neoantigens that hold promise for effective immunotherapy, ultimately improving the prospects for tailored and effective cancer treatment.

Author contributions

Pham Thi Mong Quynh, Nguyen Thanh Nhan, Nguyen Bui Que Tran, Nguyen Hoang Thien Phuc, Tran Thi Phuong Diem, Pham Nguyen My Diem, Ho Thi Kim Cuong, and Nguyen Dinh Viet Linh were engaged in formal analysis, data curation, and methodology development for the manuscript draft. Huu Thinh Nguyen provided conceptualization and patient consultation. Duc Huy Tran, Tran Thanh Sang, Truong-Vinh Ngoc Pham, and Minh-Triet Le were responsible for patient recruitment and tissue histological analysis. Nguyen Thi Tuong Vy contributed to formal analysis and data curation. Minh-Duy Phan supported in conceptualization and manuscript editing. Hoa Giang contributed to conceptualization and methodology. Hoai Nghia Nguyen provided supervision and conceptualization guidance. Le Son Tran contributed to conceptualization and manuscript writing.

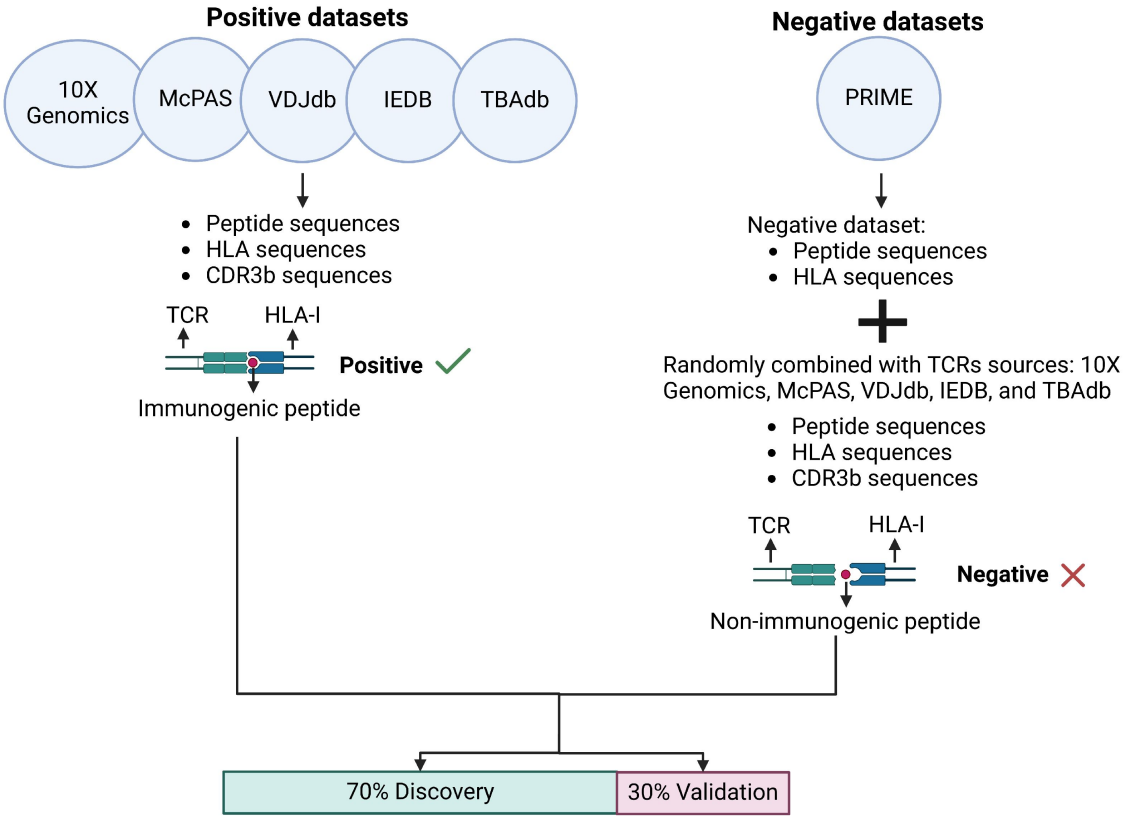
Acknowledgements

The authors thank all participants who agreed to take part in this study. We thank Dr. Kien Nguyen for proofreading our manuscript.

Financial & competing interests' disclosure

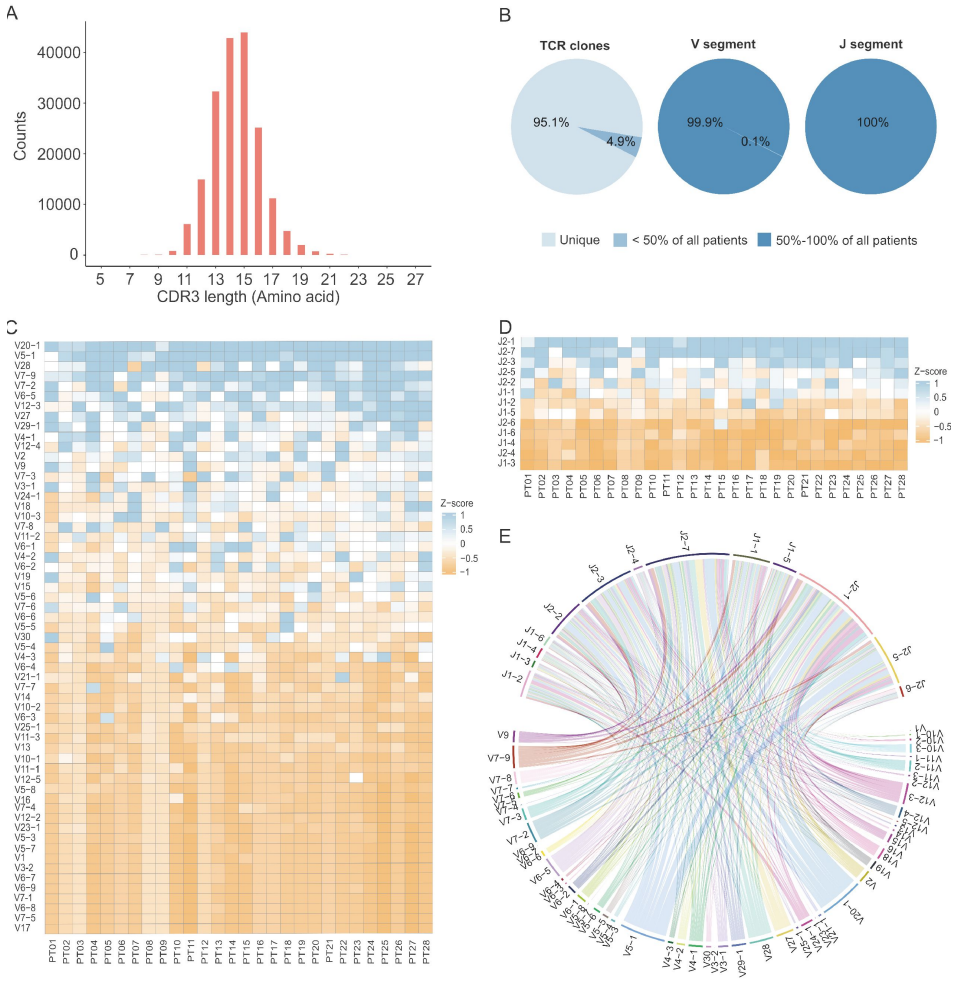
This research was funded by a NexCalibur Therapeutic grant (NC01). The authors including Le Son Tran and Minh Duy Phan hold the equity in NexCalibur Therapeutic.

Supplementary figures



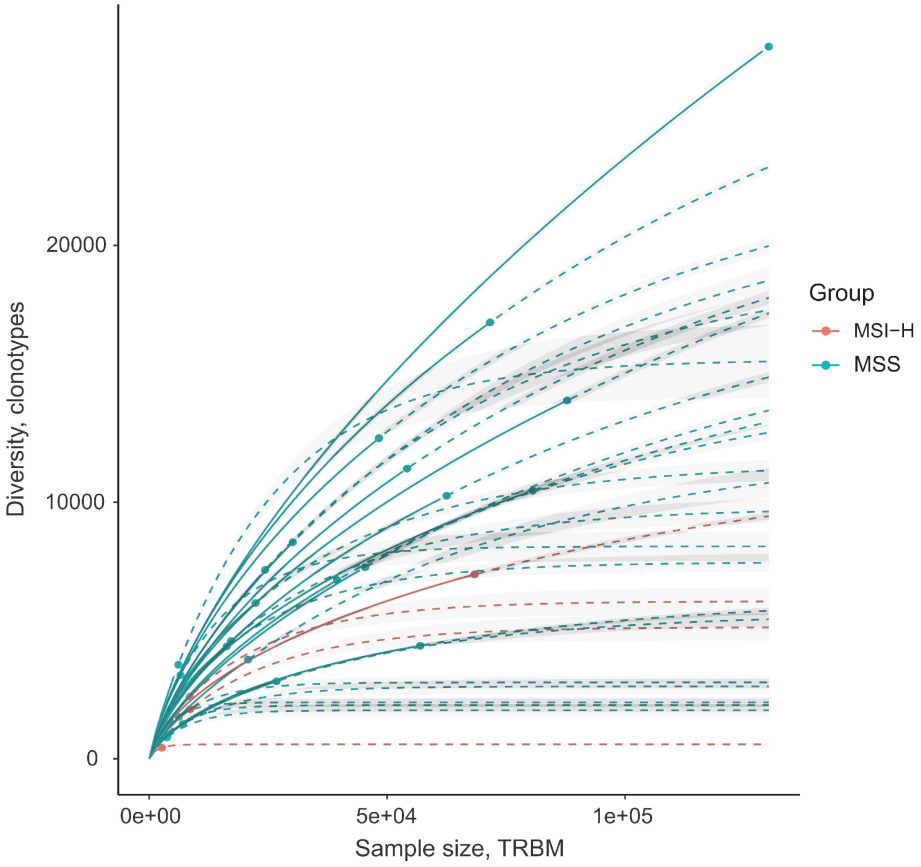
Supplementary Figure 1.

Dataset construction workflow.



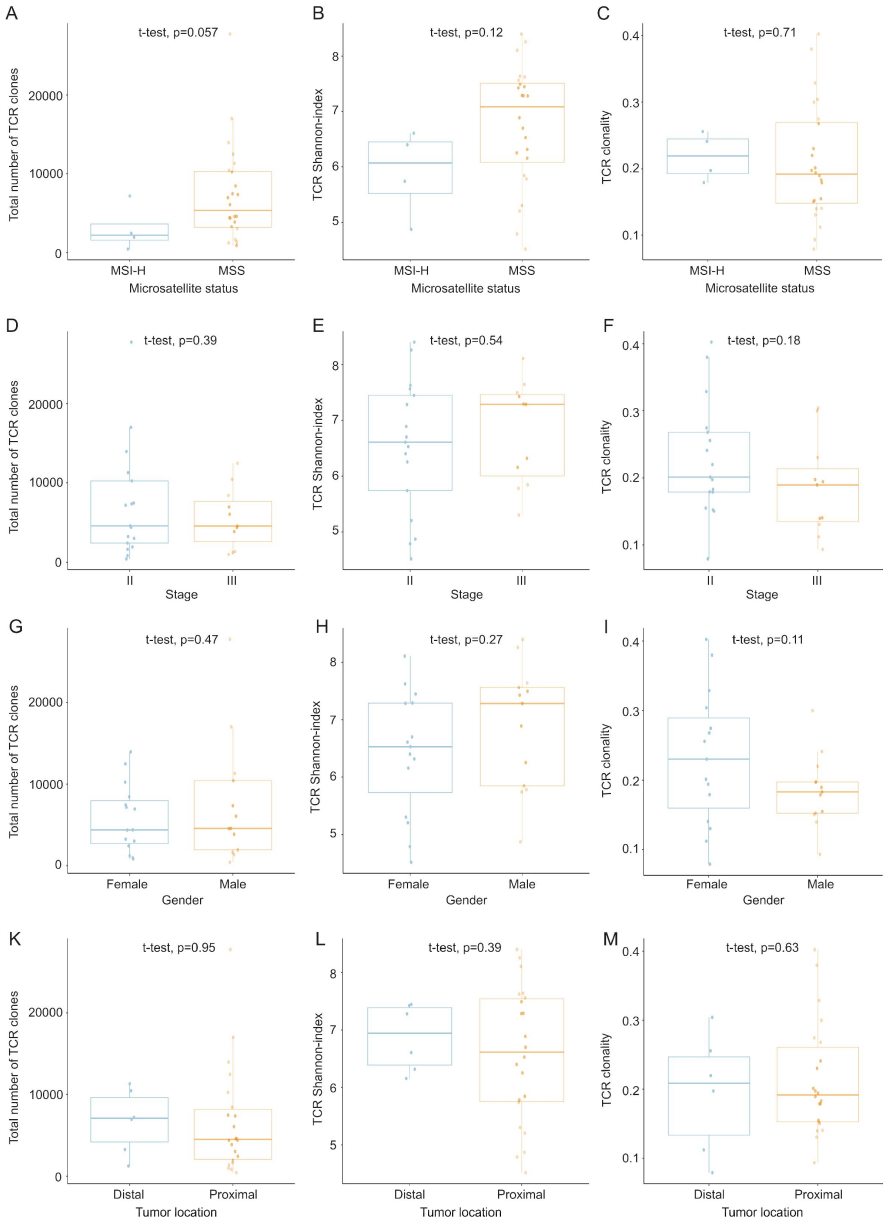
Supplementary Figure 2.

Quality control metrics for Tumor-Infiltrating Lymphocyte (TIL) TCRβ analysis.
(A) Distribution of CDR3β lengths in total TCR clones. (B) The pie chart displays the recurrence rates of TCR clones, V segments, and J segments when the read count of TCR clones exceeds 01. The graph illustrates the uniqueness of TCR clones and the shared presence of both V and J segments. (C) The heatmap depicts the Z-scored read counts of V segments or (D) J segments across 28 samples. Some V and J segments were found to be dominant in all samples. (E) The chord diagram illustrates the rearrangement of V and J segments, revealing random V and J combinations, with a few combinations exhibiting high frequencies.



Supplementary Figure 3.

Rarefaction between MSI and MSS samples.
The rarefaction plot illustrates the sample size and diversity of samples in two groups: MSI and MSS.

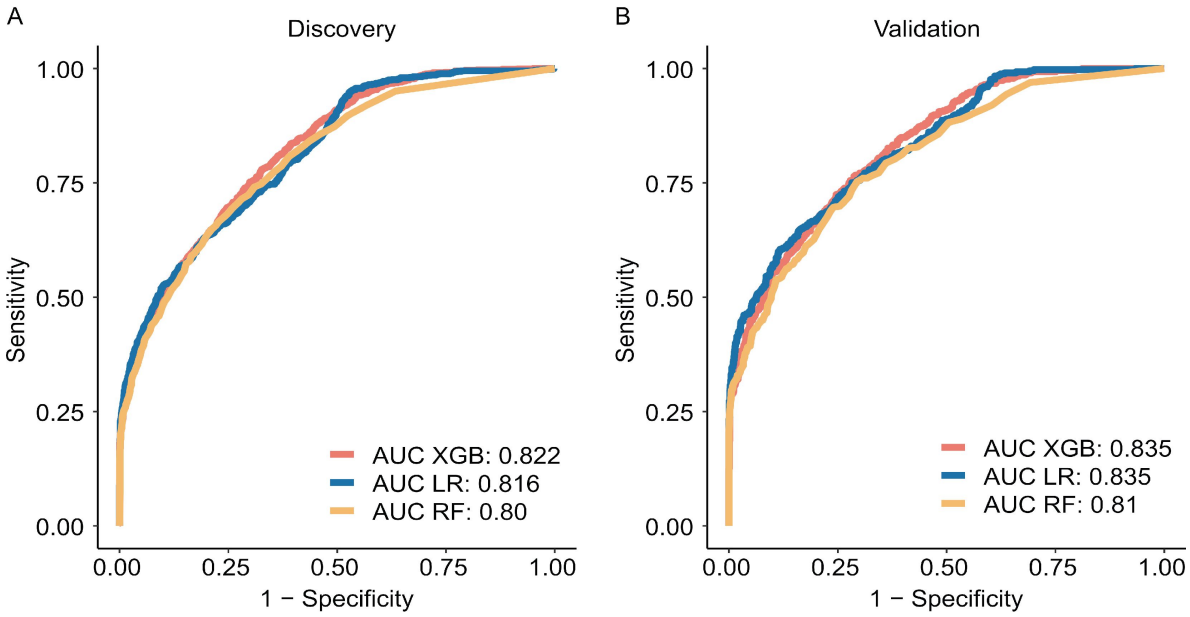


Supplementary Figure 4.

Association between TIL TCRβ profiles and patients' characteristics.
The bar plot and dot plot compare TCR clones, Shannon-index, and clonality between MSI-H and MSS (A, B, C), stage II and III (D, E, F), female and male gender (G, H, I), and distal and proximal tumor locations (K, L, M).

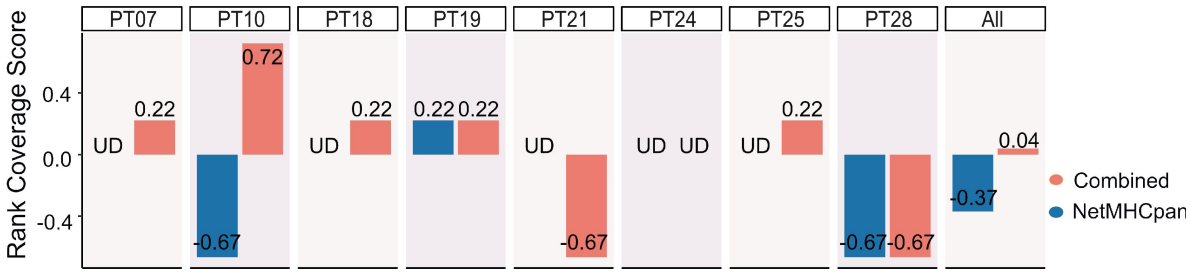
Supplementary Figure 5.

The performance of three machine learning models with three different algorithms is evaluated using receiver operating characteristic (ROC) curves.
The curves depict the performance of the combined model in the discovery cohort (A) and the validation cohort (B)



Supplementary Figure 6.

The rank coverage score of the combined model compared to NetMHCpan.
The bar graphs display rank coverage scores of validated long peptides identified by the NetMHCpan tool (blue bars) and the combined method (red bars) for individual patients and all patients collectively.



References

1. Ganesh K., et al. (2019) **Immunotherapy in colorectal cancer: rationale, challenges and potential** *Nat Rev Gastroenterol Hepatol* **16**:361–375

2. Dudley J. C., Lin M. T., Le D. T., Eshleman J. R. (2016) **Microsatellite Instability as a Biomarker for PD-1 Blockade** *Clin Cancer Res* **22**:813–820

3. Le D. T., et al. (2015) **PD-1 Blockade in Tumors with Mismatch-Repair Deficiency** *N Engl J Med* **372**:2509–2520

4. Overman M. J., et al. (2017) **Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study** *Lancet Oncol* **18**:1182–1191

5. Le D. T., et al. (2017) **Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade** *Science* **357**:409–413

6. Yu Y., et al. (2022) **Neoantigen-reactive T cells exhibit effective anti-tumor activity against colorectal cancer** *Hum Vaccin Immunother* **18**:1–11

7. Kim V. M., et al. (2020) **Neoantigen-based EpiGVAX vaccine initiates antitumor immunity in colorectal cancer** *JCI Insight* **5**

8. Blass E., Ott P. A. (2021) **Advances in the development of personalized neoantigen-based therapeutic cancer vaccines** *Nat Rev Clin Oncol* **18**:215–229

9. Miao D., et al. (2018) **Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma** *Science* **359**:801–806

10. Yarchoan M., Hopkins A., Jaffee E. M. (2017) **Tumor Mutational Burden and Response Rate to PD-1 Inhibition** *N Engl J Med* **377**:2500–2501

11. Hundal J., et al. (2020) **pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens** *Cancer Immunol Res* **8**:409–420

12. Chheda Z. S., et al. (2018) **Novel and shared neoantigen derived from histone 3 variant H3.3K27M mutation for glioma T cell therapy** *J Exp Med* **215**:141–157

13. Reynisson B., Alvarez B., Paul S., Peters B., Nielsen M. (2020) **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* **48**:W449–W454

14. Schumacher T. N., Schreiber R. D. (2015) **Neoantigens in cancer immunotherapy** *Science* **348**:69–74

15. Chen I., Chen M. Y., Goedegebuure S. P., Gillanders W. E. (2021) **Challenges targeting cancer neoantigens in 2021: a systematic literature review** *Expert Rev Vaccines* **20**:827–837

16. Szeto C., Lobos C. A., Nguyen A. T., Gras S. (2020) **TCR Recognition of Peptide-MHC-I: Rule Makers and Breakers** *Int J Mol Sci* **22**

17. Guerder S., Flavell R. A. (1995) **T-cell activation. Two for T** *Curr Biol* **5**:866–868

18. Kuhns M. S., Badgandi H. B. (2012) **Piecing together the family portrait of TCR-CD3 complexes** *Immunol Rev* **250**:120–143

19. Rast J. P., et al. (1997) **, alpha, beta, gamma, and delta T cell antigen receptor genes arose early in vertebrate phylogeny** *Immunity* **6**:1–11

20. Rosati E., et al. (2017) **Overview of methodologies for T-cell receptor repertoire analysis** *BMC Biotechnol* **17**

21. Wucherpfennig K. W., Gagnon E., Call M. J., Huseby E. S., Call M. E. (2010) **Structural biology of the T-cell receptor: insights into receptor assembly, ligand recognition, and initiation of signaling** *Cold Spring Harb Perspect Biol* **2**

22. Porciello N., Franzese O., D'Ambrosio L., Palermo B., Nistico P. (2022) **T-cell repertoire diversity: friend or foe for protective antitumor response?** *J Exp Clin Cancer Res* **41**

23. Lu Y. C., et al. (2021) **Direct identification of neoantigen-specific TCRs from tumor specimens by high-throughput single-cell sequencing** *J Immunother Cancer* **9**

24. Mazzotti L., et al. (2022) **T-Cell Receptor Repertoire Sequencing and Its Applications: Focus on Infectious Diseases and Cancer** *Int J Mol Sci* **23**

25. 10x Genomics (2019) **10x Genomicshttps://www.10xgenomics.com/resources/datasets/cd-8-plus-t-cells-of-healthy-donor-1-1-standard-3-0-2. 2019**

26. 10x Genomics (2019) **10x Genomicshttps://www.10xgenomics.com/resources/datasets/cd-8-plus-t-cells-of-healthy-donor-2-1-standard-3-0-2. 2019**

27. 10x Genomics (2019) **10x Genomicshttps://www.10xgenomics.com/resources/datasets/cd-8-plus-t-cells-of-healthy-donor-3-1-standard-3-0-2. 2019**

28. 10x Genomics (2019) **10x Genomicshttps://www.10xgenomics.com/resources/datasets/cd-8-plus-t-cells-of-healthy-donor-4-1-standard-3-0-2. 2019**

29. Tickotsky N., Sagiv T., Prilusky J., Shifrut E., Friedman N. (2017) **McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences** *Bioinformatics* **33**:2924–2929

30. Schmidt J., et al. (2021) **Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting** *Cell Rep Med* **2**

31. Shugay M., et al. (2018) **VDJdb: a curated database of T-cell receptor sequences with known antigen specificity** *Nucleic Acids Res* **46**:D419–D427

32. Vita R., et al. (2019) **The Immune Epitope Database (IEDB): 2018 update** *Nucleic Acids Res* **47**:D339–D343

33. Zhang W., et al. (2020) **PIRD: Pan Immune Repertoire Database** *Bioinformatics* **36**:897–903

34. Amin M. B., et al. (2017) **The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging** *CA Cancer J Clin* **67**:93–99

35. Tong G. J., et al. (2018) **Comparison of the eighth version of the American Joint Committee on Cancer manual to the seventh version for colorectal cancer: A retrospective review of our data** *World J Clin Oncol* **9**:148–161

36. Nguyen B. Q. T., et al. (2023) **Improvement in neoantigen prediction via integration of RNA sequencing data for variant calling** *Front Immunol* **14**

37. Severine Catreux V. J., Murray Lisa, Mehio Rami, Parnaby Gavin, Roddey Cooper, Ruehle Michael, Chen Wei-Ting, Zhang Fan (2022) **V. J. Severine Catreux, Lisa Murray, Rami Mehio, Gavin Parnaby, Cooper Roddey, Michael Ruehle, Wei-Ting Chen, Fan Zhang. https://www.illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html. 2022**

38. Bolger A. M., Lohse M., Usadel B. (2014) **Trimmomatic: a flexible trimmer for Illumina sequence data** *Bioinformatics* **30**:2114–2120

39. Dobin A., et al. (2013) **STAR: ultrafast universal RNA-seq aligner** *Bioinformatics* **29**:15–21

40. Andrews S. (2010) **FastQC: a quality control tool for high throughput sequence data. (Babraham Bioinformatics, Babraham Institute**

41. Koboldt D. C., et al. (2012) **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing** *Genome Res* **22**:568–576

42. Li H., et al. (2009) **The Sequence Alignment/Map format and SAMtools** *Bioinformatics* **25**:2078–2079

43. Broad Institute (2021) **Broad Institutehttp://broadinstitute.github.io/picard. 2021**

44. McLaren W., et al. (2016) **The Ensembl Variant Effect Predictor** *Genome Biol* **17**

45. Takara Bio USA (2023) **Takara Bio USAhttps://www.takarabio.com/learning-centers/next-generation-sequencing/bioinformatics-resources/cogent-ngs-immune-profiler. 2023**

46. Szolek A., et al. (2014) **OptiType: precision HLA typing from next-generation sequencing data** *Bioinformatics* **30**:3310–3316

47. Li G., et al. (2024) **Splicing neoantigen discovery with SNAF reveals shared targets for cancer immunotherapy** *Sci Transl Med* **16**

48. Hundal J., et al. (2016) **pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens** *Genome Med* **8**

49. Hundal J., et al. (2019) **Accounting for proximal variants improves neoantigen prediction** *Nat Genet* **51**:175–179

50. Lu T., et al. (2021) **Deep learning-based prediction of the T cell receptor-antigen binding specificity** *Nat Mach Intell* **3**:864–875

51. Zhou C., et al. (2019) **pTuneos: prioritizing tumor neoantigens from next-generation sequencing data** *Genome Med* **11**

52. Moodie Z., et al. (2010) **Response definition criteria for ELISPOT assays revisited** *Cancer Immunol Immunother* **59**:1489–1501

53. Chen F., et al. (2019) **Neoantigen identification strategies enable personalized immunotherapy in refractory solid tumors** *J Clin Invest* **129**:2056–2070

54. Hey S., et al. (2023) **Analysis of CDR3 Sequences from T-Cell Receptor beta in Acute Respiratory Distress Syndrome** *Biomolecules* **13**

55. Roudko V., et al. (2020) **Shared Immunogenic Poly-Epitope Frameshift Mutations in Microsatellite Unstable Tumors** *Cell* **183**:1634–1649

56. Maleki Vareki S. (2018) **High and low mutational burden tumors versus immunologically hot and cold tumors and response to immune checkpoint inhibitors** *J Immunother Cancer* **6**

57. Laghi L., et al. (2020) **Prognostic and Predictive Cross-Roads of Microsatellite Instability and Immune Response to Colon Cancer** *Int J Mol Sci* **21**

58. Yassai M. B., Naumov Y. N., Naumova E. N., Gorski J. (2009) **A clonotype nomenclature for T cell receptors** *Immunogenetics* **61**:493–502

59. Vitiello A., Zanetti M. (2017) **Neoantigen prediction and the need for validation** *Nat Biotechnol* **35**:815–817

60. Xia J., et al. (2021) **NEPdb: A Database of T-Cell Experimentally-Validated Neoantigens and Pan-Cancer Predicted Neoepitopes for Cancer Immunotherapy** *Front Immunol* **12**

61. Zhou W. J., et al. (2019) **NeoPeptide: an immunoinformatic database of T-cell-defined neoantigens** *Database (Oxford)* **2019**

62. Tan X., et al. (2020) **dbPepNeo: a manually curated database for human tumor neoantigen peptides** *Database (Oxford)* **2020**

63. Zhang G., Chitkushev L., Olsen L. R., Keskin D. B., Brusic V. (2021) **TANTIGEN 2.0: a knowledge base of tumor T cell antigens and epitopes** *BMC Bioinformatics* **22**

64. Wu J., et al. (2018) **TSNAdb: A Database for Tumor-specific Neoantigens from Immunogenomics Data Analysis** *Genomics Proteomics Bioinformatics* **16**:276–282

65. Montemurro A., et al. (2021) **NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRalpha and beta sequence data** *Commun Biol* **4**

66. Borden E. S., Buetow K. H., Wilson M. A., Hastings K. T. (2022) **Cancer Neoantigens: Challenges and Future Directions for Prediction, Prioritization, and Validation** *Front Oncol* **12**

67. Joshi K., et al. (2019) **Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer** *Nat Med* **25**:1549–1559

68. Motta R., et al. (2021) **Immunotherapy in microsatellite instability metastatic colorectal cancer: Current status and future perspectives** *J Clin Transl Res* **7**:511–522

69. Xie N., et al. (2023) **Neoantigens: promising targets for cancer therapy** *Signal Transduct Target Ther* **8**

70. Muller M., et al. (2023) **Machine learning methods and harmonized datasets improve immunogenic neoantigen prediction** *Immunity* https://doi.org/10.1016/j.immuni.2023.09.002

71. Malekzadeh P., et al. (2020) **Antigen Experienced T Cells from Peripheral Blood Recognize p53 Neoantigens** *Clin Cancer Res* **26**:1267–1276

Editors

Reviewing Editor
Annemarie Woolston
NA, NA, United Kingdom

Senior Editor
Tony Ng
King's College London, London, United Kingdom

Reviewer #2 (Public Review):

Summary:

This paper introduces a novel approach for improving personalized cancer immunotherapy by integrating TCR profiling with traditional pHLA binding predictions, addressing the need for more precise neoantigen CRC patients. By analyzing TCR repertoires from tumor-infiltrating lymphocytes and applying machine learning algorithms, the authors developed a predictive model that outperforms conventional methods in specificity and sensitivity. The validation of the model through ELISpot assays confirmed its potential in identifying more effective neoantigens, highlighting the significance of combining TCR and pHLA data for advancing personalized immunotherapy strategies.

Strengths:

(1) Comprehensive Patient Data Collection: The study meticulously collected and analyzed clinical data from 27 CRC patients, ensuring a robust foundation for research findings. The detailed documentation of patient demographics, cancer stages, and pathology information enhances the study's credibility and potential applicability to broader patient populations.
(2) The use of machine learning classifiers (RF, LR, XGB) and the combination of pHLA and pHLA-TCR binding predictions significantly enhance the model's accuracy in identifying immunogenic neoantigens, as evidenced by the high AUC values and improved sensitivity, NPV, and PPV.
(3) The use of experimental validation through ELISpot assays adds a practical dimension to the study, confirming the computational predictions with actual immune responses. The calculation of ranking coverage scores and the comparative analysis between the combined model and the conventional NetMHCpan method demonstrate the superior performance of the combined approach in accurately ranking immunogenic neoantigens.
(4) The use of experimental validation through ELISpot assays adds a practical dimension to the study, confirming the computational predictions with actual immune responses.

Weakness:

The authors have made comprehensive revisions to the original version of the article, and this version has now addressed my concerns.

https://doi.org/10.7554/eLife.94658.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *This paper reports a number of somewhat disparate findings on a set of colorectal tumour and infiltrating T-cells. The main finding is a combined machine-learning tool which combines two previous state-of-the-art tools, MHC prediction, and T-cell binding prediction to predict immunogenicity. This is then applied to a small set of neoantigens and there is a small-scale validation of the prediciton at the end.*
>
> *Strengths:*
>
> *The prediction of immunogenic neoepitopes is an important and unresolved question.*
>
> *Weaknesses:*
>
> *The paper contains a lot of extraneous material not relevant to the main claim. Conversely, it lacks important detail on the major claim.*
>
> *(1) The analysis of T cell repertoire in Figure 2 seems irrelevant to the rest of the paper. As far as I could ascertain, this data is not used further.*

We appreciate the reviewer for their valuable feedback. We concur with the reviewer's observation that the analysis of the TCR repertoire in Figure 2 should be moved to the supplementary section. We have moved Figures 2B to 2F to Supplementary Figure 2.

However, the analysis of TCR profiles is still presented in Figure 2, as it plays a pivotal role in the process of neoantigen selection. This is because the TCR profiles of eight (out of 28) patients were used for neoantigen prediction. We have added the following sentences to the results section to explain the importance of TCR profiling: "Furthermore, characterizing T cell receptors (TCRs) can complement efforts to predict immunogenicity." (Results, Lines 311-312, Page 11)

> *(2) The key claim of the paper rests on the performance of the ML algorithm combining NETMHC and pmtNET. In turn, this depends on the selection of peptides for training. I am unclear about how the negative peptides were selected. Are they peptides from the same databases as immunogenic petpides but randomised for MHC? It seems as though there will be a lot of overlap between the peptides used for testing the combined algorithm, and the peptides used for training MHCNet and pmtMHC. If this is so, and depending on the choice of negative peptides, it is surely expected that the tools perform better on immunogenic than on non-immunogenic peptides in Figure 3. I don't fully understand panel G, but there seems very little difference between the TCR ranking and the combined. Why does including the TCR ranking have such a deleterious effect on sensitivity?*

We thank the reviewer for their valuable feedback. We believe the reviewer implies 'MHCNet' as NetMHCpan and 'pmtMHC' as pMTnet tools. First, the negative peptides, which have been excluded from PRIME (1), were not randomized with MHC (HLA-I) but were randomized with TCR only. Secondly, the positive peptides selected for our combined algorithms are chosen from many databases such as 10X Genomics, McPAS, VDJdb, IEDB, and TBAdb, while MHCNet uses peptides from the IEDB database and pMTNet uses a totally different dataset from ours for training. Therefore, there is not much overlap between our training data and the training datasets for MHCNet and pMTNet. Thus, the better performance of our tool is not due to overlapping training datasets with these tools or the selection of negative peptides.

To enhance the clarity of the dataset construction, we have added Supplementary Figure 1, which demonstrates the workflow of peptide collection and the random splitting of data to generate the discovery and validation datasets. Additionally, we have revised the following sentence: "To objectively train and evaluate the model, we separated the dataset mentioned above into two subsets: a discovery dataset (70%) and a validation dataset (30%). These subsets are mutually exclusive and do not overlap." (Methods, lines 221-223, page 8).

Initially, the "combine" label in Figure 3G was confusing and potentially misleading when compared to our subsequent approach using a combined machine learning model. In Figure 3G, the "combine" approach simply aggregates the pHLA and pHLA-TCR criteria, whereas our combined machine learning model employs a more sophisticated algorithm to integrate these criteria effectively. The combined analysis in Figure 3G utilizes a basic "AND" algorithm between pHLA and pHLA-TCR criteria, aiming for high sensitivity in HLA binding and high specificity. However, this approach demonstrated lower efficacy in practice, underscoring the necessity for a more refined integration method through machine learning. This was the key point we intended to convey with Figure 3G. To address this issue, we have revised Figure 3G to replace "combined" with "HLA percentile & TCR ranking" to clarify its purpose and minimize confusion.

> *(3) The key validation of the model is Figure 5. In 4 patients, the authors report that 6 out 21 neo-antigen peptides give interferon responses > 2 fold above background. Using NETMHC alone (I presume the tool was used to rank peptides according to binding to the respective HLAs in each individual, but this is not clear), identified 2; using the combined tool identified 4. I don't think this is significant by any measure. I don't understand the score shown in panel E but I don't think it alters the underlying statistic.*

Acknowledging the limitations of our study's sample size, we proceeded to further validate our findings with four additional patients to acquire more data. The final results revealed that our combined model identified seven peptides eliciting interferon responses greater than a two-fold increase, compared to only three peptides identified by NetMHCpan (Figure 5)

> *In conclusion, the paper demonstrates that combining MHCNET and pmtMHC results in a modest increase in the ability to discriminate 'immunogenic' from 'non-immunogenic' peptide; however, the strength of this claim is difficult to evaluate without more knowledge about the negative peptides. The experimental validation of this approach in the context of CRC is not convincing.*
>
> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *This paper introduces a novel approach for improving personalized cancer immunotherapy by integrating TCR profiling with traditional pHLA binding predictions, addressing the need for more precise neoantigen CRC patients. By analyzing TCR*

> *repertoires from tumor-infiltrating lymphocytes and applying machine learning algorithms, the authors developed a predictive model that outperforms conventional methods in specificity and sensitivity. The validation of the model through ELISpot assays confirmed its potential in identifying more effective neoantigens, highlighting the significance of combining TCR and pHLA data for advancing personalized immunotherapy strategies.*
>
> *Strengths:*
>
> *(1) Comprehensive Patient Data Collection: The study meticulously collected and analyzed clinical data from 27 CRC patients, ensuring a robust foundation for research findings. The detailed documentation of patient demographics, cancer stages, and pathology information enhances the study's credibility and potential applicability to broader patient populations.*
>
> *(2) The use of machine learning classifiers (RF, LR, XGB) and the combination of pHLA and pHLA-TCR binding predictions significantly enhance the model's accuracy in identifying immunogenic neoantigens, as evidenced by the high AUC values and improved sensitivity, NPV, and PPV.*
>
> *(3) The use of experimental validation through ELISpot assays adds a practical dimension to the study, confirming the computational predictions with actual immune responses. The calculation of ranking coverage scores and the comparative analysis between the combined model and the conventional NetMHCpan method demonstrate the superior performance of the combined approach in accurately ranking immunogenic neoantigens.*
>
> *(4) The use of experimental validation through ELISpot assays adds a practical dimension to the study, confirming the computational predictions with actual immune responses.*
>
> *Weaknesses:*
>
> *(1) While multiple advanced tools and algorithms are used, the study could benefit from a more detailed explanation of the rationale behind algorithm choice and parameter settings, ensuring reproducibility and transparency.*

We thank the reviewer for their comment. We have revised the explanation regarding the rationale behind algorithm choice and parameter settings as follows: “We examined three machine learning algorithms - Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGB) - for each feature type (pHLA binding, pHLA-TCR binding), as well as for combined features. Feature selection was tested using a k-fold cross-validation approach on the discovery dataset with 'k' set to 10-fold. This process splits the discovery dataset into 10 equal-sized folds, iteratively using 9 folds for training and 1 fold for validation. Model performance was evaluated using the ‘roc_auc’ (Receiver Operating Characteristic Area Under the Curve) metric, which measures the model's ability to distinguish between positive and negative peptides. The average of these scores provides a robust estimate of the model's performance and generalizability. The model with the highest ‘roc_auc’ average score, XGB, was chosen for all features.” (Method, lines 225-234, page 8).

> *(2) While pHLA-TCR binding displayed higher specificity, its lower sensitivity compared to pHLA binding suggests a trade-off between the two measures. Optimizing the balance between sensitivity and specificity could be crucial for the practical application of these predictions in clinical settings.*

We appreciate the reviewer's suggestion. Due to the limited availability of patient blood samples and time constraints for validation, we have chosen to prioritize high specificity and

positive predictive value to enhance the selection of neoantigens.

> *(3) The experimental validation was performed on a limited number of patients (four), which might affect the generalizability of the findings. Increasing the number of patients for validation could provide a more comprehensive assessment of the model's performance.*

This has been addressed earlier. Here, we restate it as follows: Acknowledging the limitations of our study's sample size, we proceeded to further validate our findings with four additional patients to acquire more data. The final results revealed that our combined model identified seven peptides eliciting interferon responses greater than a two-fold increase, compared to only three peptides identified by NetMHCpan (Figure 5).

> ***Reviewer #3 (Public Review):***
>
> *Summary:*
>
> *This study presents a new approach of combining two measurements (pHLA binding and pHLA-TCR binding) in order to refine predictions of which patient mutations are likely presented to and recognized by the immune system. Improving such predictions would play an important role in making personalized anti-cancer vaccinations more effective.*
>
> *Strengths:*
>
> *The study combines data from pre-existing tools pVACseq and pMTNet and applies them to a CRC patient population, which the authors show may improve the chance of identifying immunogenic, cancer-derived neoepitopes. Making the datasets collected publicly available would expand beyond the current datasets that typically describe caucasian patients.*
>
> *Weaknesses:*
>
> *It is unclear whether the pNetMHCpan and pMTNet tools used by the authors are entirely independent, as they appear to have been trained on overlapping datasets, which may explain their similar scores. The pHLA-TCR score seems to be driving the effects, but this not discussed in detail.*

The HLA percentile from NetMHCpan and the TCR ranking from pMTNet are independent. NetMHCpan predicts the interaction between peptides and MHC class I, while pMTNet predicts the TCR binding specificity of class I MHCs and peptides.Additionally, we partitioned the dataset mentioned above into two subsets: a discovery dataset (70%) and a validation dataset (30%), ensuring no overlap between the training and testing datasets.

To enhance the clarity of the dataset construction, we have added Supplementary Figure 1, which demonstrates the workflow of peptide collection and the random splitting of data to generate the discovery and validation datasets. Additionally, we have revised the following sentence: "To objectively train and evaluate the model, we separated the dataset mentioned above into two subsets: a discovery dataset (70%) and a validation dataset (30%). These subsets are mutually exclusive and do not overlap." (Methods, lines 221-223, page 8). We also included the dataset construction workflow in Supplementary Figure 1.

> *Due to sample constraints, the authors were only able to do a limited amount of experimental validation to support their model; this raises questions as to how generalizable the presented results are. It would be desirable to use statistical thresholds to justify cutoffs in ELISPOT data.*

We chose a cutoff of 2 for ELISPOT, following the recommendation of the study by Moodie et al. (2). The study provides standardized cutoffs for defining positive responses in ELISPOT assays. It presents revised criteria based on a comprehensive analysis of data from multiple studies, aiming to improve the precision and consistency of immune response measurements across various applications.

> *Some of the TCR repertoire metrics presented in Figure 2 are incorrectly described as independent variables and do not meaningfully contribute to the paper. The TCR repertoires may have benefitted from deeper sequencing coverage, as many TCRs appear to be supported only by a single read.*

We appreciate the reviewer's feedback. We have moved Figures 2B through 2F to Supplementary Figure 2. We agree with the reviewer that deeper sequencing coverage could potentially benefit the repertoires. However, based on our current sequencing depth, we have observed that many of our samples (14 out of 28) have reached sufficient saturation, as indicated by Figure 2C. The TCR clones selected in our studies are unique molecular identifier (UMI)-collapsed reads, each representing at least three raw reads sharing the same UMI. This approach ensures that the data is robust despite the variability. It is important to note that Tumor-Infiltrating Lymphocytes (TILs) differ across samples, resulting in non-uniform sequencing coverage among them.

> ***Recommendations for the authors**:*
>
> ***Reviewer #2 (Recommendations For The Authors):***
>
> *(1) Please open source the raw and processed data, code, and software output (NetMHCpan, pMTnet), which are important to verify the results.*

NetMHCpan and pMTNet are publicly available software tools (3, 4). In our GitHub repository, we have included links to the GitHub repositories for NetMHCpan and pMTNet (https://github.com/QuynhPham1220/Combined-model).

> *(2) Comparison with more state-of-the-art neoantigen prediction models could provide a more comprehensive view of the combined model's performance relative to the current field.*

To further evaluate our model, we gathered additional public data and assessed its effectiveness in comparison to other models. We utilized immunogenic peptides from databases such as NEPdb (5), NeoPeptide (6), dbPepneo (7), Tantigen (8), and TSNAdb (9), ensuring there was no overlap with the datasets used for training and validation. For non-immunogenic peptides, we used data from 10X Genomics Chromium Single Cell Immune Profiling (10-13).The findings indicate that the combined model from pMTNet and NetMHCpan outperforms NetTCR tool (14). To address the reviewer's inquiry, we have incorporated these results in Supplementary Table 6.

> *(3) While the combined model shows a positive overall rank coverage score, indicating improved ranking accuracy, the scores are relatively low. Further refinement of the model or the inclusion of additional predictive features might enhance the ranking accuracy.*

We appreciate the reviewer's suggestion. The RankCoverageScore provides an objective evaluation of the rank results derived from the final peptide list generated by the two tools. The combined model achieved a higher RankCoverageScore than pMTNet, indicating its superior ability to identify immunogenic peptides compared to existing in silico tools. In

order to provide a more comprehensive assessment, we included an additional four validated samples to recalculate the rank coverage score. The results demonstrate a notable difference between NetMHCpan and the Combined model (-0.37 and 0.04, respectively). We have incorporated these findings into Supplementary Figure 6 to address the reviewer's question. Additionally, we have modified Figure 5E to present a simplified demonstration of the superior performance of the combined model compared to NetMHCpan.

> *(4) Collect more public data and fine-tune the model. Then you will get a SOTA model for neoantigen selection. I strongly recommend you write Python scripts and open source.*

We thank the reviewer for their feedback. We have made the raw and processed data, as well as the model, available on GitHub. Additionally, we have gathered more public data and conducted evaluations to assess its efficiency compared to other methods. You can find the repository here: https://github.com/QuynhPham1220/Combined-model.

> ***Reviewer #3 (Recommendations For The Authors):***
>
> *The Methods section seems good, though HLA calling is more accurate using arcasHLA than OptiType. This would be difficult to correct as OptiType is integrated into pVACtools.*

We chose Optitype for its exceptional accuracy, surpassing 99%, in identifying HLA-I alleles from RNA-Seq data. This decision was informed by a recent extensive benchmarking study that evaluated its performance against "gold-standard" HLA genotyping data, as described in the study by Li et al.(15). Furthermore, we have tested two tools using the same RNA-Seq data from FFPE samples. The allele calling accuracy of Optitype was found to be superior to that of Acras-HLA. To address the reviewer's question, we have included these results in Supplementary Table 2, along with the reference to this decision (Method, line 200, page 07).

> *I am not sufficiently expert in machine learning to assess this part of the methods. TCR beta repertoire analysis of biopsy is highly variable; though my expertise lies largely in sequencing using the 10X genomics platform, typically one sees multiple RNAs per cell. Seeing the majority of TCRs supported by only a single read suggests either problems with RNA capture (particularly in this case where the recovered RNA was split to allow both RNAseq and targeted TCR seq) or that the TCR library was not sequenced deeply enough. I'd like to have seen rarefaction plots of TCR repertoire diversity vs the number of reads to ensure that sufficiently deep sequencing was performed.*

We appreciate the suggestions provided by the reviewer. We agree that deeper sequencing coverage could potentially benefit the repertoires. However, based on our current sequencing depth, we have observed that many of our samples (14 out of 28) have reached sufficient saturation, as indicated by Figure 2C. In addition, the TCR clones selected in our studies are unique molecular identifier (UMI)-collapsed reads, each representing at least three raw reads sharing the same UMI. This approach ensures that the data is robust despite variability. It is important to note that Tumor-Infiltrating Lymphocytes (TILs) differ across samples, resulting in non-uniform sequencing coverage among them. We have already added the rarefaction plots of TCR repertoire diversity versus the number of reads in Figure 2C. These have been added to the main text (lines 329-335).

> *In order to support the authors' conclusions that MSI-H tumors have fewer TCR clonotypes than MSS tumors (Figure S2a) I would have liked to see Figure 2a annotated so that it was easy to distinguish which patient was in which group, as well as the rarefaction plots suggested above, to be sure that the difference represented a real difference between samples and not technical variance (which might occur due to only 4 samples being in the MSI-H group).*

We thank the reviewer for their recommendation. Indeed, it's worth noting that the number of MSI-H tumors is fewer than the MSS groups, which is consistent with the distribution observed in colorectal cancer, typically around 15%. This distribution pattern aligns with findings from several previous studies, as highlighted in these studies (16, 17). To provide further clarification on this point, we have included rarefaction plots illustrating TCR repertoire diversity versus the number of reads in Supplementary Figure 3 (line 339). Additionally, MSI-H and MSS samples have been appropriately labeled for clarity.

> *The authors write: "in accordance with prior investigations, we identified an inverse relationship between TCR clonality and the Shannon index (Supplementary Figure S1)" >> Shannon index is measure of TCR clonality, not an independent variable. The authors may have meant TCR repertoire richness (the absolute number of TCRs), and the Shannon index (a measure of how many unique TCRs are present in the index).*

We thank the reviewer for their comment regarding the correlation between the number of TCRs and the Shannon index. We have revised the figure to illustrate the relationship between the number of TCRs and the Shannon index, and we have relocated it to Figure 2B.

> *The authors continue: "As anticipated, we identified only 58 distinct V (Figure 2C) and 13 distinct J segments (Figure 2D), that collectively generated 184,396 clones across the 27 tumor tissue samples, underscoring the conservation of these segments (Figure 2C & D)" >> it is not clear to me what point the authors are making: it is well known that TCR V and J genes are largely shared between Caucasian populations (https://pubmed.ncbi.nlm.nih.gov/10810226/), and though IMGT lists additional forms of these genes, many are quite rare and are typically not included in the reference sequences used by repertoire analysis software. I would clarify the language in this section to avoid the impression that patient repertoires are only using a restricted set of J genes.*

We thank for the reviewer's feedback. We have revised the sentence as follows: " As anticipated, we identified 59 distinct V segments (Supplementary Figure 2C) and 13 distinct J segments (Supplementary Figure 2D), collectively sharing 185,627 clones across the 28 tumor tissue samples. This underscores the conservation of these segments (Supplementary Figure 2C & D)" (Result, lines 354-356, page 12)

> *As a result I would suggest moving Figure 2 with the exception of 2A into the supplementals - I would have been more interested in a plot showing the distribution of TCRs by frequency, i.e. how what proportion of clones are hyperexpanded, moderately expanded etc. This would be a better measure of the likely immune responses.*

We thank the reviewer for their comment. With the exception of Figure 2A, we have relocated Figures 2B through 2F to Supplementary Figure 2.

> *The authors write "To accomplish this, we gathered HLA and TCRβ sequences from established datasets containing immunogenic and non-immunogenic peptides (Supplementary Table 3)" >> The authors mean to refer to Table S4.*

We appreciate the reviewer's feedback. Here's the revised sentence: "To accomplish this, we gathered HLA and TCRβ sequences from established datasets containing immunogenic and non-immunogenic pHLA-TCR complexes (Supplementary Table 5)" (lines 368-370).

> *The authors write "As anticipated, our analysis revealed a significantly higher prevalence of peptides with robust HLA binding (percentile rank < 2%) among immunogenic peptides in contrast to their non-immunogenic counterparts (Figure 3A & B, p< 0.00001)"*

> *>> this is not surprising, as tools such as NetMHCpan are trained on databases of immunogenic peptides, and thus it is likely that these aren't independent measures (in https://academic.oup.com/nar/article/48/W1/W449/5837056 the authors state that "The training data have been vastly extended by accumulating MHC BA and EL data from the public domain. In particular, EL data were extended to include MA data"). In the pMTNet paper it is stated that pMNet encoded pMHC information using "the exact data that were used to train the netMHCpan model" >> While I am not sufficiently expert to review details on machine learning training models, it would seem that the pHLA scores from NetMHCpan and pMTNet may not be independent, which would explain the concordance in scores that the authors describe in Figures 3B and 3D. I would invite the authors to comment on this.*

The HLA percentiles from NetMHCpan and TCR rankings from pMTNet are independent. NetMHCpan predicts the interaction between peptides and MHC class I, while pMTNet predicts the TCR binding specificity of class I MHCs and peptides. NetMHCpan is trained to predict peptide-MHC class I interactions by integrating binding affinity and MS eluted ligand data, using a second output neuron in the NNAlign approach. This setup produces scores for both binding affinity and ligand elution. In contrast, pMTNet predicts TCR binding specificity of class I pMHCs through three steps:

(1) Training a numeric embedding of pMHCs (class I only) to numerically represent protein sequences of antigens and MHCs.

(2) Training an embedding of TCR sequences using stacked auto-encoders to numerically encode TCR sequence text strings.

(3) Creating a deep neural network combining these two embeddings to integrate knowledge from TCRs, antigenic peptide sequences, and MHC alleles. Fine-tuning is employed to finalize the prediction model for TCR-pMHC pairing.

Therefore, pHLA scores from NetMHCpan and pMTNet are independent. Furthermore, Figures 3B and 3D do not show concordance in scores, as there was no equivalence in the percentage of immunogenic and non-immunogenic peptides in the two groups (≥2 HLA percentile and ≥2 TCR percentile).

> *Many of the authors of this paper were also authors of the epiTCR paper, would this not have been a better choice of tool for assessing pHLA-TCR binding than pMTNet?*

When we started this project, EpiTCR had not been completed. Therefore, we chose pMTNet, which had demonstrated good performance and high accuracy at that time. The validated performance of EpiTCR is an ongoing project that will implement immunogenic assays (ELISpot and single-cell sequencing) to assess the prediction and ranking of neoantigens. This study is also mentioned in the discussion: "Moreover, to improve the accuracy and effectiveness of the machine learning model in predicting and ranking neoantigens, we have developed an in-house tool called EpiTCR. This tool will utilize immunogenic assays, such as ELISpot and single-cell sequencing, for validation." (lines 532-535).

> *In Figure 3G it would appear that the pHLA-TCR score is driving the interaction, could the authors comment on this?*

The authors sincerely appreciate the reviewer for their valuable feedback. Initially, the "combine" label in Figure 3G was confusing and potentially misleading when compared to our subsequent approach using a combined machine learning model. In Figure 3G, the "combine" approach simply aggregates the pHLA and pHLA-TCR criteria, whereas our

combined machine learning model employs a more sophisticated algorithm to integrate these criteria effectively.

The combined analysis in Figure 3G utilizes a basic "AND" algorithm between pHLA and pHLA-TCR criteria, aiming for high sensitivity in HLA binding and high specificity. However, this approach demonstrated lower efficacy in practice, underscoring the necessity for a more refined integration method through machine learning. This was the key point we intended to convey with Figure 3G. To address this issue, we have revised Figure 3G to replace "combined" with "HLA percentile & TCR ranking" to clarify its purpose and minimize confusion.

> *In Figure 4A I would invite the authors to comment on how they chose the sample sizes they did for the discovery and validation datasets: the numbers seem rather random. I would question whether a training dataset in which 20% of the peptides are immunogenic accurately represents the case in patients, where I believe immunogenic peptides are less frequent (as in Figure 5).*

We aimed to maximize the number of experimentally validated immunogenic peptides, including those from viruses, with only a small percentage from tumors available for training. This limitation is inherent in the field. However, our ultimate objective is to develop a tool capable of accurately predicting peptide immunogenicity irrespective of their source. Therefore, the current percentage of immunogenic peptides may not accurately reflect real-world patient cases, but this is not crucial to our development goals.

> *For Figure 5C I would invite the authors to consider adding a statistical test to justify the cutoff at 2fold enrichments.*

Thank you for your feedback. Instead of conducting a statistical test, we have implemented standardized cutoffs as defined in the cited study (2). This research introduces refined criteria for identifying positive responses in ELISPOT assays through a comprehensive analysis of data from multiple studies. These criteria aim to improve the accuracy and consistency of immune response measurements across various applications. The reference to this study has been properly incorporated into the manuscript (Method, line 281, page 10).

> *Minor points:*
>
> *"paired white blood cells" >> use "paired Peripheral Blood Mononuclear Cells".*

We appreciate the reviewer for the feedback. We agree with the reviewer's observation. The sentence has been revised as follows: "Initially, DNA sequencing of tumor tissues and paired Peripheral Blood Mononuclear Cells identifies cancer-associated genomic mutations. RNA sequencing then determines the patient's HLA-I allele profile and the gene expression levels of mutated genes." (Introduction, lines 55-58, page 2).

> *"while RNA sequencing determines the patient's HLA-I allele profile and gene expression levels of mutated genes." >> RNA sequencing covers both the mutant and reference form of the gene, allowing assessment of variant allele frequency.*
>
> *"the current approach's impact on patient outcomes remains limited due to the scarcity of effective immunogenic neoantigens identified for each patient" >> Some clearer language here would have been preferred as different tumor types have different mutational loads*

We thank the reviewer for their valuable feedback. We agree with the reviewer's observation. The passage has been revised accordingly: “The current approach's impact on

patient outcomes remains limited due to the scarcity of mutations in cancer patients that lead to effective immunogenic neoantigens." (Introduction, lines 62-64, page 3).

References

(1) J. Schmidt *et al.*, Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep Med* **2**, 100194 (2021).

(2) Z. Moodie *et al.*, Response definition criteria for ELISPOT assays revisited. *Cancer Immunol Immunother* **59**, 1489-1501 (2010).

(3) V. Jurtz *et al.*, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* **199**, 3360-3368 (2017).

(4) T. Lu *et al.*, Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nat Mach Intell* **3**, 864-875 (2021).

(5) J. Xia *et al.*, NEPdb: A Database of T-Cell Experimentally-Validated Neoantigens and Pan-Cancer Predicted Neoepitopes for Cancer Immunotherapy. *Front Immunol* **12**, 644637 (2021).

(6) W. J. Zhou *et al.*, NeoPeptide: an immunoinformatic database of T-cell-defined neoantigens. *Database (Oxford)* **2019** (2019).

(7) X. Tan *et al.*, dbPepNeo: a manually curated database for human tumor neoantigen peptides. *Database (Oxford)* **2020** (2020).

(8) G. Zhang, L. Chitkushev, L. R. Olsen, D. B. Keskin, V. Brusic, TANTIGEN 2.0: a knowledge base of tumor T cell antigens and epitopes. *BMC Bioinformatics* **22**, 40 (2021).

(9) J. Wu *et al.*, TSNAdb: A Database for Tumor-specific Neoantigens from Immunogenomics Data Analysis. *Genomics Proteomics Bioinformatics* **16**, 276-282 (2018).

(10) https://www.10xgenomics.com/resources/datasets/cd-8-plus-t-cells-of-healthy-donor-1-1-standard-3-0-2.

(11) https://www.10xgenomics.com/resources/datasets/cd-8-plus-t-cells-of-healthy-donor-2-1-standard-3-0-2.

(12) https://www.10xgenomics.com/resources/datasets/cd-8-plus-t-cells-of-healthy-donor-3-1-standard-3-0-2.

(13) https://www.10xgenomics.com/resources/datasets/cd-8-plus-t-cells-of-healthy-donor-4-1-standard-3-0-2.

(14) A. Montemurro *et al.*, NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRalpha and beta sequence data. *Commun Biol* **4**, 1060 (2021).

(15) G. Li *et al.*, Splicing neoantigen discovery with SNAF reveals shared targets for cancer immunotherapy. *Sci Transl Med* **16**, eade2886 (2024).

(16) Z. Gatalica, S. Vranic, J. Xiu, J. Swensen, S. Reddy, High microsatellite instability (MSI-H) colorectal carcinoma: a brief review of predictive biomarkers in the era of personalized medicine. *Fam Cancer* **15**, 405-412 (2016).

(17) N. Mulet-Margalef *et al.*, Challenges and Therapeutic Opportunities in the dMMR/MSI-H Colorectal Cancer Landscape. *Cancers (Basel)* **15** (2023).

https://doi.org/10.7554/eLife.94658.2.sa0