frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 04 September 2023
DOI 10.3389/fimmu.2023.1251603

OPEN ACCESS

EDITED BY
Patrick Schmidt,
National Center for Tumor Diseases (NCT), Germany

REVIEWED BY
Ming Yi,
Zhejiang University, China
Ngoc Hieu Tran,
University of Waterloo, Canada

*CORRESPONDENCE
Le Son Tran
leson1808@gmail.com
Hoai-Nghia Nguyen
nhnghia81@gmail.com

†These authors have contributed equally to this work

RECEIVED 02 July 2023
ACCEPTED 17 August 2023
PUBLISHED 04 September 2023

CITATION
Nguyen BQT, Tran TPD, Nguyen HT, Nguyen TN, Pham TMQ, Nguyen HTP, Tran DH, Nguyen V, Tran TS, Pham T-VN, Le M-T, Phan M-D, Giang H, Nguyen H-N and Tran LS (2023) Improvement in neoantigen prediction via integration of RNA sequencing data for variant calling. *Front. Immunol.* 14:1251603. doi: 10.3389/fimmu.2023.1251603

Improvement in neoantigen prediction via integration of RNA sequencing data for variant calling

Bui Que Tran Nguyen[1†], Thi Phuong Diem Tran[1†], Huu Thinh Nguyen[2], Thanh Nhan Nguyen[1], Thi Mong Quynh Pham[1], Hoang Thien Phuc Nguyen[1], Duc Huy Tran[2], Vy Nguyen[1], Thanh Sang Tran[2], Truong-Vinh Ngoc Pham[2], Minh-Triet Le[2], Minh-Duy Phan[1], Hoa Giang[1], Hoai-Nghia Nguyen[1*] and Le Son Tran[1*]

[1]Medical Genetics Institute, Ho Chi Minh, Vietnam, [2]University Medical Center Ho Chi Minh City, Ho Chi Minh, Vietnam

Introduction: Neoantigen-based immunotherapy has emerged as a promising strategy for improving the life expectancy of cancer patients. This therapeutic approach heavily relies on accurate identification of cancer mutations using DNA sequencing (DNAseq) data. However, current workflows tend to provide a large number of neoantigen candidates, of which only a limited number elicit efficient and immunogenic T-cell responses suitable for downstream clinical evaluation. To overcome this limitation and increase the number of high-quality immunogenic neoantigens, we propose integrating RNA sequencing (RNAseq) data into the mutation identification step in the neoantigen prediction workflow.

Methods: In this study, we characterize the mutation profiles identified from DNAseq and/or RNAseq data in tumor tissues of 25 patients with colorectal cancer (CRC). Immunogenicity was then validated by ELISpot assay using long synthesis peptides (sLP).

Results: We detected only 22.4% of variants shared between the two methods. In contrast, RNAseq-derived variants displayed unique features of affinity and immunogenicity. We further established that neoantigen candidates identified by RNAseq data significantly increased the number of highly immunogenic neoantigens (confirmed by ELISpot) that would otherwise be overlooked if relying solely on DNAseq data.

Discussion: This integrative approach holds great potential for improving the selection of neoantigens for personalized cancer immunotherapy, ultimately leading to enhanced treatment outcomes and improved survival rates for cancer patients.

KEYWORDS

neoantigen, colorectal cancer (CRC), RNA sequencing (RNAseq), tumor variant calling, neoantigen identification workflow, Neoantigen prioritization, cancer immunotherapy

Introduction

Colorectal cancer (CRC) is a major global health concern, being the third most common cancer in the world and the fifth leading cause of cancer-related mortality among the Vietnamese population (1, 2). Traditional treatments, such as surgery, chemotherapy, and radiation therapy, have limited efficacy and are poorly tolerant, particularly in advanced stages of CRC (3). Immunotherapy, while not a cure for CRC, has the potential to significantly improve patient survival rates and quality of life (4, 5). In metastatic CRC patients, immunotherapy has demonstrated promise in improving outcomes. 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 (6–8). However, alternative immunotherapy strategies are urgently required for CRC patients, as patients with proficient mismatch repair (pMMR) or microsatellite stability (MSS) have not shown significant responses to immune checkpoint inhibitors (6, 9).

Neoantigens (neopeptides) have emerged as potential targets for personalized cancer immunotherapy, including CRC (10–12). Neoantigens are peptides resulting from somatic mutations, capable of being presented by class I human leukocyte antigen (HLA-I) molecules on cancer cell surface and by class II HLA molecules on professional antigen-presenting cells, thereby activating anti-tumor immune responses (13). Recent studies have demonstrated that the presence of neoantigens is associated with better responses to immune checkpoint inhibitor (ICI) therapy in CRC patients (14, 15). A high neoantigen burden has been linked to improved overall survival and progression-free survival in patients with various solid tumors, including CRC (14, 15). Therefore, neoantigen-based immunotherapies are thought to have the potential to significantly improve treatment outcomes for CRC patients.

The identification of neoantigens with strong binding affinity to their respective HLA-I molecules and high immunogenicity is critical for the development of effective neoantigen-based therapies. This process involves the use of next-generation sequencing (NGS) and bioinformatics tools. Initially, DNA sequencing of tumor tissues and paired white blood cells enables the identification of cancer associated genomic mutations, while RNA sequencing is used to determine patient's HLA-I allele profile and to quantify expression levels of genes carrying mutations. Next, tumor somatic variant, HLA-I allele, and gene expression data are analyzed using *in silico* tools based on machine learning algorithms to predict the binding affinity of neoantigens to patients' HLA-I alleles and their potential to activate T cell responses (16–18). This standard workflow has been exploited in numerous studies to identify clinically relevant neoantigens in melanoma, lung cancer, and other malignancies (17, 19).

Despite promising results, only small portions of patients benefit from the current approach due to the limited number of effective immunogenic neoantigens identified for each patient. To maximize the detection of potential neoantigens, whole exosome sequencing (WES) has been employed to comprehensively profile the cancer-specific landscape (20–22). While WES allows a much larger search space for mutations within the genome, it is not a cost- and time-effective approach. Moreover, a significant proportion of identified tumor DNA mutations, especially those which are not actively transcribed or transcribed at very low levels, might not result in the formation of neoantigens (19). Lastly, WES-based mutation calling is inefficient in capturing all tumor somatic mutations, especially clonal mutations with low frequencies and underrepresentation in the sequencing data (23), while targeting combined neoantigens derived from both clonal and subclonal mutations is necessary to evoke efficient immune-mediated cell death in a broader range of tumor cells. Therefore, relying solely on DNAseq data for tumor mutation calling, which has traditionally been the basis for identifying neoantigens, may not capture the full extent of tumor-related mutations, resulting in an incomplete identification of neoantigens.

Genetic variants at the RNA level are frequently excluded from conventional bioinformatic workflows, despite several studies indicating that neoantigens can be derived from RNA mutations, such as splicing, polyadenylation dysregulation, or RNA editing (24, 25). In addition, recent studies have shown that the presence of variant-bearing transcripts is an important factor for accurate identification of immunogenic neoantigen candidates (26, 27). Therefore, integrating RNAseq data into tumor mutation calling holds promise for unveiling a more comprehensive repertoire of neoantigens and, consequently, advancing the development of personalized immunotherapies for cancer. However, the feasibility and effectiveness of this approach require further examination.

To assess the utility of RNAseq analysis for neoantigen identification, we compared the cancer mutation profiles, binding affinity to HLA-I of neoantigens identified from RNAseq and DNAseq, and their predicted immunogenicity across 25 CRC patients. Moreover, we performed experimental validation to assess the effectiveness of utilizing RNAseq for the identification of immunogenic neoantigens. This validation utilized the ELISpot assay to measure the ability of neoantigen candidates, predicted from DNAseq and RNAseq-derived variants, to activate T cells in PBMCs obtained from four CRC patients.

Abbreviations: CRC, colorectal cancer; dMMR, deficient mismatch repair; DNAseq, DNA sequencing; FDA, the US Food and Drug Administration; FPKM, Fragments Per Kilobase of transcript per Million mapped reads; GATK, Genome Analysis Toolkit; HLA, human leukocyte antigens; ICI, immune checkpoint inhibitor; IFN-γ, interferon-gamma; LPs, long peptides; MAF, mutant-allele fraction; MSI-H, high microsatellite instability; MSS, microsatellite stability; NGS, next-generation sequencing;PBMCs, peripheral blood mononuclear cells; pMMR, proficient mismatch repair; RNAseq, RNA sequencing; SNPs, single nucleotide polymorphisms; TCR, T cell receptor; VAF, variant allele frequency; WES, whole exosome sequencing.

Materials and methods

Tumor biopsy and peripheral blood collection

A total of 25 patients diagnosed with colorectal cancer (CRC) were enrolled in this study from the University Medical Center at Ho Chi

Minh city between June 2022 and April 2023. The confirmation of CRC was based on abnormal colonoscopies and histopathological analysis confirming the presence of malignancy. The stages of CRC were determined following the guidelines provided by the American Joint Committee on Cancer and the International Union for Cancer Control. Prior to participation, all patients provided written informed consent for the collection of tumor and whole blood samples. 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 Table S1. The Ethics Committee of The University of Medicine and Pharmacy at Ho Chi Minh City, Vietnam approved this study. Fresh CRC specimens were collected immediately after biopsy or tumor resection and were placed in microtubes containing RNAlater, an RNA stabilization solution (Thermo Fisher Scientific, Japan). For four patients, ten mL of peripheral blood was collected serially before surgery and stored in Heparin tubes.

Targeted DNA and RNA 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 2x100 bp with an average target coverage of 200X (with actual coverage from 89 to 968X).

Isolated total RNA was subjected to a NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, MA, USA) to isolate intact poly(A)+ RNA as per manufacturer instructions. RNA libraries were constructed using NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs). These libraries were subsequently sequenced for paired end reads of 2x100 bp on an MGI system at 50X depth coverage.

Variant calling from DNAseq and RNAseq data

To select the optimal variant calling tool for DNAseq data, we evaluated the performance of three different pipelines including Dragen, VarScan and MuTect2, which are commonly used for somatic variant calling (28, 29). Among the three pipelines, Dragen demonstrated superior performance for detecting a set of validated ground truth variants in a standard dataset downloaded from a public repository, NCBI Sequencing Read Archive SRA (ID: SRR7890830) (Figure S1A). Therefore, we utilized Dragen (Illumina) (30) in tumor-normal mode to call somatic mutations from DNAseq data. The default filtering thresholds of Dragen were used to call SNPs and indels. SNPs were further filtered using the dbSNP and 1000 Genome datasets. Germline mutations in tumor tissues were identified by comparing them with matched WBC-DNA samples. Mutations within immunoglobulin and HLA genes were excluded due to alignment difficulties in these highly polymorphic regions that require specialized analysis tools (31). Additionally, synonymous mutations were removed from downstream analysis. Included for analysis were somatic mutations that surpassed a minimum threshold of ≥2% variant allele frequency (VAF) in DNA extracted from fresh frozen tissues.

To identify the most suitable variant calling tools for RNAseq data, we assessed the performance of two different pipelines, VarScan and MuTect2 by comparing the proportions of variants that overlapped with DNA-derived variants. Sequencing reads were trimmed using Trimmomatic (32) and aligned to the human reference genome using STAR (version 2.6.0c) (33). Prior to alignment, the raw sequencing reads underwent quality checks using FastQC version 0.11.9 (34). VarScan 2 (28), which accepts both DNA and RNAseq data, was used to call mutations in paired tumor and WBC samples in 95 cancer-associated genes, again in the tumor-normal mode. Four filtering steps were applied: (i) only calls with a PASS status were used, (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% 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 and indexed using Samtools version 1.10 (35), and PCR duplicates were eliminated using Picard tools version 2.25.6 (36). The mutations from RNAseq data were also called using MuTect2, a variant caller from the Genome Analysis Toolkit (GATK) pipeline. Like VarScan, the MuTect2 pipeline was run in tumor versus normal mode, utilizing default settings. Following variant calling, a similar variant filtration step was also applied to eliminate potential false positives. Somatic variants from the two pipelines were manually checked using Integrative Genomics Viewer (v2.8.2). The VCF files generated by Dragen (for DNAseq) and by MuTect2 and VarScan (for RNAseq) were subsequently annotated using the Ensembl Variant Effect Predictor (VEP version 105) (37) to extract the potential effect of variants on the phenotypic outcome.

Gene expression quantification and tumor purity estimation

We used the Cufflinks (38) to analyze the tumor RNAseq data using the Ensemble human reference transcriptomes (GRCh38) for assessing gene expression. The expression data was used to calculate the tumor purity via ESTIMATE (v1.0.13) package, (R-v3.6.3) (39).

In silico prediction of HLA binding affinity and immunogenicity

Class I HLA alleles (HLA-A/B/C) with two-digit resolution were identified from patient tumor RNAseq data using OptiType tool (40). The annotated VCF files were analyzed using pVAC-Seq, a tool of pVACtools (v1.5.9) (16, 41, 42) with the default settings, except for disabling the coverage and MAF filters. We used all HLA-I binding algorithms that were implemented in pVAC-Seq to predict 8 to 11-mer epitopes binding to HLA-I (A, B, or C) for downstream analysis. Neoantigen candidates were subjected to MHC binding predictions and subsequent prioritization based on their binding affinity scores (measured in nM) using NetMHCpan-4.1 (18). The prioritization process involved calculating the percentile ranking of each neoantigen's binding affinity score within the distribution of scores for the corresponding HLA allele. Neoantigen candidates with a percentile rank lower than 2% were selected for our immunogenicity analysis.

The immunogenicity of neoantigens was validated by the PRIME tool (43) with default settings. To predict the immunogenicity of neoantigen candidates, a two-step ranking process was employed, involving ranking the neoantigen candidates based on their immunogenicity score and estimating percentiles for each HLA allele. These scores represented the predicted likelihood of a neoantigen being immunogenic. The neoantigens were then ranked in descending order based on their immunogenicity scores, enabling the prioritization of neoantigen candidates with higher predicted immunogenicity for further analysis. A ranking value for immunogenicity was assigned to each neoantigen candidate by determining the percentile rank of its immunogenicity score within the group of neoantigens predicted to bind to the same HLA allele. The percentile rank of binding affinity score in NetMHCpan or immunogenicity score in PRIME for a peptide is the fraction of random peptides that would have a score higher or equal to the peptide given in input. Therefore, a peptide with lower percentile rank value of NetMHCpan or PRIME indicate better binding affinity and immunogenicity, respectively. To identify public neoantigens, we conducted a comprehensive search of several databases, including TSNAdb (44, 45), NeoPeptide (46), dbPepNeo (47, 48), NEPdb (49), TANTIGEN (50, 51), and IEDB (52). All databases contained epitopes from published studies where their immunogenicity was validated by immunological assays.

Isolation, culture, and stimulation of PBMCs with long peptides

Peripheral blood samples from four patients were collected 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. PBMCs were then resuspended in FBS/10% DMSO solution with a concentration of 7-10x10^6 cells/mL for freezing 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 µg/mL) solution. 10^5 PBMCs were allowed to rest in 96-round bottom well-plate containing AIM V media supplemented with 10% FBS, 10 mM HEPES, and 50 µM β-mercaptoethanol overnight before stimulation with synthesized long peptides at a concentration of 5 µM in a humidified incubator at 37°C with 5% CO_2. 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, Sigma-Aldrich, 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 a 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 ELISpot plate (Mabtech, Sweden) and incubated for 20 hours at 37°C. PBMCs cultured with DMSO were used as a negative control group, while PBMCs stimulated with anti-CD3 were used as a positive control group. ELISpot assay was performed on treated PBMCs using ELISpot Pro: Human IFN-γ (ALP) kit (Mabtech, Sweden), following manufacture's protocol. Developed spots on the ELISpot plate were then enumerated using an ELISpot reader (Mabtech, Sweden). The reactivity was determined by measuring the fold increase in the number of spots of PBMCs treated with mutant peptides relative to those treated with wild type peptides. A fold change of two was selected as the cut off for positivity (53).

Flow cytometry intracellular staining for IFN-γ

Cells from ELISpot plate were collected in media supplemented with GolgiStop Protein Transport Inhibitor (BD Biosciences, NJ, USA) and incubated for 6 hr at 37°C. Positive control group was treated with 50 µM PMA (Abcam, UK), 1 mg/mL Ionomycin (Abcam, UK). Cells were then washed, blocked with Fc receptor (Biolegend, CA, USA), and stained with CD3-PE (clone HIT3a, Biolegend), CD4-PE/Cyanine7 (clone RPA-T4, Biolegend), CD8-FITC (clone RPA-T8, Cell Signaling) antibodies for 2 hr at 4°C. Cells were permeabilized for 20 mins at 4°C and then stained overnight with IFN-γ-APC (clone 4S.B3, Biolegend) antibody at 4°C.

Statistical analysis

The Wilcoxon rank-sum test was used to compare the coverage, VAF, and immunogenicity percentile among three groups for three

mutation groups (DNA-unique, RNA-unique and Shared). All statistical analyses were carried out using R (v2.6.3).

Results

Comparison of mutation profiles from DNA sequencing and RNA sequencing data

RNA sequencing (RNAseq) data, which is commonly used for analysis of mutated gene expression in the current standard workflow of neoantigen identification, have been exploited to identify cancer-specific mutations in recent studies (27, 54, 55). However, the properties of RNAseq derived variants and neoantigens have not been fully characterized. To assess the utility of RNAseq in calling cancer-specific somatic mutations for neoantigen prediction, we sought to compare the mutation profiles obtained from RNAseq and DNAseq data across 25 CRC patients (Table S1), with a focus on all single nucleotide variants (SNVs) and indel variants (Figure 1). To achieve a balance between cost and mutation detection efficiency, we used a targeted sequencing panel consisting of 95 commonly mutated cancer-associated genes (Table S2). As a result, our comparison of RNAseq and DNAseq analysis was limited to these genes (Figure 1). The DNAseq and RNAseq data obtained from all 25 CRC patients have successfully met quality metrics, ensuring reliable datasets for mutation calling (Tables S3, S4). To identify mutations in DNAseq data, we used Dragen as our primary tool due to its superior performance in both SNV and indel mutation calling from a reference sample compared to other tools used in the analysis of DNAseq data (Figure S1A) (56).

To determine the most effective tool for calling mutations from RNAseq data, we compared the performance of VarScan and MuTect2. We found that VarScan yielded a higher proportion of variants overlapping with mutations detected from DNAseq compared to MuTect2 (18.3% versus 0.8%, Figure S1B). Furthermore, while MuTect2 tended to call a high percentage of indels with abnormal length, VarScan yielded a higher proportion of SNVs that were comparable to the mutation profiles identified from DNAseq (Figures S1C, D). These data suggested that VarScan exhibited higher sensitivity in detecting SNVs and produced fewer artifact indels. Thus, we decided to use VarScan as the variant calling tool for RNAseq data from the 25 CRC patients.

Out of the total 1,520 variants identified, only 340 (22.4%) were common between the two mutation calling methods, while most variants (77.6%) were exclusively detected by either DNAseq (DNA-unique) or RNAseq (RNA-unique) data. DNA-unique variants were more frequent than RNA-unique variants (56.1% versus 21.5.%, Figure 2A). Shared variants were detected in 16 out of the 25 CRC patients, accounting for 1% to 47% of the total identified variants (Figure 2B, Table S5). Interestingly, we found that RNA-unique variants were the major source of variants in 4 out of 25 (16%) patients (Figure 2B), while DNA-unique variants were identified as the major source of variants in the remaining 21 patients.

When comparing the distribution of variant types between DNAseq and RNAseq, we observed a consistent pattern where missense variants were the most prevalent variant type (>50% of all

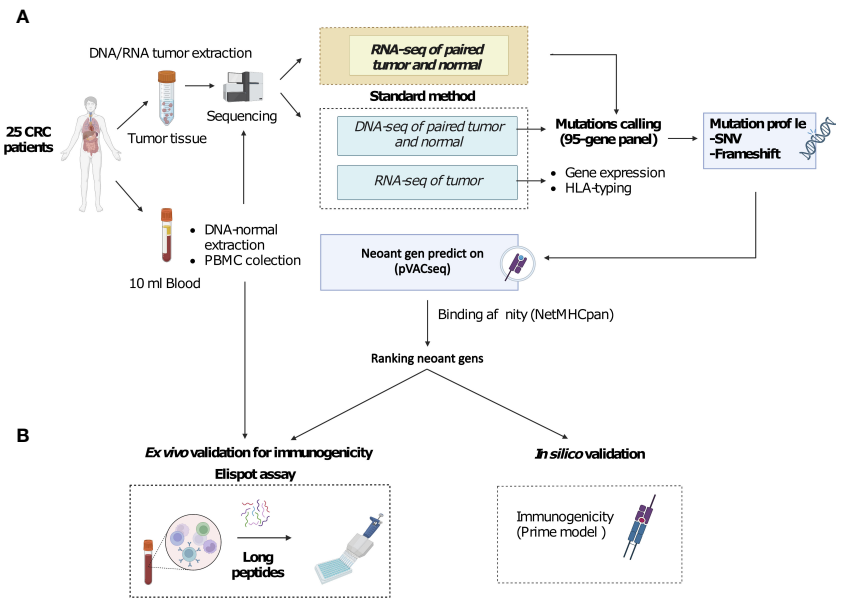

FIGURE 1
A novel workflow for CRC neoantigen identification and validation that integrates RNAseq data into somatic mutation calling. **(A)**. Schematic diagram of the new workflow. Tumor biopsies and blood samples from CRC patients are subjected to targeted DNA and RNA sequencing, which focuses on a panel of 95 genes, for somatic mutation calling. Additionally, RNAseq data is used to determine gene expression and HLA-typing information. pVAC-Seq tool is then utilized for neoantigen prediction using DNA and RNA-derived somatic mutation data, gene expression data, and patient-specific HLA-typing data as inputs. **(B)**. Methods to validate the advantages of the workflow. Predicted neoantigens from the workflow are subsequently validated by *ex vivo* ELISpot assay measuring IFN-γ secretion from PBMCs stimulated with long peptides carrying predicted variants and by *in silico* prediction of immunogenicity by PRIME tool.

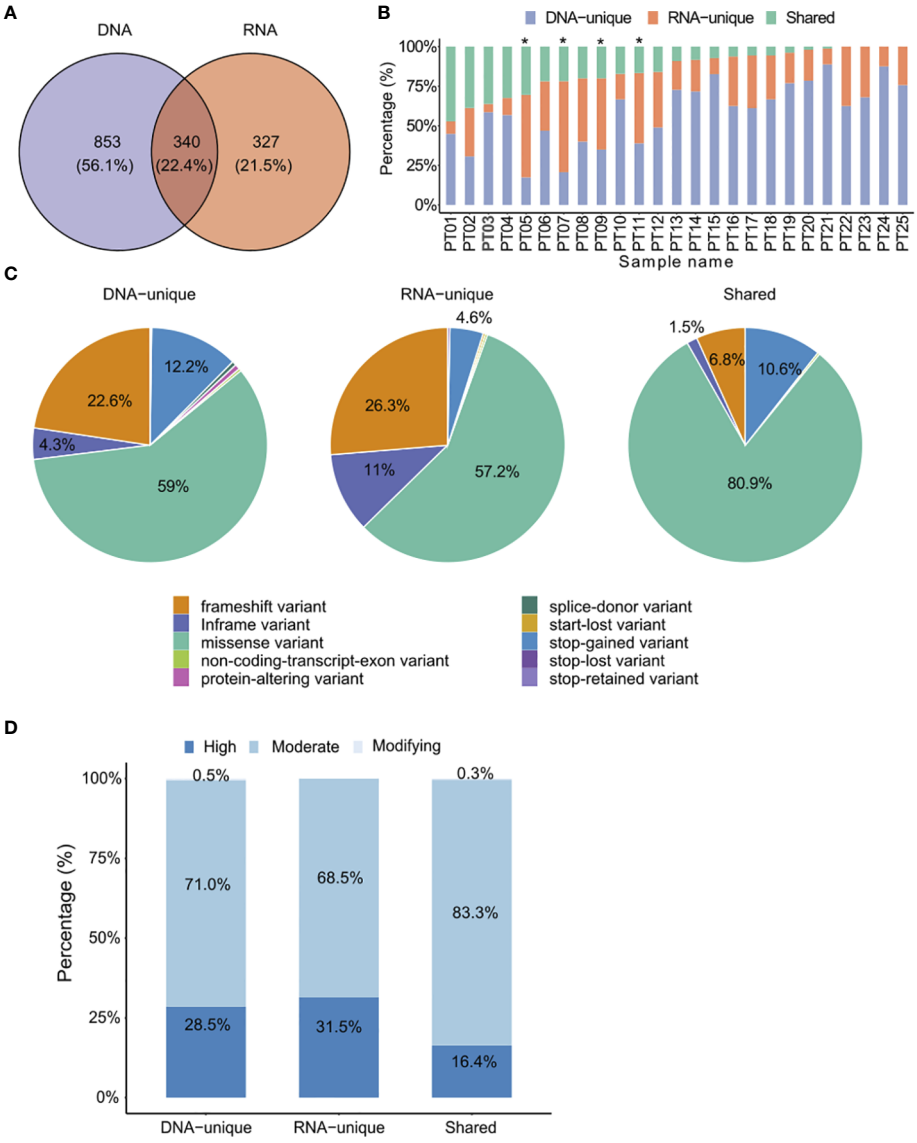

FIGURE 2
Comparison of identified somatic mutations between DNAseq data and RNAseq data. **(A)** Venn diagrams display the numbers of DNA and RNA mutations called by the specified mutation callers on matched tumor-normal DNAseq and RNAseq data from 25 CRC patients. **(B)** Proportions of each type of variants identified from both DNAseq and RNAseq data for each patient. The graph is presented in descending order based on the proportion of shared variants. Patients marked with an asterisk exhibited a higher proportion of RNA-unique variants compared to DNA-unique variants. **(C)** Pie charts presenting the percentages of mutation types. **(D)** The proportions of indicated types of variants in relation to their phenotypic impacts.

variants in each group, Figure 2C). However, we did notice some notable differences. Specifically, RNA-unique variants exhibited a higher frequency of in-frame variants (11% compared to 4.3% in DNAseq, Figure 2C) and frameshift variants (26.3% versus 22.6%, Figure 2C). On the other hand, DNA-unique variants had a higher occurrence of stop-gained variants (12.2% versus 4.6%, Figure 2C). In the shared-variant group, most variants consisted of missense variants (80.9%) and stop-gained variants (10.6%), collectively accounting for approximately 91.5% of all variants. To predict the functional impact of the three variant groups, we employed the Ensembl's Variant Effect Predictor tool (37). Our analysis revealed that the phenotypic outcome was most significantly affected by RNA-unique variants in the high impact category, followed by DNA-unique and shared variants (Figure 2D). These results indicate a clear distinction between the tumor variant landscapes profiled by RNAseq and DNAseq, wherein RNAseq reveals a greater proportion of clinically relevant variants compared to DNAseq. Therefore, RNAseq appears to be particularly valuable in identifying variants with potential clinical significance.

To gain deeper insights into the variants identified by both sequencing methods, we conducted an analysis of their depth coverage and mutation allele frequency (MAF). Despite having lower coverage levels ($P= 9.1x10^{-5}$, Figure 3A), the shared variants exhibited significantly higher MAFs ($P= 2.22x10^{-16}$, Figure 3B) compared to the DNA-unique. This observation suggests that the shared variants are likely derived from major clones of somatic mutation clones, while the

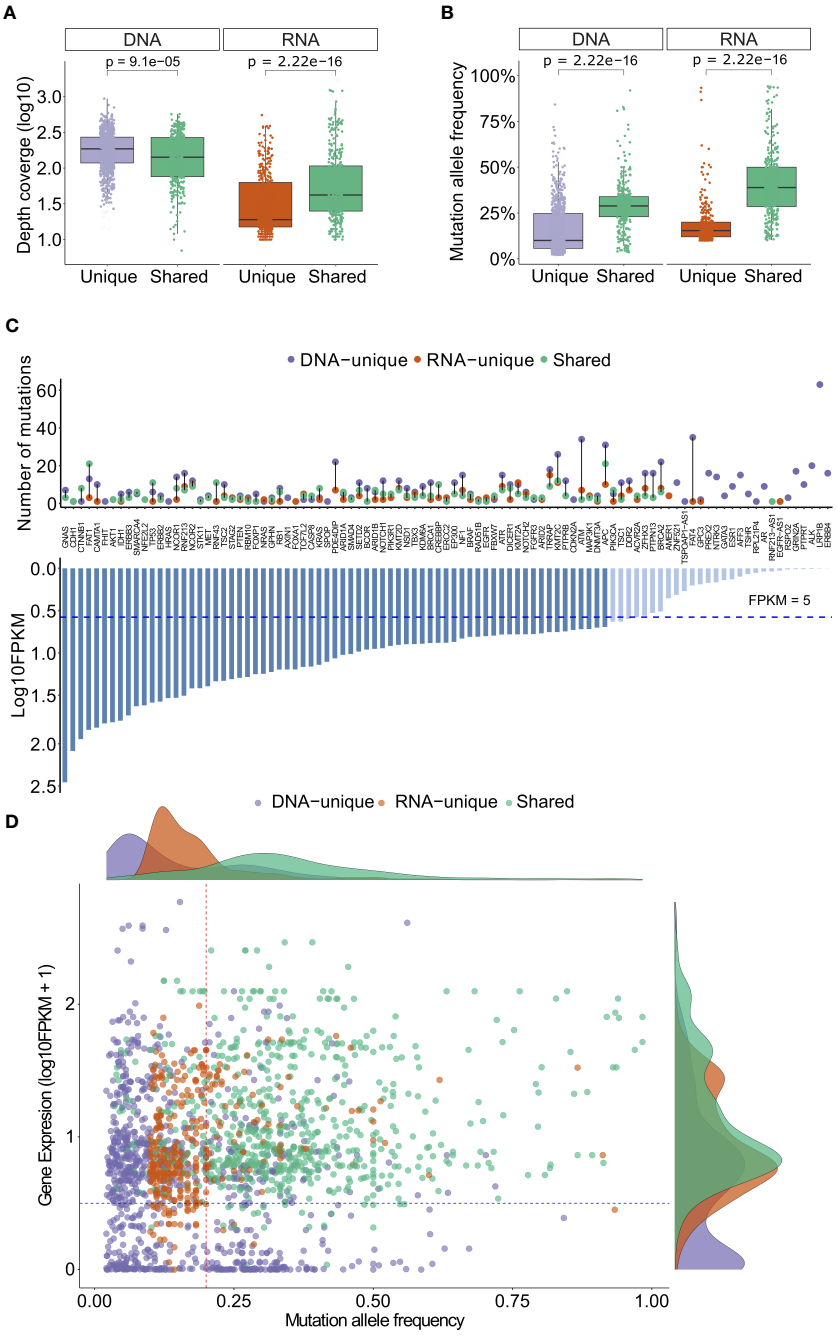

FIGURE 3
Depth coverage, MAFs and gene expression levels of variants from DNAseq and RNAseq data. **(A)** Depth coverage of the indicated groups of variants based on DNAseq and RNAseq data. **(B)** Mutation allele frequency of the indicated groups of variants. **(C)** A list of genes with indicated variants, along with their corresponding FPKM. **(D)** Gene expression levels of different groups of variants in relation to their mutation allele frequency. In **(A, B)**, the boxes represent the median value, as well as the lower and upper quartiles (25th and 75th percentiles). The p-values were obtained from the Wilcoxon rank-sum test.

DNA-unique variants, characterized by significantly lower MAF ($P < 2.22x10^{-16}$, Figure 3B), may originate from minor tumor clones.

RNA-unique variants displayed a notably lower median depth of coverage ($P<2.22x10^{-}16$, Figure 3A) and MAF (20% versus 40%, $p<2.22x10^{-6}$, Figure 3B) compared to the shared variants. These findings suggest that RNA-unique variants may originate from genes with low expression levels, resulting in a smaller number of variant transcripts. It is notable that the majority of shared variants and RNA-unique variants were identified in genes with high expression levels (FPKM >5, dashed line, Figure 3C), while unique variants identified through DNAseq (494/853, 58%, Table S5) were more commonly found in genes with low expression levels (FPKM <5, Figure 3C). Furthermore, when examining the MAF of variants in relation to their gene expression levels, shared variants (green dots, Figure 3D) exhibited higher levels of gene expression (FPKM >5) and MAF (> 24%) compared to other mutation types. In contrast, RNA-unique variants

(orange dots, Figure 3D) tended to have similar gene expression levels but lower MAF, while a substantial number of DNA-unique variants (purple dots, Figure 3D) displayed both low gene expression and MAF. These observations strongly suggest that the MAF and transcriptional activity of mutated genes are significant factors contributing to the disparities observed between RNAseq and DNAseq. Notably, shared variants with high numbers of MAF may arise from dominant tumor clones and are highly expressed, making them potential neoantigen candidates. On the other hand, unique variants displaying low MAFs may be derived from subclonal mutations or poorly expressed mutations, further emphasizing the influence of MAF and gene expression on the distinct characteristics of the identified variants.

In silico analysis of HLA-I binding affinity and immunogenicity of neoantigens derived from DNAseq and RNAseq

To identify neoantigen candidates, we utilized the pVAC-Seq pipeline, a well-established computation tool, to predict the binding affinity of 8-13 mer peptides generated from DNA or RNA variants to patient-specific HLA class I molecules (42). The HLA-I allele profiles of 25 patients were presented in Table S6. Through our analysis, we identified a total of 48,155 DNA-unique variants derived neoantigen candidates (61.7%), 15,584 shared-variant derived neoantigen candidates (20%), and 14,532 RNA-unique derived neoantigen candidates (18.4%) (Figure 4A, Table S7). As expected, the proportions of candidates from each group showed a significant correlation with the proportions of nucleotide mutations (Figure S2A).

It is well established that effective activation of T cell responses relies on the presentation of neoantigens on the patient's HLA-I molecules (57). Here, we assessed the binding affinity of predicted neoantigen candidates from each group of tumor variants to HLA-I using NetMHCpan 4.1 (18). For this analysis, only neoantigen candidates with predicted percentile ranks of less than 2% were considered, in accordance with the recommendations provided by NetMHCpan. We further considered 0.5 and 2 as percentile rank cutoffs to identify strong binding and weak binding epitopes, respectively. In Figure 4B, we presented the density distribution of

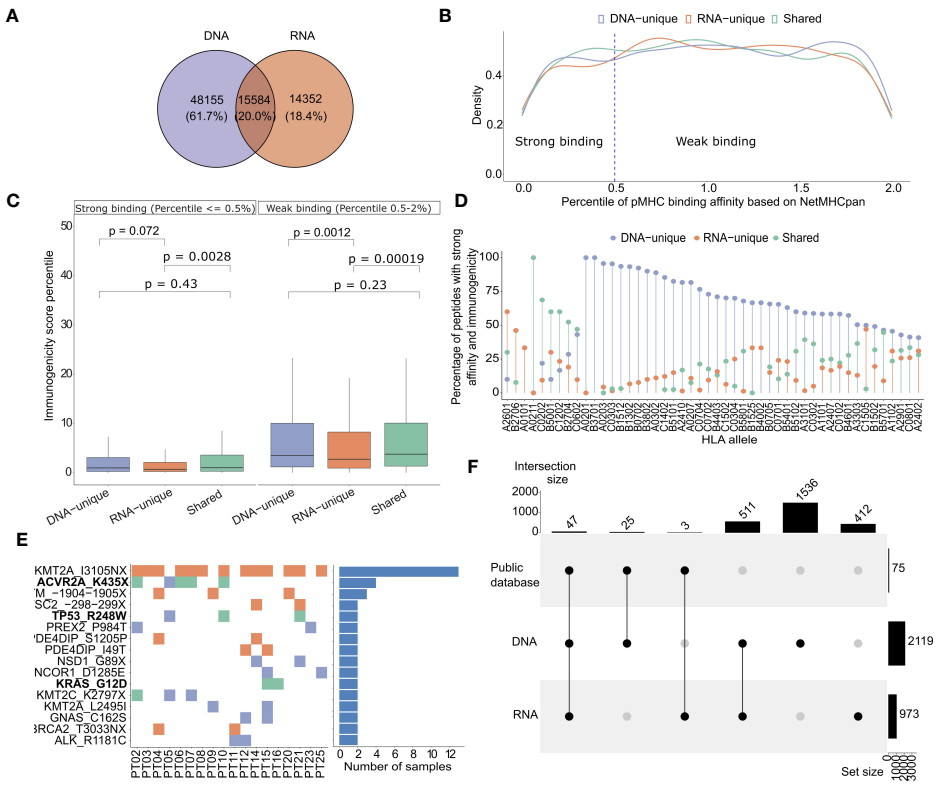

FIGURE 4
HLA-I binding affinity and immunogenicity of predicted neoantigens derived from DNAseq and RNAseq data. **(A)** A Venn diagram illustrates the proportion of each type of neoantigens identified from DNAseq and RNAseq data. **(B)** Histograms showing the density distribution of neoantigens with percentile ranks for HLA-I binding affinity calculated by NetMHCpan, that fall below 2%. The threshold value of 0.5% rank, designated for distinguishing strong and weak binders, is indicated by dashed lines. This distinction aligns with the recommendation provided by NetMHCpan. **(C)** Predicted immunogenicity, as calculated by the PRIME tool, for both strong binding and weak binding neoantigens. The box plot represents the median value, along with the lower and upper quartiles (25th and 75th percentiles). Outliers are not displayed for clarity of visualization. The p-values were estimated using the Wilcoxon rank-sum test. **(D)** A Lollipop plot depicts the distribution of specific groups of neoantigens based on their percentage, focusing on indicated HLA-I alleles. These plots highlight neoantigens that fall within the top 2% in terms of strong binding affinity to HLA-I and demonstrate high immunogenicity. **(E)** A map illustrates the frequency of indicated mutations on 25 CRC patients. The ones highlighted in bold have been previously validated as highly immunogenic through immunological assays in previous studies. **(F)** An UpSet plot illustrates the frequency distribution of the indicated groups of variants identified from public datasets.

predicted neoantigen candidates originating from DNA-unique, RNA-unique, or shared variants based on their percentile ranks of HLA-I binding affinity as predicted by NetMHCpan 4.1 (18). We observed that neoantigen candidates from RNA-unique variants exhibited a lower proportion a lower proportion of strong binding neoantigen (< 0.5%rank) compared to those from shared and DNA-unique variants (Figure 4B). This suggests that, in comparison to neoantigen candidates derived from DNA-unique variants, those originating from RNA-unique variants exhibited lower HLA-I binding affinity, as indicated by the NetMHCpan predictions. It has been reported that the binding affinity to HLA-I is determined by specific anchor residues in neopeptides (58). When comparing DNA-unique and shared neoantigens with RNA-unique neoantigens, it was observed that the latter exhibited a reduced proportion of mutations at P2 (Figure S2B). Notably, P2 serves as a crucial anchor residue involved in the primary interactions between the peptide and HLA-I molecule, and mutations occured within this position increase the binding affinity to HLA-I. This observation suggests that the decreased frequency of RNA-unique derived neoantigens carrying mutations at this anchor site, in comparison to other sources of neoantigens, may account for their lower binding affinity.

To assess the immunogenicity of the predicted candidates, we employed the PRIME tool which captures biophysical properties of both antigen presentation and TCR recognition to evaluate their potential to elicit a $CD8^+$ T cell-specific immune response (43). The predicted immunogenicity of neoantigen candidates was evaluated in relation to their predicted binding affinity to HLA-I (Figure 4C). We observed a positive correlation between the predicted binding affinity to HLA-I using NetMHCpan and the predicted immunogenicity assessed by the PRIME tool, irrespective of the neoantigen candidate class. Notably, strong binding neoantigen candidates exhibited lower percentile ranks of immunogenicity (Figure 4C). However, among the neoantigen candidates with strong HLA-I binding affinity, the RNA-unique neoantigen candidates showed significantly lower percentiles of immunogenicity compared to both DNA-unique (P=0.0075, Figure 4C) and shared neoantigen candidates (P= 0.0045, Figure 4C). Within the weak binding neoantigen candidates, RNA-unique neoantigen candidates consistently demonstrated lower percentiles of immunogenicity compared to DNA-unique (P=0.0012, Figure 4C) and shared neoantigen candidates (P=0.0011, Figure 4C). Subsequently, neoantigen candidates meeting the criteria for predicted binding affinity and immunogenicity within the top two percentile for both parameters were profiled based on the specific HLA-I alleles identified in our cohort of 25 CRC patients. As shown in Figure 4D, we observed that the binding affinity of predicted neoantigen candidates to HLA-I was influenced by both the specific neoantigen candidate's sequence and the HLA-I allele. For instance, we observed that the HLA-I allele A02011 exhibited a higher binding affinity to shared neoantigen candidates, as this allele showed the highest proportion of detected neoantigen candidates in this group. Similarly, the HLA-I allele A2601 displayed a stronger binding affinity for RNA-unique derived neoantigen candidates; while the HLA-I allele A0201 showed a stronger binding affinity for DNA-unique derived neoantigen candidates, in comparison to shared and RNA-unique neoantigen candidates (Figure 4D). Among the neoantigen candidates displaying strong predicted affinity and immunogenicity, a noteworthy subset of 16 neoantigen candidates was consistently identified in at least two patients (Figure 4E). Of those, neoantigen candidates derived from three shared mutations (ACVR2A_K435X, TP53_R428W, and KRAS_G12D) have been experimentally validated in previous studies and reported in public databases of immunogenic neoantigens. Notably, the KMT2A_IN3105X neoantigen candidate predicted from an RNA-unique variant, exhibited the highest frequency among these frequently detected neoantigen candidates, being present in 13 out of 25 (52%) patients. This suggests that this neoantigen candidate has the potential to serve as a public neoantigen, capable of eliciting immune responses across multiple individuals. Additionally, a total of 75 strong affinity and immunogenic neoantigen candidates were previously reported in public databases of immunogenic peptides. Among these, the majority (47/75, 62.7%%) could be found from shared variants, while 25 and 3 neoantigens were predicted from DNA-unique and RNA-unique variants, respectively (Figure 4F). These findings underscore the presence of both shared and unique neoantigen candidates with strong binding and immunogenicity in 25 analyzed patients, further highlighting the importance of considering different sources of NGS data for mutation identification in neoantigen-based immunotherapy approaches.

Taken together, these findings emphasize the distinct binding affinity and immunogenic potential of neoantigen candidates originating from different variant groups. Particularly, our data suggests that despite their low predicted binding affinity, neoantigen candidates derived from RNA somatic mutations still exhibit high immunogenicity, indicating their potential to elicit an immune response for immunotherapy. These observations underscore the importance of considering not only DNAseq but also RNAseq derived variants for selecting candidate neoantigens.

Experimental validation of predicted neoantigen candidates by ELISpot

To evaluate the effectiveness of integrating RNAseq variant calling into the current standard method, we conducted ELISpot assay on four CRC patients using autologous PBMCs following the procedure outlined in Figure 5A. Initially, we identified 431 nonsynonymous variants from both DNAseq and RNAseq data, resulting in a total of 18,479 predicted neoantigen candidates using the pVAC-Seq tool. To accommodate the limited availability of PBMCs, only the top ten mutations resulting in neoantigen candidates with the highest predicted binding affinity to HLA-I were chosen for each patient. As a result, a total of 40 synthesized long peptides (LPs) carrying the corresponding mutations were synthesized and used in an *ex vivo* ELISpot assay to measure the release of IFN-γ from patients' PBMCs (Figure 5A, Table S8).

Among the 40 designed LPs, those originating from shared neoantigen candidates were detected in all patients, whereas LPs derived from DNA-unique or RNA-unique variants were only detected in three out of four patients (Figure 5B). However, no LPs were identified within the DNA-unique group for patient PT10 and within the RNA-unique group for patient PT03 (Figure 5B). When

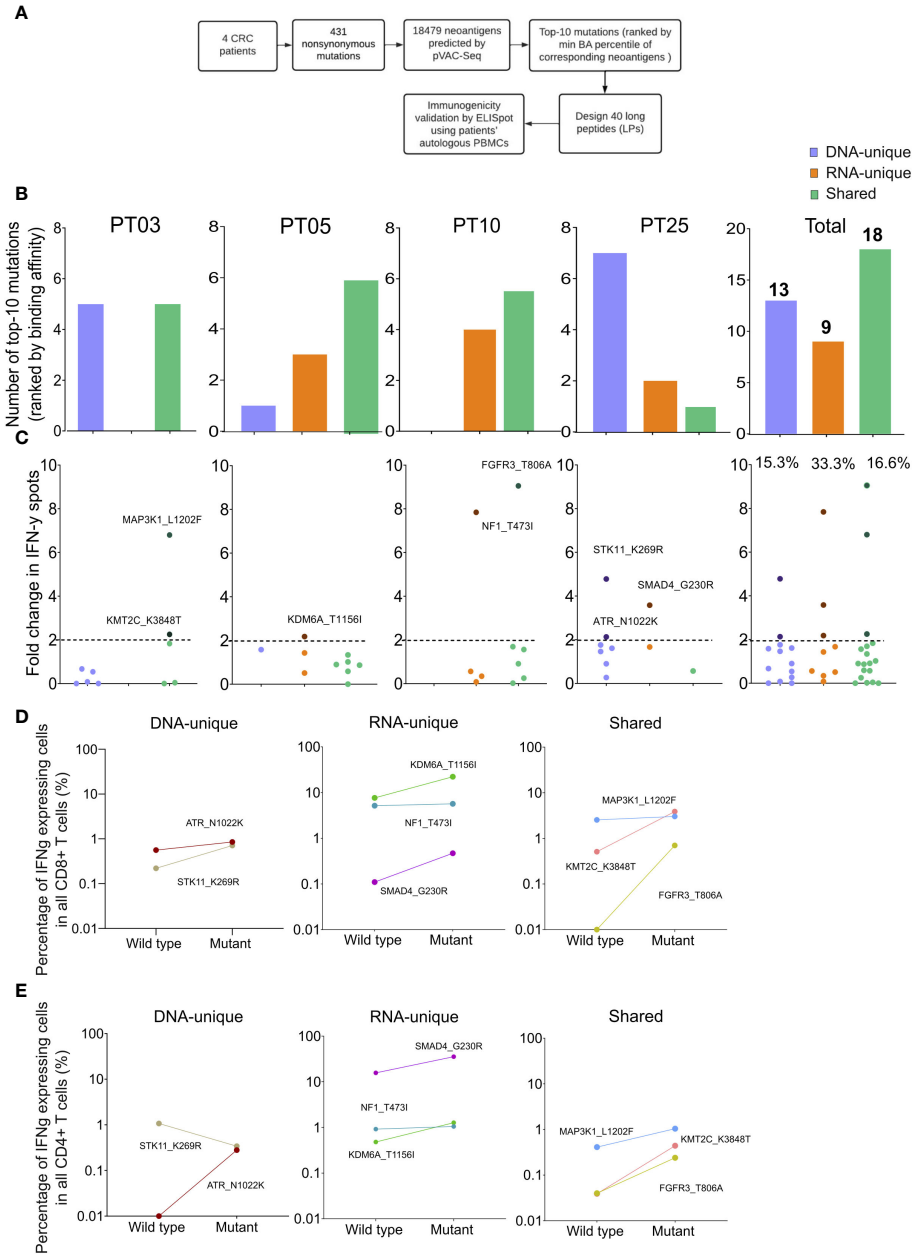

FIGURE 5
Validation of neoantigens *in silico* identified from the modified workflow by ELISpot assays on four CRC patients. **(A)** A schematic diagram illustrates the procedural steps of neoantigen prioritization and the ELISpot assay. **(B)** The number of each type of neoantigens identified from each CRC patient. **(C)** The fold change in IFN-γ spots, relative to the wildtype peptides, for 40 long peptides. Note: only the mutants that result in a positive value in ELISpot are depicted with their corresponding amino acid change. **(D)** The percentage of IFN-γ expressing CD4+ T cells induced by indicated long peptides. Note: these long peptides induce a more than 2-fold change in IFN-γ spots as observed in the ELISpot assay. **(E)** The percentage of IFN-γ expressing CD8+ T cells induced by indicated long peptides.

considering the cumulative number of LPs across all patients, it was observed that shared-variants yielded the highest number (18 out of 40), while RNA-unique variants yielded the fewest (9 out of 40, Figure 5B).

The PBMCs from four patients were subjected to three rounds of stimulation with 40 LPs carrying mutations or their corresponding wildtype counterparts to measure the secretion of IFN-γ. The ELISpot results for the 40 tested LPs were presented in Figure 5C and Table S8. A fold change of two in the number of IFN-γ spots from LPs relative to their corresponding wildtype peptides was chosen as the positivity cutoff, with LPs resulting in an ELISpot fold change value of two or higher considered as immunogenic (53). Among 40 tested LPs, we identified eight immunogenic LPs, with three originating from RNA-unique variants, three from shared variants, and two from DNA-unique variants (Figures 5C, S3). Notably, all four patients had at least one LP capable of inducing IFN-γ production by PBMCs. Among the LPs derived from RNA-

unique variants, three out of nine (33.3%) were positive for IFN-γ activation, while the proportions of positive LPs were lower for those derived from shared variants (three out of 18, 16.7%, Figure 5C) or DNA-unique variants (two out of 13, 15.4%, Figure 5C). The findings suggest that RNA-unique variants may result in fewer neoantigen candidates with strong binding affinity to HLA-I, but they are more likely to activate T cells compared to shared or DNA-unique neoantigen candidates.

Intracellular flow cytometry staining of IFN-γ in T cells further demonstrated that all LPs showing positive results in the ELISpot assay effectively activated CD8+ T cells. This activation led to a significant increase in the percentage of IFN-γ positive cells, with a fold increase greater than 1 compared to their corresponding wildtype peptides (Figures 5D, S4). Moreover, consistent with the activation of $CD8^+$ T cells, all LPs exhibited increased production of IFN-γ by $CD4^+$T cells, except for the LP carrying STK11_K269R, which originated from a DNA-unique variant (Figure 5E). Although this LP did not exhibit detectable changes in intracellular IFN- γ levels in $CD4^+$ T cells, it still demonstrated $CD8^+$ T cell activation. Overall, these findings suggested that the integration of RNAseq data for variant calling into the current neoantigen prediction workflow could enhance the identification of effective and immunogenic neoantigen candidates for the development of cancer immunotherapies.

Discussion

The identification of highly immunogenic neoantigens capable of eliciting T-cell-mediated responses is essential for the development of effective personalized immunotherapies for cancer. However, the current challenge lies in accurately identifying these neoantigens due to the limited number of highly immunogenic neopeptides predicted by conventional bioinformatic workflows. These workflows solely rely on genomic sequencing data for tumor mutation calling, overlooking the potential contribution of transcriptomic variants in generating neoantigens. To address this limitation, we aimed to enhance the identification of highly immunogenic neoantigens by integrating RNA sequencing data into the conventional bioinformatic workflow (Figure 1). By considering tumor mutations at the transcriptional level, we sought to expand the pool of valuable immunogenic neopeptides for colorectal cancer (CRC) patients. In our study, we successfully demonstrated that integrating RNAseq data into the conventional workflow for variant calling significantly increased the number of valuable immunogenic neopeptides for CRC patients. This improvement provides a promising avenue for the development of more effective cancer treatments.

Our analysis of tumor variants using DNAseq and RNAseq data obtained from 25 CRC patients identified a moderate proportion (22.4%) of shared somatic variants (Figure 2A). This finding is consistent with a previous study that reported a similar trend in two datasets (59). The differences in variants identified by DNAseq and RNAseq could be attributed to variations in sequencing technologies or variant calling tools, as reported in previous studies (60). To mitigate the impact of differences in sequencing technology and *in silico* tools on mutation results, we conducted both DNAseq and RNAseq on the same sequencing platform and selected the optimal variant calling tools for RNAseq data that exhibit the highest concordance with the DNAseq mutation profile (Figure S1A). However, we believe that more validation studies are required to improve the variant calling tools and standardize their use for RNA sequencing data. In addition to these technical factors, it has been reported that RNA mutations could be generated from a post-transcriptional modification process known as RNA editing (61, 62). Such mutations exclusively occur in transcribed RNA and have been shown to result in a new source of neoantigens in cancer patients (63, 64).

Additionally, the proportions of shared mutations exhibited significant variation among patients (Figure 2B), highlighting the intrinsic diversity of cancer mutations and the heterogeneity of clonal expansion within each patient. Furthermore, different variant groups displayed distinct characteristics, with RNA-variants showing an enrichment for frameshift and inframe variants and displaying more profound impact on the phenotypic outcome (Figures 2C, D). Neoantigens derived from frameshift or indel variants, which are greatly distinct from self peptides, have been shown to generate highly immunogenic tumor neoantigens and thereby expand the pool of ideal candidates for immunotherapy (65, 66).

Both DNA-unique and RNA-unique variants displayed significantly lower MAFs compared to shared variants (Figure 3B). This observation implies that these unique variants likely originated from tumor clones with low frequencies, which might not be consistently detected at both genomic and transcriptomic levels due to the limited sensitivity of sequencing methods. Notably, our analysis revealed that DNA-unique variants were more frequently associated with genes characterized by low FPKMs, unlike shared or RNA-unique variants (Figures 3C, D). These findings suggest that DNA-unique variants may arise from genes with low expression or those displaying mono-allelic expression of the wild-type allele. Conversely, RNA-unique or shared variants tend to occur in genes exhibiting high expression levels, implying their abundant transcription. Previous studies have demonstrated a correlation between the expression levels of neoantigens and their likelihood of being presented by HLA-I on the surface of tumor cells, which can trigger immune responses leading to the eradication of tumor cells (67, 68). Hence, neoantigens arising from RNA-unique or shared variants might be superior, as they are more likely to be presented and recognized by the immune system. The discrepancies in mutation profiles between RNAseq and DNAseq could be attributed to the low MAFs, low quantities of transcripts harboring variants, and/or insufficient sequencing coverage.

The proportions of neoantigens predicted by the pVAC-Seq tool are similar to those of nucleotide variants (Figures 3A, 4A). Currently, the prediction of peptide binding affinity for HLA-I is a pivotal criterion in the selection of neoantigens for experimental validation (18). Employing NetMHCpan 4.1, we discovered that neoantigen candidates originating from RNA-unique variants exhibited lower percentile ranks of binding affinity compared to those derived from shared or DNA-unique variants (Figure 4B). This finding suggests that neoantigen candidates resulting from

RNA variants tend to display reduced levels of HLA-I binding affinity in comparison to those arising from DNA variants. Prior research has indicated that the position of mutations within mutant peptides can influence their binding affinity to HLA-I molecules, with specific residues in the peptides, known as anchor residues, serving as key determinants of binding affinity (69). Therefore, it is plausible that amino acid changes in neoantigen candidates predicted from RNA mutations may arise from positions that do not lead to enhanced binding affinity, in contrast to those arising from DNA mutations. Interestingly, our findings revealed a lower proportion of RNA-derived neoantigen candidates with mutations occurring at the primary anchor site P2, which is recognized as a critical factor influencing peptide affinity for various HLA-I types. This distinction was observed when comparing RNA-derived neoantigen candidates with both shared and DNA-unique derived ones (Figure S2B) (70). Another possible explanation for the lower binding affinity of RNA-unique neoantigen candidates could be attributed to the fact that current prediction tools have not been specifically trained on this particular group of candidates (71).

While predicted HLA-I binding affinity serves as a crucial indicator for the presentation of neoantigens on tumor cells, it is not the sole determinant of neoantigen immunogenicity. The immunogenicity of neoantigens is also influenced by the interaction between peptide-HLA complexes and T cell receptors (TCR) (43, 72, 73). Therefore, in our study, we initially selected neopeptides with strong binding affinity (< 2% percentile rank). Subsequently, we employed the PRIME tool (43), which captures molecular properties related to both antigen presentation and TCR recognition, to estimate the immunogenicity of these selected peptides. Interestingly, we observed that neoantigen candidates derived from RNA-unique mutations or shared mutations exhibited significantly higher immunogenicity compared to those derived from DNA-unique mutations (Figure 4C). Schmidt et al. have identified specific amino acid positions within the neopeptide sequence, known as minimally impacting on HLA-I affinity positions. These positions have been found to have significant roles in binding to the T cell receptor (TCR) (43). Therefore, it is plausible that amino acid changes in neopeptides derived from RNA mutations may occur at such positions, resulting in enhanced TCR affinity and consequently explaining their stronger immunogenicity. Analysis of neoantigen candidates' immunogenicity, considering the HLA-I allele panels obtained from our CRC patient cohort, revealed a notable dependence on specific HLA-I alleles, thereby emphasizing the significance of profiling the HLA-I genotype of cancer patients for personalized immunotherapy (Figure 4D). The notable immunogenicity scores of neoantigen candidates derived from RNA variants suggest their potential to effectively activate T cell-mediated immune responses, rendering them valuable candidates for clinical evaluation. Our *in silico* analysis successfully identified a recurrent RNA-derived neoantigen candidate (KMT2A_IN3105X) in 25 CRC patients. Additionally, we discovered three shared candidates (ACVR2A_K435X, TP53_R428W, and KRAS_G12D) that have been experimentally validated as highly immunogenic in publicly available databases (Figures 4E, F). These neopeptides hold potential as public neoantigens, making them suitable candidates for an off-the-shelf vaccine strategy. Thus, we speculate that incorporating RNA-unique variants, which exhibit strong binding affinity and higher transcription abundance, can serve as a strategy to identify more effective targets for neoantigen-based vaccination.

To validate our hypothesis regarding the effectiveness of neoantigen candidates derived from RNA variants compared to DNA-derived candidates, we conducted *ex vivo* ELISpot assays on four patients with available blood samples for PBMC collection. The purpose was to assess the immunogenicity of predicted neoantigen candidates originating from different mutation sources. For each patient, we selected the top 10 mutations based on the predicted binding affinity of the corresponding neopeptides to the patients' HLA-I profile. To evaluate immunogenicity, we designed LPs incorporating these mutations (Figure 5A). Consistent with our analysis on 25 CRC patients, the proportion of LPs derived from RNA-unique mutations with strong binding affinity was lower compared to those derived from DNA-unique or shared mutations (Figure 5B). However, in the *ex vivo* ELISpot assays, three out of nine LPs (33.3%) carrying RNA-unique variants triggered IFN-γ production in PBMCs of three out of four patients, while only two out of 13 LPs (15.3%) carrying DNA-unique variants induced IFN-γ production in a single patient (Figure 5C). In line with the ELISpot data, we detected IFN-γ activation not only in $CD8^+$ T cells but also in $CD4^+$ T cells for most of the tested long peptides. However, one LP derived from a DNA-unique mutation exclusively activated $CD8^+$ T cells (Figures 5D, E). Our selection and design of LPs was based on the rank of neopeptide candidates' HLA-I binding affinity, aiming to specifically activate $CD8^+$ T cells. However, our findings align with a previous study demonstrating that LPs covering target mutations could be intracellularly processed to peptides of differrent lengths and subsequenty presented to both $CD4^+$ and $CD8^+$ T cells (74). Our *ex vivo* validation of neoantigens' immunogenicity using patients' PBMCs provides compelling experimental evidence that relying solely on DNAseq data for tumor mutation calling would overlook valuable neoantigen candidates derived from RNA variants and that integrating variant calling by RNAseq into this process significantly enhances the likelihood of detecting immunogenic neoantigens.

This study has several limitations that should be acknowledged. Firstly, in order to develop a cost-effective workflow for neoantigen identification, the analysis was focused on SNV and indel variants within only 95 cancer-associated genes. Consequently, other types of mutations, such as gene fusions and alternative splicing, and other genes were not explored (75, 76). Secondly, while RNAseq holds the potential to identify mutations on a genome-wide scale, its sensitivity and specificity are influenced by many factors such as sequencing depth, tumor purity, and the variant calling pipeline. To mitigate the potential impact of these biases, we carefully selected the optimal mutation caller for RNAseq data, VarScan, after comparing its performance with MuTect2. However, more validation studies are necessary to

improve the variant calling tools for RNAseq data and standardize their use. Thirdly, the study was conducted with a limited sample size of 25 CRC patients, and the experimental validation of predicted neoantigens through *ex-vivo* ELISpot assays was performed on only four patients due to the availability of blood samples. As a result, the generalizability of the findings may be constrained. Finally, the assessment of the immunogenicity of candidate LPs relied exclusively on *ex-vivo* stimulation of patients' PBMCs, which may not accurately reflect the natural presentation of neoantigens by HLA-I molecules expressed in patients' tumor cells. Therefore, additional experimental validation using liquid chromatography mass spectrometry-based immunopeptidomics may be required to confirm the presentation of predicted neoantigens on HLA-I molecules in tumor cells.

Taken together, in this proof-of concept study, we provide compelling evidence for the benefits of utilizing RNAseq-guided mutations for neoantigen prediction, as it allows for the identification of a larger pool of potential and highly immunogenic neoantigens by leveraging additional information from RNAseq data beyond conventional gene expression levels.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: BioProject via accession ID PRJNA1005034.

Ethics statement

This study was approved by the Ethics Committee of University of Medicine and Pharmacy at Ho Chi Minh City, Vietnam. The patients/participants provided written informed consent for the collection of tumor and whole blood samples.

Author contributions

BN and TPDT conduct experiments, perform formal analysis, curate data, and develop methodologies. HTN is responsible for patient recruitment and conceptualization. TN specializes in data curation and formal analysis. TP conducts experiments and performs formal analysis. HTPN and V.N conduct experiments, perform formal analysis, and curate data. DT, TST, TP, and ML recruit patients and analyze data. MP, HG, and HNN conceptualize the study and edit writings. LT conceptualizes the study, writes the original manuscript, and edits the final document. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by a NexCalibur Therapeutic grant.

Acknowledgments

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

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1251603/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Evaluation of mutation calling tools for DNAseq and RNAseq data **(A)** Comparison of performance of three indicated mutation callers on a reference DNAseq dataset. **(B)** A Venn diagram illustrates the number of mutations identified by Dragen and two RNA mutation callers, VarScan and MuTect2. **(C)** Proportions of SNV and indel mutations called by indicated tools. **(D)** Length distribution of INDEL mutations called by indicated tools

SUPPLEMENTARY FIGURE 2
Distribution of mutation positions of DNAseq and RNAseq derived neoantigen **(A)** Correlation between the numbers of variants and neoantigens within the indicated groups. **(B)** A lollipop plot displays the percentage of neoantigens from the indicated groups that contain mutations at positions 1 to 12. The blue box represents the anchor site of the peptide and HLA-I molecule.

SUPPLEMENTARY FIGURE 3
ELISpot assays on eight long peptides which result in 2-fold change of IFN-γ spots.

SUPPLEMENTARY FIGURE 4
Gating strategy for detecting IFN-γ production from $CD4^+$ and $CD8^+$ T cells in LP-stimulated PBMCs of 4 CRC patients.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2018) 68:394–424. doi: 10.3322/caac.21492

2. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: An overview. *Int J Cancer* (2021) 149:778–89. doi: 10.1002/ijc.33588

3. Biller LH, Schrag D. Diagnosis and treatment of metastatic colorectal cancer: A review. *JAMA* (2021) 325:669–85. doi: 10.1001/jama.2021.0106

4. Ciardiello D, Vitiello P P, Cardone C, Martini G, Troiani T, Martinelli E, et al. Immunotherapy of colorectal cancer: Challenges for therapeutic efficacy. *Cancer Treat Rev* (2019) 76:22–32. doi: 10.1016/j.ctrv.2019.04.003

5. Overman MJ, Ernstoff MS, Morse MA. Where we stand with immunotherapy in colorectal cancer: deficient mismatch repair, proficient mismatch repair, and toxicity management. *Am Soc Clin Oncol Educ Book* (2018) 38:239–47. doi: 10.1200/EDBK_200821

6. Dudley JC, Lin MT, Le DT, Eshleman JR. Microsatellite instability as a biomarker for PD-1 blockade. *Clin Cancer Res* (2016) 22:813–20. doi: 10.1158/1078-0432.CCR-15-1678

7. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N Engl J Med* (2015) 372:2509–20. doi: 10.1056/NEJMoa1500596

8. Overman MJ, McDermott R, Leach JL, Lonardi S, Lenz HJ, Morse MA, et al. 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* (2017) 18:1182–91. doi: 10.1016/S1470-2045(17)30422-9

9. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* (2017) 357:409–13. doi: 10.1126/science.aan6733

10. Yu Y, Zhang J, Ni L, Zhu Y, Yu H, Teng Y, et al. Neoantigen-reactive T cells exhibit effective anti-tumor activity against colorectal cancer. *Hum Vaccin Immunother* (2022) 18:1–11. doi: 10.1080/21645515.2021.1891814

11. Kim VM, Pan X, Soares KC, Azad NS, Ahuja N, Gamper CJ, et al. Neoantigen-based EpiGVAX vaccine initiates antitumor immunity in colorectal cancer. *JCI Insight* (2020) 5(9):e136368. doi: 10.1172/jci.insight.136368

12. Yi M, Qin S, Zhao W, Yu S, Chu Q, Wu K. The role of neoantigen in immune checkpoint blockade therapy. *Exp Hematol Oncol* (2018) 7:28. doi: 10.1186/s40164-018-0120-y

13. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nat Rev Clin Oncol* (2021) 18:215–29. doi: 10.1038/s41571-020-00460-2

14. Miao D, Margolis CA, Gao W, Voss MH, Li W, Martini DJ, et al. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* (2018) 359:801–6. doi: 10.1126/science.aan5951

15. Yarchoan M, Hopkins A, Jaffee EM. Tumor mutational burden and response rate to PD-1 inhibition. *N Engl J Med* (2017) 377:2500–1. doi: 10.1056/NEJMc1713444

16. Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, et al. pVACtools: A computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol Res* (2020) 8:409–20. doi: 10.1158/2326-6066.CIR-19-0401

17. Chheda ZS, Kohanbash G, Okada K, Jahan N, Sidney J, Pecoraro M, et al. Novel and shared neoantigen derived from histone 3 variant H3.3K27M mutation for glioma T cell therapy. *J Exp Med* (2018) 215:141–57. doi: 10.1084/jem.20171046

18. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* (2020) 48:W449–54. doi: 10.1093/nar/gkaa379

19. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science* (2015) 348:69–74. doi: 10.1126/science.aaa4971

20. Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* (2014) 515:572–6. doi: 10.1038/nature14001

21. Tran E, Turcotte S, Gros A, Robbins PF, Lu YC, Dudley ME, et al. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* (2014) 344:641–5. doi: 10.1126/science.1251102

22. van Rooij N, van Buuren MM, Philips D, Velds A, Toebes M, Heemskerk B, et al. Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J Clin Oncol* (2013) 31:e439–442. doi: 10.1200/JCO.2012.47.7521

23. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* (2012) 366:883–92. doi: 10.1056/NEJMoa1113205

24. Yang HD, Nam SW. Pathogenic diversity of RNA variants and RNA variation-associated factors in cancer development. *Exp Mol Med* (2020) 52:582–93. doi: 10.1038/s12276-020-0429-6

25. Obeng EA, Stewart C, Abdel-Wahab O. Altered RNA processing in cancer pathogenesis and therapy. *Cancer Discovery* (2019) 9:1493–510. doi: 10.1158/2159-8290.CD-19-0399

26. Borden ES, Ghafoor S, Buetow KH, LaFleur BJ, Wilson MA, Hastings KT. NeoScore integrates characteristics of the neoantigen : MHC class I interaction and expression to accurately prioritize immunogenic neoantigens. *J Immunol* (2022) 208:1813–27. doi: 10.4049/jimmunol.2100700

27. Hashimoto S, Noguchi E, Bando H, Miyadera H, Morii W, Nakamura T, et al. Neoantigen prediction in human breast cancer using RNA sequencing data. *Cancer Sci* (2021) 112:465–75. doi: 10.1111/cas.14720

28. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* (2012) 22:568–76. doi: 10.1101/gr.129684.111

29. Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci Rep* (2020) 10:20222. doi: 10.1038/s41598-020-77218-4

30. Severine Catreux VJ, Murray L, Mehio R, Parnaby G, Roddey C, Ruehle M, et alAvailable at: https://www.illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html

31. Richards NM. Secretary upholds FDA on generics. *Pa Med* (1990) 93:28.

32. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170

33. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) 29:15–21. doi: 10.1093/bioinformatics/bts635

34. FastQC, A. Quality control tool for high throughput sequence data. *BibSonomy* (2015). Available online: https://www. bibsonomy. org/bibtex/f230a919c34360709aa298734d63dca3. (Accessed March 17, 2022).

35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* (2009) 25:2078–9. doi: 10.1093/bioinformatics/btp352

36. Available at: http://broadinstitute.github.io/picard.

37. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. *Genome Biol* (2016) 17:122. doi: 10.1186/s13059-016-0974-4

38. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* (2010) 28:511–5. doi: 10.1038/nbt.1621

39. Available at: https://www.R-project.org/.

40. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* (2014) 30:3310–6. doi: 10.1093/bioinformatics/btu548

41. Hundal J, Kiwala S, Feng YY, Liu CJ, Govindan R, Chapman WC, et al. Accounting for proximal variants improves neoantigen prediction. *Nat Genet* (2019) 51:175–9. doi: 10.1038/s41588-018-0283-9

42. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med* (2016) 8:11. doi: 10.1186/s13073-016-0264-5

43. Schmidt J, Smith AR, Magnin M, Racle J, Devlin JR, Bobisse S, et al. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep Med* (2021) 2:100194. doi: 10.1016/j.xcrm.2021.100194

44. Wu J, Zhao W, Zhou B, Su Z, Gu X, Zhou Z, et al. TSNAdb: A database for tumor-specific neoantigens from immunogenomics data analysis. *Genomics Proteomics Bioinf* (2018) 16:276–82. doi: 10.1016/j.gpb.2018.06.003

45. Wu J, Chen W, Zhou Y, Chi Y, Hua X, Wu J, et al. TSNAdb v2.0: the updated version of tumor-specific neoantigen database. *Genomics Proteomics Bioinf* (2022) S1672-0229(22)00128-0. doi: 10.1016/j.gpb.2022.09.012

46. Zhou WJ, Qu Z, Song CY, Sun Y, Lai AL, Luo MY, et al. NeoPeptide: an immunoinformatic database of T-cell-defined neoantigens. *Database* (2019) 2019: baz128. doi: 10.1093/database/baz128

47. Lu M, Xu L, Jian X, Tan X, Zhao J, Liu Z, et al. dbPepNeo2.0: A database for human tumor neoantigen peptides from mass spectrometry and TCR recognition. *Front Immunol* (2022) 13:855976. doi: 10.3389/fimmu.2022.855976

48. Tan X, Li D, Huang P, Jian X, Wan H, Wang G, et al. dbPepNeo: a manually curated database for human tumor neoantigen peptides. *Database* (2020) 2020: baaa004. doi: 10.1093/database/baaa004

49. Xia J, Bai P, Fan W, Li Q, Li Y, Wang D, et al. NEPdb: A database of T-cell experimentally-validated neoantigens and pan-cancer predicted neoepitopes for cancer immunotherapy. *Front Immunol* (2021) 12:644637. doi: 10.3389/fimmu.2021.644637

50. Zhang G, Chitkushev L, Olsen LR, Keskin DB, Brusic V. TANTIGEN 2.0: a knowledge base of tumor T cell antigens and epitopes. *BMC Bioinf* (2021) 22:40. doi: 10.1186/s12859-021-03962-7

51. Olsen LR, Tongchusak S, Lin H, Reinherz EL, Brusic V, Zhang GL. TANTIGEN: a comprehensive database of tumor T cell antigens. *Cancer Immunol Immunother* (2017) 66:731–5. doi: 10.1007/s00262-017-1978-y

52. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* (2019) 47:D339–43. doi: 10.1093/nar/gky1006

53. Moodie Z, Price L, Gouttefangeas C, Mander A, Janetzki S, Löwer M, et al. Response definition criteria for ELISPOT assays revisited. *Cancer Immunol Immunother* (2010) 59:1489–501. doi: 10.1007/s00262-010-0875-4

54. Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res* (2009) 37: e106. doi: 10.1093/nar/gkp507

55. Katzir R, Rudberg N, Yizhak K. Estimating tumor mutational burden from RNA-sequencing without a matched-normal sample. *Nat Commun* (2022) 13:3092. doi: 10.1038/s41467-022-30753-2

56. Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nat Biotechnol* (2021) 39:1151–60. doi: 10.1038/s41587-021-00993-6

57. Xie N, Shen G, Gao W, Huang Z, Huang C, Fu L. Neoantigens: promising targets for cancer therapy. *Signal Transduct Target Ther* (2023) 8:9. doi: 10.1038/s41392-022-01270-x

58. Nguyen AT, Szeto C, Gras S. The pockets guide to HLA class I molecules. *Biochem Soc Trans* (2021) 49:2319–31. doi: 10.1042/BST20210410

59. Tretter C, de Andrade Kraetzig N, Pecoraro M, Lange S, Seifert P, von Frankenberg C, et al. Proteogenomic analysis reveals RNA as an important source for tumor-agnostic neoantigen identification correlating with T-cell infiltration. *bioRxiv* (2022) 14(1):4632. doi: 10.1101/2022.09.17.508207

60. Koboldt DC. Best practices for variant calling in clinical sequencing. *Genome Med* (2020) 12:91. doi: 10.1186/s13073-020-00791-w

61. Guo Y, Yu H, Samuels DC, Yue W, Ness S, Zhao YY. Single-nucleotide variants in human RNA: RNA editing and beyond. *Brief Funct Genomics* (2019) 18:30–9. doi: 10.1093/bfgp/ely032

62. O'Brien TD, Jia P, Xia J, Saxena U, Jin H, Vuong H, et al. Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer. *Methods* (2015) 83:118–27. doi: 10.1016/j.ymeth.2015.04.016

63. Komatsu Y, Shigeyasu K, Yano S, Takeda S, Takahashi K, Hata N, et al. RNA editing facilitates the enhanced production of neoantigens during the simultaneous administration of oxaliplatin and radiotherapy in colorectal cancer. *Sci Rep* (2022) 12:13540. doi: 10.1038/s41598-022-17773-0

64. Wang W, Zhou W, Lu Q, Ding T, Fan L. Investigating the clinical relevance of RNA editing events and their derived neoantigens in patients with melanoma treated with immunotherapy. *J Clin Oncol* (2023) 41:e21580–0.

65. Mattos-Arruda L, Vazquez M, Finotello F, Lepore R, Porta E, Hundal J, et al. Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the ESMO Precision Medicine Working Group. *Ann Oncol* (2020) 31:978–90. doi: 10.1016/j.annonc.2020.05.008

66. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol* (2017) 18:1009–21. doi: 10.1016/S1470-2045(17)30516-8

67. Lo AA, Wallace A, Oreper D, Lounsbury N, Havnar C, Pechuan-Jorge X, et al. Indication-specific tumor evolution and its impact on neoantigen targeting and biomarkers for individualized cancer immunotherapies. *J Immunother Cancer* (2021) 9(10):e003001. doi: 10.1101/2021.03.15.434617

68. Yi M, Dong B, Chu Q, Wu K. Immune pressures drive the promoter hypermethylation of neoantigen genes. *Exp Hematol Oncol* (2019) 8:32. doi: 10.1186/s40164-019-0156-7

69. Capietto AH, Jhunjhunwala S, Pollock SB, Lupardus P, Wong J, Hänsch L, et al. Mutation position is an important determinant for predicting cancer neoantigens. *J Exp Med* (2020) 217(4):e20190179. doi: 10.1084/jem.20190179

70. Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A, et al. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res* (2008) 4:2. doi: 10.1186/1745-7580-4-2

71. Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M. Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome Med* (2019) 11:56. doi: 10.1186/s13073-019-0666-2

72. Montemurro A, Schuster V, Povlsen HR, Bentzen AK, Jurtz V, Chronister WD, et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRalpha and beta sequence data. *Commun Biol* (2021) 4:1060. doi: 10.1038/s42003-021-02610-3

73. Pham MN, Nguyen TN, Tran LS, Nguyen QTB, Nguyen TPH, Pham TMQ, et al. epiTCR: a highly sensitive predictor for TCR-peptide binding. *Bioinformatics* (2023) 39 (5):btad284. doi: 10.1093/bioinformatics/btad284

74. Chen X, Yang J, Wang L, Liu B. Personalized neoantigen vaccination with synthetic long peptides: recent advances and future perspectives. *Theranostics* (2020) 10:6011–23. doi: 10.7150/thno.38742

75. Yang W, Lee KW, Srivastava RM, Kuo F, Krishna C, Chowell D, et al. Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nat Med* (2019) 25:767–75. doi: 10.1038/s41591-019-0434-2

76. Slansky JE, Spellman PT. Alternative splicing in tumors - A path to immunogenicity? *N Engl J Med* (2019) 380:877–80. doi: 10.1056/NEJMcibr1814237