

# NeoTCR: An Immunoinformatic Database of Experimentally-supported Functional Neoantigen-specific TCR Sequences

Weijun Zhou (周炜均) <sup>1,2,#</sup>, Wenting Xiang (向文婷) <sup>3,#</sup>, Jinyi Yu (喻瑾怡) <sup>2</sup>,  
Zhihan Ruan (阮志涵) <sup>3</sup>, Yichen Pan (潘逸辰) <sup>3</sup>, Kankan Wang (王侃侃) <sup>2,\*</sup>,  
Jian Liu (刘健) <sup>3,4,\*</sup>

<sup>1</sup>Department of Hematology, Zhujiang Hospital, Southern Medical University, Guangzhou 510282, China

<sup>2</sup>Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

<sup>3</sup>College of Computer Science, Nankai University, Tianjin 300350, China

<sup>4</sup>Centre for Bioinformatics and Intelligent Medicine, Institute of Big Data, Nankai University, Tianjin 300350, China

\*Corresponding authors: kankanwang@shsmu.edu.cn (Wang K), jianliu@nankai.edu.cn (Liu J).

#Equal contribution.

Handling Editor: Qian-Fei Wang

## Abstract

Neoantigen-based immunotherapy has demonstrated long-lasting antitumor activity. The recognition of neoantigens by T cell receptors (TCRs) is considered a trigger for antitumor responses. Due to the overwhelming number of TCR repertoires in the human genome, pinpointing neoantigen-specific TCRs is a formidable challenge. Recent studies have identified a number of functional neoantigen-specific TCRs, but the corresponding information is scattered across published literature and is difficult to retrieve. To improve access to these data, we developed an immunoinformatic database (NeoTCR) containing a unified description of publicly available neoantigen-specific TCR sequences, as well as relevant information on targeted neoantigens, from experimentally-supported studies across 17 cancer subtypes. A user-friendly web interface allows interactive browsing and running of complex database queries. To facilitate rapid identification of neoantigen-specific TCRs from raw sequencing data, NeoTCR offers a one-stop analysis for annotation and visualization of TCR clonotypes, discovery of existing neoantigen-specific TCRs, and exclusion of bystander virus-associated TCRs. NeoTCR represents a unique tool to expedite future studies of neoantigen-specific TCRs and the development of neoantigen-based immunotherapy. NeoTCR is available at <http://neotcrdb.bioxai.cn/> and <https://github.com/lyotvincent/NeoTCR>.

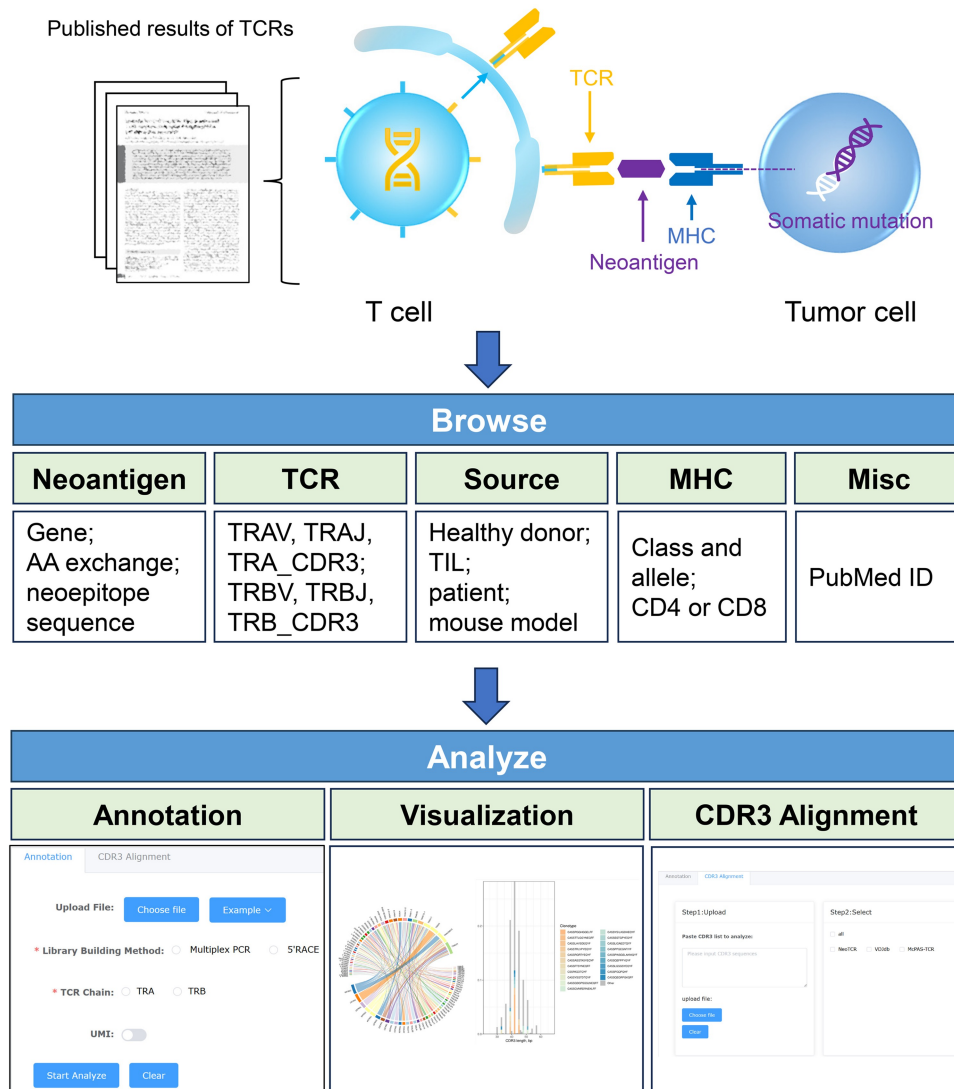
**Key words:** T cell receptor; Neoantigen; CDR3; Immunotherapy; Database.

Received: 17 October 2022; Revised: 21 April 2023; Accepted: 26 May 2023.

© The Author(s) 2024. Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

## Graphical abstract



## Introduction

Neoantigens arise from somatic mutation events that are exclusively present in tumor cells and thus represent ideal targets for immunotherapy [1]. Neoantigen-based immunotherapies have been shown to produce durable responses and long-term remission in patients with metastatic or advanced cancer [2–5]. The responses to these immunotherapies are triggered by T cell receptors (TCRs) specific to neoantigens presented by human leukocyte antigen (HLA) on tumor cells [6–9].

Recent advances in immunology and high-throughput TCR sequencing have facilitated experimental efforts to map TCR repertoires for a given epitope, including neoantigens [10–12]. However, the major challenge in this field is identifying tumor-reactive and neoantigen-specific TCRs from enormous TCR sequencing data. While several methods for TCR specificity prediction [13,14] have been established, they do not allow for predicting TCR-specific neoantigens, largely due to the scarcity of training data computationally linking a given TCR sequence to a specific neoantigen. Currently, identifying neoantigen-specific TCRs relies on time-consuming biological validation, e.g., *in vitro* assays based on T cell recognition of targets

expressing specified peptide–HLA molecules (peptide–HLA multimers) and specific cytotoxic trails using TCR-engineered T cells. More importantly, a particular neoantigen may trigger several DNA rearrangement processes, leading to a diversity of TCR repertoires, and a particular neoantigen-specific TCR may reappear in various cancer patients with the same mutation [15,16]. Additionally, the samples submitted for TCR sequencing may contain bystander T cells with virus-specific TCRs [9].

Several databases of TCR sequences, including VDJdb [17,18], McPAS-TCR [19], and TCRdb [20], are currently available. These databases contain a large number of TCR sequences from humans and other species. VDJdb and McPAS-TCR also include information on TCR-related antigens. However, each TCR sequence in VDJdb is an  $\alpha$  (TRA) or a  $\beta$  (TRB) chain, rather than a paired  $\alpha\beta$ TCR. McPAS-TCR provides paired  $\alpha\beta$ TCRs, but the majority of these are associated with viruses. TCRdb holds more than 277 million highly reliable TCR sequences but no information on related antigens and HLA alleles. In essence, the existing databases lack TCR sequences specific to neoantigens, especially the TCRs with complete  $\alpha$  and  $\beta$  chains that can be directly used

in the development of TCR–T cell-based immunotherapy for clinical application.

We developed an immunoinformatic database of TCR sequences associated with numerous neoantigens in a variety of malignancies, referred to as NeoTCR. This database contains comprehensive details on each TCR sequence, including the related neoantigen and corresponding neoepitope, the restriction of the HLA allele, the tissue of origin, and the original publication. NeoTCR assists in the detection of neoantigen-specific TCRs from raw sequencing data and represents a one-stop platform for annotation and visualization of TCR clonotypes, identification of existing neoantigen-specific TCRs, and exclusion of bystander virus-associated TCRs.

## Implementation

The NeoTCR database was developed based on the web and browser/server architecture and implemented using Spring Boot and Vue. All TCR sequences related to neoantigens were stored in MySQL. The user interface was developed using Element UI and the data visualization was developed using ECharts, which provides intuitive figures to describe the follow-up analysis results (Figure 1A).

## Database content and usage

### Overview of NeoTCR and data summary

Currently, NeoTCR contains 988 entries curated from 52 publications, encompassing TCR sequences from 17 types of cancer, including solid tumors and hematologic malignancies (Figure 1B). Over 99% of the sequences are from human data (Figure 1C), including peripheral lymphocytes of cancer patients ( $n = 538$ ) and healthy donors ( $n = 55$ ), tumor-infiltrating lymphocytes (TILs;  $n = 389$ ), and humanized mouse models ( $n = 6$ ). Somatic mutations in NeoTCR are mostly single nucleotide mutations (SNMs;  $n = 181$ ), with additional mutations including insertions and deletions (indels;  $n = 9$ ), splicing variants ( $n = 1$ ), and fusions ( $n = 2$ ).

### Functional description of NeoTCR

NeoTCR provides three functional modules (Figure 2): (1) the Browse module enables discovery and comparative analysis of TCR sequences across different conditions; (2) the Annotation module offers one-stop analysis of user-submitted TCR sequencing data to extract general features of TCR repertoires and interactively visualize the annotation results [including TCR diversity, complementarity-determining region 3 (CDR3) length distribution, frequency of CDR3 sequences, usage frequency of variable (V)/joining (J) gene segment, and V-diversity (D)-J gene segment utilization]; and (3) the CDR3 Alignment module, supports alignment of annotated CDR3 sequences with NeoTCR and other existing TCR databases.

### Application of NeoTCR

#### Performing a query for neoantigen-specific TCRs

NeoTCR provides a tool to investigate the associations between specific neoantigens and their cognate TCRs. To demonstrate the process, *TP53*-derived neoantigens were used as an example. A fuzzy search was conducted by entering “*TP53*” in the “Gene” query box (Figure 3A), which retrieved 23 entries of TCR sequences derived from TILs of patients with non-small cell lung cancer, ovarian cancer, or

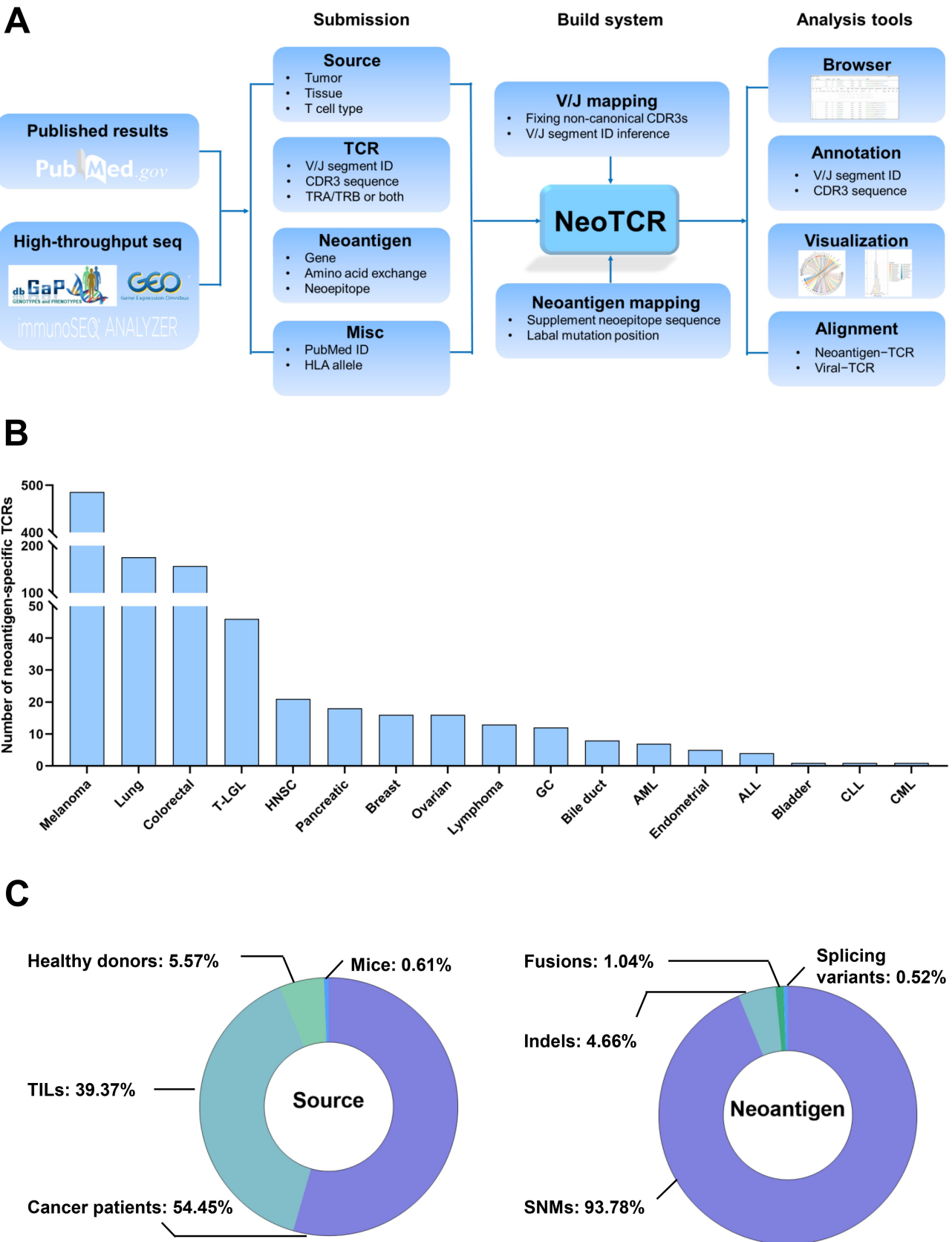
colorectal cancer. These TCRs were specific to 11 neoepitopes arising from six SNMs of *TP53* (R248L/T172I/G245S/R175H/Y220C/R248W) and thus exhibited distinct HLA restrictions (Figure 3B). Importantly, different cancer patients sharing the *HLA-A\*68:01* allele might have potent antitumor TCR rearrangements at the same somatic position (R248) but with different amino acid substitutions (L vs. W). The majority of R248L-specific TCR sequences shared identical TRAV/TRBV segments but displayed variability in the TRAJ/TRBJ segments. In contrast, the R248W-reactive TCRs demonstrated differences in TRAJ/TRBV segments, while sharing identical TRAV/TRBJ segments (Figure 3C). The V and J segments of the  $\alpha$  and  $\beta$  chains in R175H, however, did not show any similarities (Table 1). Notably, for a single *TP53* mutation, the CDR3 sequences of TRA or TRB were different, despite the similarity in V or J segments.

### TCR repertoire annotation and clonotype visualization

To illustrate the functionality of NeoTCR in annotating high-throughput TCR sequencing data, we utilized bulk TRA sequencing data from neoantigen-induced T cell subsets, which were isolated from peripheral blood mononuclear cells of a cancer patient, as an example. Bulk TCR sequencing data were loaded using the “Upload file” option in the “Annotation” window. The analysis using the “5’ Race” option produced 203 TRA sequences, each representing 0.000072% to 3.036% of the total sequences (Table S1). The CDR3 length was primarily restricted to 30–45 amino acids (Figure 4A). The top 5 CDR3 sequences were “CAGDDRRGGYNKLIF”, “CAHTGANSKLTF”, “CAASAWGGADGLTF”, “CAGNQFYF”, and “CAVMDSSYKLIF” (Figure 4B). The most frequently used V and J segments were TRAV29/DV5 and TRAJ45, respectively (Figure 4C and D), and these two segments also exhibited the highest combination frequency (Figure 4E).

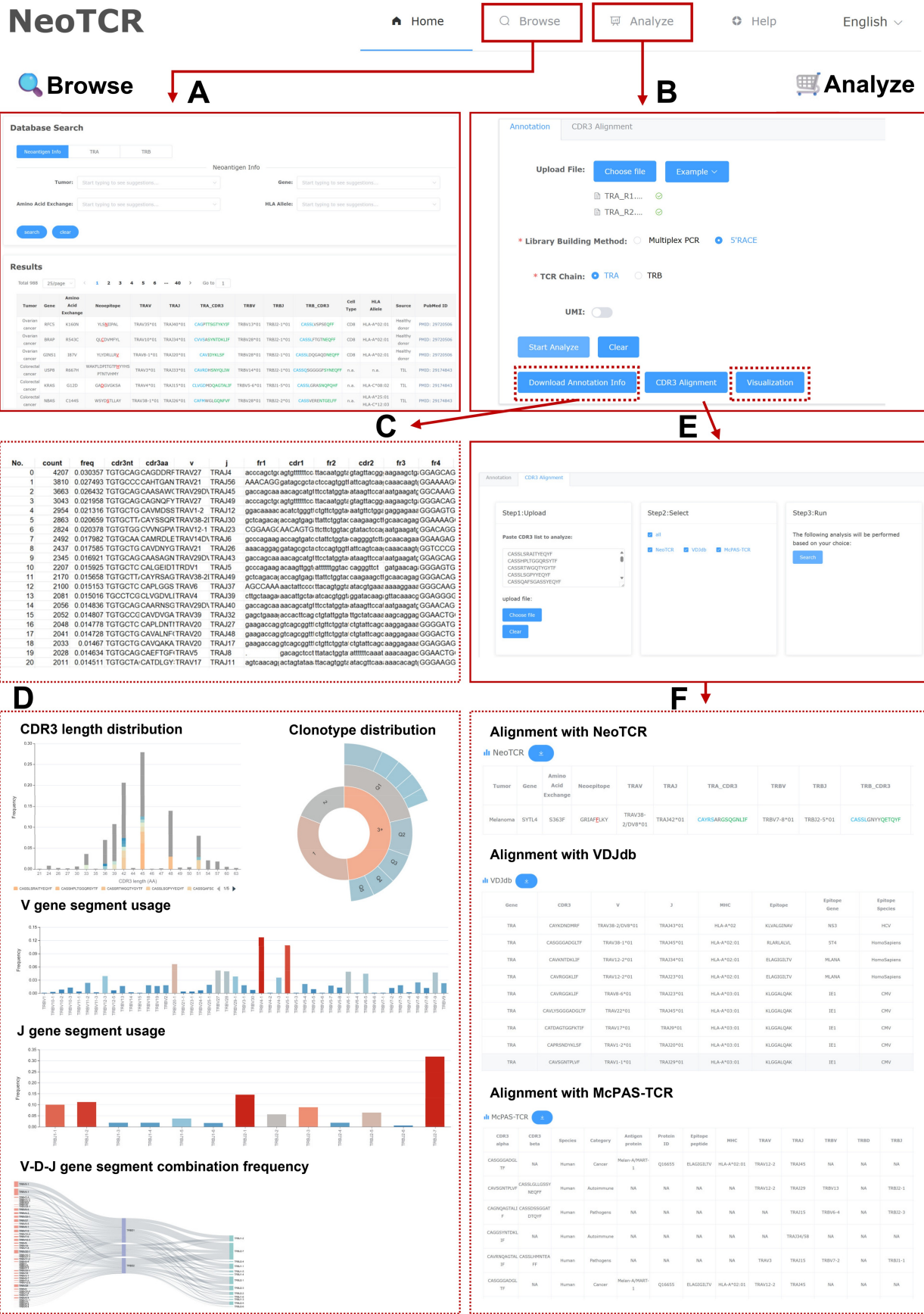
### CDR3 sequence alignment

Distinguishing neoantigen-specific TCRs from virus-associated TCRs can streamline the verification process of TCR binding and antigen recognition. CDR3, a product of the TCR rearrangement, is in charge of antigen recognition and thus forms the cornerstone of TCR specificity. The “CDR3 Alignment” section in NeoTCR was used to analyze the annotated CDR3 sequences derived from bulk TCR sequencing data of a specific patient (Figure 5A). The analysis showed that “CAYRSARGSQGNLIF” targeted the neoantigen *SYTL4*<sup>S363F</sup> in an *HLA-B\*27:05*-dependent manner, and provided information on corresponding TRB (Figure 5B; Table S2). A previous study has reported that this *SYTL4*<sup>S363F</sup>-specific TCR can induce immune responses to cancer cells expressing both *SYTL4*<sup>S363F</sup> and *HLA-B\*27:05* [21]. Based on such information, T-cell therapy targeting this paired TCR may be appropriate for this cancer patient. In this example, bystander virus-associated TCRs, including those from the cytomegalovirus (CMV), Epstein–Barr virus (EBV), yellow fever virus (YFV), and influenza A, were intermingled. CMV-associated TCRs were prevalent in the alignment results from VDJdb and McPAS-TCR, particularly those specific to pp65 (Figure 5C; Tables S3 and S4). Notably, 2 and 15 distinct TCRs were found in McPAS-TCR and VDJdb, respectively (Figure 5C).



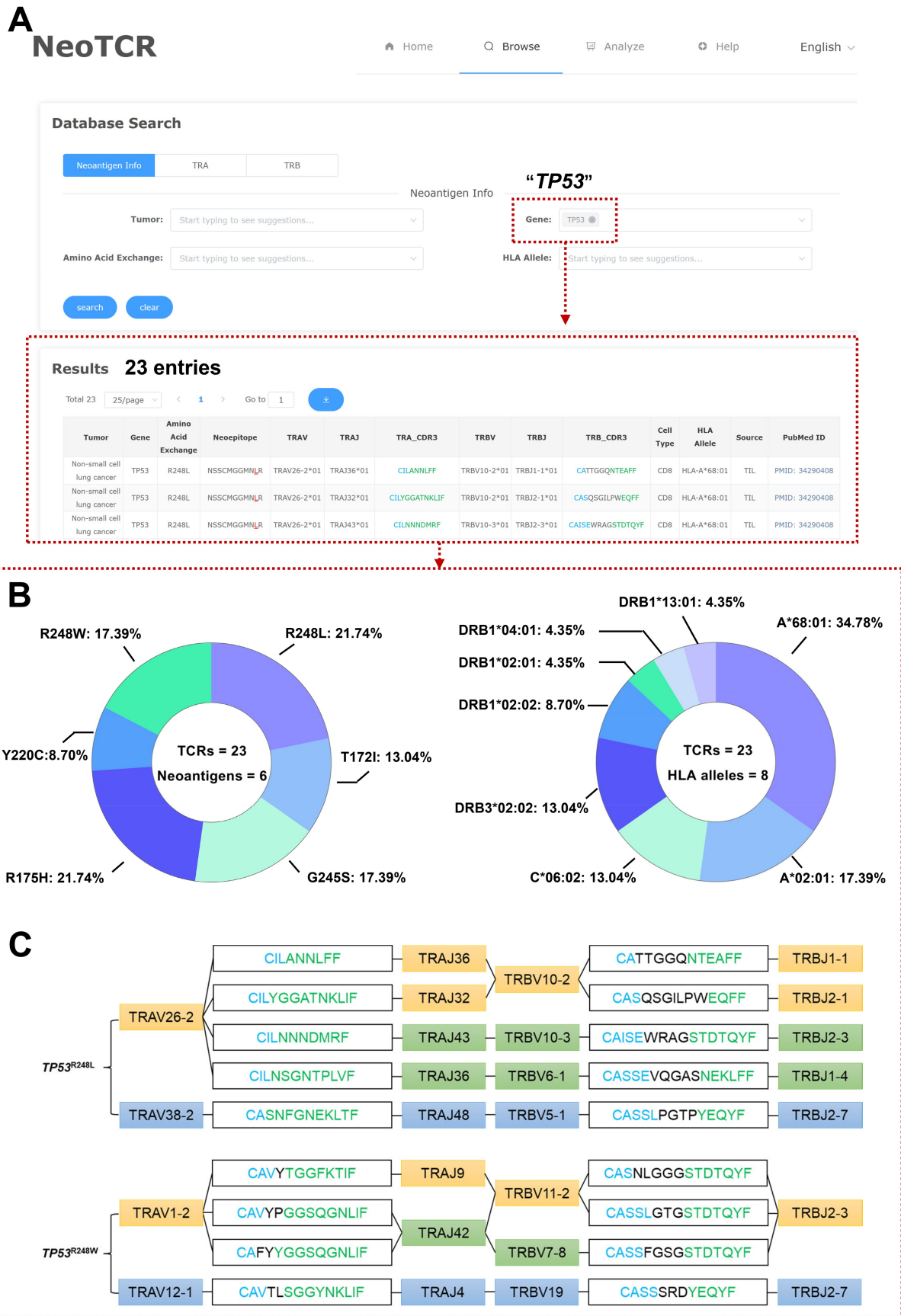
**Figure 1 Overview of NeoTCR**

**A.** Overall design and construction of NeoTCR. **B.** Number of neoantigen-specific TCRs across different cancer types. **C.** Source of neoantigen-specific TCRs and TCR-related neoantigen compositions in NeoTCR. T-LGL, T large granular lymphocyte leukemia; HNSC, head and neck squamous cell carcinoma; GC, gastroesophageal cancer; AML, acute myeloid leukemia; ALL, acute lymphoblastic leukemia; CLL, chronic lymphocytic leukemia; CML, chronic myeloid leukemia; TCR, T cell receptor; CDR3, complementarity-determining region 3; V, variable; J, joining; D, diversity; ID, identity document; HLA, human leukocyte antigen; TRA, T cell receptor alpha; TRB, T cell receptor beta; TIL, tumor-infiltrating lymphocyte; SNM, single nucleotide mutation; indel, insertion and deletion; seq, sequencing.



**Figure 2 Functional web pages of NeoTCR**

**A.** The Browse page: flexible searching of TCR sequences and their corresponding neoantigens, with results displayed in user-friendly interactive pages. **B-F.** The Analyze page: personalized TCR sequencing data analysis in the Annotation module (B), including downloading the annotation information in .txt format (C) and viewing the TCR clonotypes on the interactive page (D), or in the CDR3 Alignment module (E) to align the annotated CDR3 sequences with NeoTCR and other datasets in more detail (F).



**Figure 3 Searching for TCRs specific to a particular neoantigen**

**A.** The search with the keyword "TP53" yielded 23 entries. **B.** Distribution of neoantigens and HLA alleles in TP53 mutation-specific TCRs. **C.** Characteristics of TCR sequences targeting TP53<sup>R248L</sup> and TP53<sup>R248W</sup>. Blue and green boxes correspond to V and J gene segments, respectively; orange box includes the entire CDR3 sequence. Conserved residues of the relevant V and J gene segments are shown in blue and green, respectively.

**Table 1 Characteristics of TCR sequences specific to TP53<sup>R175H</sup>**

Neopeptide	V gene	J gene	CDR3 sequence	HLA allele	PMID
HMTEVVR <u>H</u> C	TRAV12-3 TRBV27	TRAJ12 TRBJ2-3	CAMSLKEDSSYLIF CASSIQQGADTQYF	HLA-A*02:01	30714987
VVR <u>H</u> CPHHERCSDSD	TRAV24	TRAJ10	CALITGGGNKLT	HLA-DRB1*13:01	30714987
QHMT <u>E</u> VVR <u>H</u> CPHHER	TRBV6-2	TRBJ1-6	CASRLQGWSPHFF		
HMTEVVR <u>H</u> C	TRAV12-1 TRBV6-1	TRAJ13 TRBJ2-7	CVVGGYQKVTF CASSEEQYF	HLA-A*02:01	30709841
HMTEVVR <u>H</u> C	TRAV6 TRBV11-2	TRAJ43 TRBJ2-2	CALDDMRF CASSLTGELFF	HLA-A*02:01	30709841
HMTEVVR <u>H</u> C	TRAV38-1 TRBV10-3	TRAJ28 TRBJ1-6	CAFMYSGAGSYQLTF CAISESPLHF	HLA-A*02:01	30709841

Note: The mutated amino acid is underlined. TCR, T cell receptor; CDR3, complementarity-determining region 3; V, variable; J, joining; D, diversity; HLA, human leukocyte antigen; PMID, PubMed Identifier; TRA, T cell receptor alpha; TRB, T cell receptor beta.

## Database construction

### Data collection and processing

We conducted a PubMed search using the following terms: [T cell receptor (Title/Abstract)] OR [TCR (Title/Abstract)] OR [CDR3 sequence (Title/Abstract)] OR [T cell repertoire (Title/Abstract)] OR [T cell (Title/Abstract)] OR [tumor-infiltrating lymphocytes (Title/Abstract)] AND (neoantigen). Additional publications were obtained by identifying the references cited by the articles from the electronic search. For the TCR sequences lacking the V and/or J specifications or containing incomplete or excessive CDR3 sequences, the raw data (if available) were downloaded and submitted to the pipeline to generate the annotated information. Generally, the neoantigen-related data in NeoTCR were collected from published experimental studies.

### Data organization

Each entry in the database contains one TCR sequence and is described by the following fields:

#### Tumor

The disease in which the TCR sequence was identified.

#### Gene, amino acid exchange, and neopeptide

These fields describe the characteristics of TCR-associated neoantigens: “gene” means the antigen protein that the TCR targets; “amino acid exchange” means the amino acid somatic mutation from the corresponding wild-type antigen protein, including SNM, indel, fusion, and splicing; “neopeptide” means the amino acid sequence of a neopeptide that the TCR binds to and recognizes, with the mutated amino acid underscored.

#### TRAV, TRAJ, and TRA\_CDR3 as well as TRBV, TRBJ, and TRB\_CDR3

These fields describe the information of TCR $\alpha$  and TCR $\beta$  chains. “TRAV” and “TRBV” mean the V gene segment identity documents (IDs) of TCR $\alpha$  and TCR $\beta$ , respectively. “TRAJ” and “TRBJ” mean the J gene segment IDs of TCR $\alpha$  and TCR $\beta$ , respectively. Each segment ID strictly follows the international immunogenetics information system (IMGT) nomenclature. “TRA\_CDR3” and “TRB\_CDR3” mean the CDR3 amino acid sequences of TCR $\alpha$  and TCR $\beta$ , respectively. The complete CDR3 sequence starts with the conserved cysteine (C) at the 5' end of the V segment and ends with the conserved phenylalanine (F) or tryptophan (W) at the 3' end of the J segment. Trimmed or excessive sequences

are fixed at the data processing stage if sufficient V/J germline parts are present.

#### Cell type

This field describes the lymphocyte subtype of the TCR, e.g., CD4 and CD8.

#### HLA allele

This field describes the HLA restriction of the TCR-associated neopeptide, as outlined in the original article.

#### Source

This field describes the tissue source of the TCR sequence, e.g., healthy donor (peripheral lymphocytes of the healthy donor), TIL, patient (peripheral lymphocytes of the cancer patient), or humanized mouse model of tumors.

#### PubMed ID

This field describes the PubMed Identifier (PMID) of the study's entry in the PubMed biomedical literature citation database (<https://pubmed.ncbi.nlm.nih.gov/>). A link to the publication is provided.

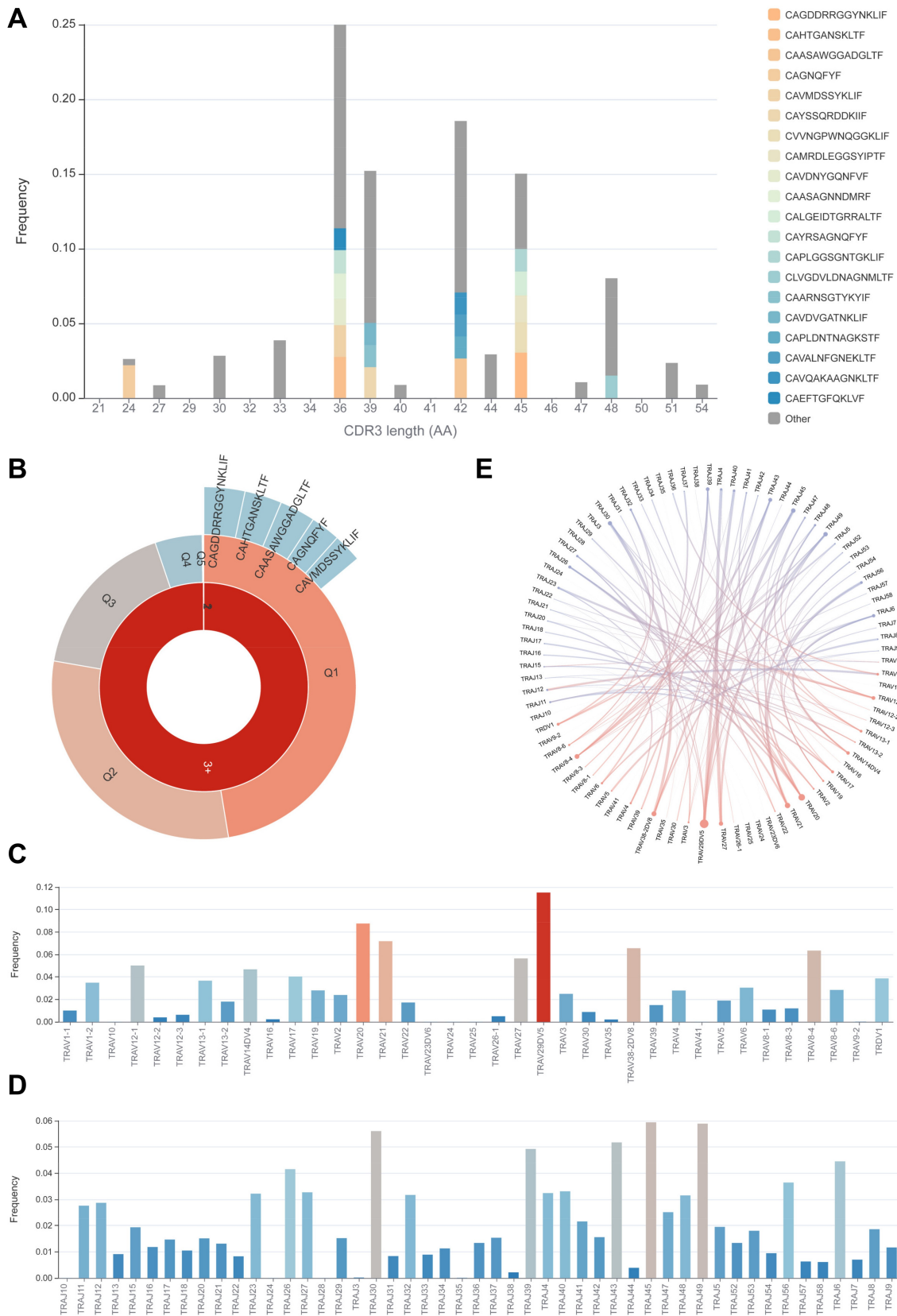
## Database analysis tools

### Annotation tool

A standardized pipeline for annotating TCR $\alpha$  and TCR $\beta$  chains was developed in NeoTCR as described below. First, raw reads of TCR sequencing were trimmed by Trimmomatic [22] to remove low-quality bases, unknown bases labeled as N, and adapter contamination. Subsequently, quality control was performed using FastQC [23]. Then UMI-tools [24] was used to process the data containing unique molecular identifiers (UMIs), which enabled the correction of polymerase chain reaction (PCR) amplification biases and quantification of expressed receptor numbers. To generate the annotation information, MiXCR [25] was employed to align sequencing reads to reference V, D, J, and constant (C) segments of the TCR chain and assemble clonotypes.

### Visualization tool

The follow-up analysis of the annotation information was performed using Python and ECharts [26] to implement data visualization. Briefly, the annotated data were grouped into statistics in different dimensions to analyze the TCR repertoires. The visualization, including the counts of V, D, and J genes of each TCR chain, the V-J gene usage of the TCR $\alpha$  chain, the V-D-J gene usage of the TCR $\beta$  chain, the CDR3



**Figure 4 Visualization of the annotated TCR sequences**

**A.** Length distribution of CDR3 sequences. **B.** Frequency of CDR3 sequences. The summed frequencies of TCR clones with 1, 2, and  $\geq 3$  sequence reads in the sample are represented by the radian of 1, 2, and 3+ in the inner layer, respectively. The distribution of quintiles (Q1–Q5) for TCR clones is shown in the middle layer. The outer layer displays the top 5 CDR3 amino acid sequences, with the frequency of corresponding TCR clones indicated by the radian of each region. **C.** and **D.** Usage frequency of V gene segments (C) and J gene segments (D). Colors represent frequency variation, ranging from deep blue (low frequency) to bright red (high frequency). **E.** Combination possibility of V and J gene segments.

# NeoTCR



**Figure 5 Performing CDR3 alignment**

**A.** Alignment using the TRA sequences annotated from the sample raw bulk TCR sequencing data. **B.** Discovered neoantigen-specific TRA sequence and the corresponding TRB sequence. **C.** Bystander virus-associated TCR sequences intermingled with different frequencies in this example data. CMV, cytomegalovirus; EBV, Epstein-Barr virus; YFV, yellow fever virus.

length, and clonotype distribution, was performed using ECharts [26].

### Alignment tool

CDR3 alignment was applied to label the known neoantigen-specific TCRs and bystander virus-associated TCRs. In brief, the query CDR3 sequences were aligned with the columns “TRA\_CDR3”/“TRB\_CDR3” of NeoTCR, “CDR3” of VDJdb [17,18], and “CDR3.alpha.aa”/“CDR3.beta.aa” of McPAS-TCR [19]. The information of TCR sequences, as well as the target genes and antigens, was returned when the query sequences matched the CDR3 sequences in the reference database.

### Discussion

TCR diversity arises from the random recombination of the V, D, and J gene segments and the pairing of TCR $\alpha$  and TCR $\beta$  chains. A vast amount of TCRs with neoantigen specificity have been identified from tumor specimens or peripheral blood [4,6,27]. The NeoTCR database was developed to accommodate the need to house the TCR sequences associated with neoantigens.

The NeoTCR database and the web server have the potential to assist biomedical researchers in developing neoantigen-induced immune responses, particularly in designing personalized TCR-T cell-based immunotherapy for cancer patients. For instance, researchers can compare TCR repertoires with identical CDR3 sequences (e.g., paired TCRs with the same TRA\_CDR3 sequence but different TRB\_CDR3, or TCR $\alpha$ / $\beta$  chains with the same CDR3 sequence but different V and/or J gene segments). For a specific clinical scenario, such as an HLA-DRB3\*02:02-positive patient harboring TP53<sup>G245S</sup> mutation, clinicians can search NeoTCR and quickly find specific TCRs (TRAV8-1\*01-CAVKGDYKLSF-TRAJ20\*01 and TRBV11-2\*01-CASSLVNTEAFF-TRBJ1-1\*01) for developing personalized TCR-T cell-based immunotherapy for this patient. NeoTCR offers two web tools for analyzing raw TCR sequencing data: “Annotation” for extracting general features of TCR repertoires and further visualizing the clonotypes, and “CDR3 Alignment” for aligning annotated CDR3 sequences with NeoTCR and other existing TCR databases to identify known neoantigen-specific TCRs and exclude bystander virus-associated TCRs.

We anticipate a rapid increase in the discovery of neoantigen-related TCR sequences. We are continuously striving to acquire more neoantigen-specific TCR sequences from raw high-throughput sequencing data and will periodically update NeoTCR. We are also planning to use deep learning technology to build mathematical models for predicting neoantigen-TCR binding for human TCRs.

### Data availability

NeoTCR is accessible at <http://neotcrdb.bioxai.cn/> and <https://github.com/lyotvincent/NeoTCR>, and the guidelines for NeoTCR are available at <http://neotcrdb.bioxai.cn/help>. NeoTCR has also been submitted to Database Commons [28] at the National Genomics Data Center (NGDC), China National Center for Bioinformatics (CNGB), which is publicly accessible at <https://ngdc.cnbc.ac.cn/databasecommons/database/id/10220>.

### CRedit author statement

**Weijun Zhou:** Conceptualization, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Wenting Xiang:** Methodology, Software, Data curation, Writing – original draft, Writing – review & editing. **Jinyi Yu:** Resources. **Zhihan Ruan:** Software. **Yichen Pan:** Data curation. **Kankan Wang:** Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Jian Liu:** Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Project administration. All authors have read and approved the final manuscript.

### Competing interests

The authors have declared no competing interests.

### Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzae010>).

### Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 81890994), the National Key R&D Program of China (Grant No. 2019YFA0905902), the Guangdong Basic and Applied Basic Research Foundation, China (Grant Nos. 2019A1515010299 and 2023A1515030298), the Science and Technology Program of Guangzhou, China (Grant No. 202102020727), the Innovative Research Team of High-level Local Universities in Shanghai, and the Dean’s fund of Zhujiang Hospital of Southern Medical University, China (Grant No. yzj2019qn05). We gratefully acknowledge Kellie N. Smith for sharing TCR sequencing data, as well as the developers of VDJdb and McPAS-TCR for providing access to their databases.

### ORCID

0000-0002-0137-7382 (Weijun Zhou)  
 0000-0002-1313-170X (Wenting Xiang)  
 0000-0002-4548-8447 (Jinyi Yu)  
 0000-0002-2902-0852 (Zhihan Ruan)  
 0000-0002-4541-4261 (Yichen Pan)  
 0000-0001-7198-2134 (Kankan Wang)  
 0000-0001-5516-0157 (Jian Liu)

### References

- [1] 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.
- [2] Carlino MS, Larkin J, Long GV. Immune checkpoint inhibitors in melanoma. *Lancet* 2021;398:1002–14.
- [3] Tran E, Robbins PF, Lu YC, Prickett TD, Gartner JJ, Jia L, et al. T-cell transfer therapy targeting mutant KRAS in cancer. *N Engl J Med* 2016;375:2255–62.
- [4] 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.

- [5] Zacharakis N, Chinnasamy H, Black M, Xu H, Lu YC, Zheng Z, et al. Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nat Med* 2018;24:724–30.
- [6] Caushi JX, Zhang J, Ji Z, Vaghasia A, Zhang B, Hsiue EH, et al. Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. *Nature* 2021;596:126–32.
- [7] Krishna S, Lowery FJ, Copeland AR, Bahadiroglu E, Mukherjee R, Jia L, et al. Stem-like CD8 T cells mediate response of adoptive cell immunotherapy against human cancer. *Science* 2020;370:1328–34.
- [8] Kristensen NP, Heeke C, Tvingsholm SA, Borch A, Draghi A, Crowther MD, et al. Neoantigen-reactive CD8<sup>+</sup> T cells affect clinical outcome of adoptive cell therapy with tumor-infiltrating lymphocytes in melanoma. *J Clin Invest* 2022;132:e150535
- [9] Lowery FJ, Krishna S, Yossef R, Parikh NB, Chatani PD, Zacharakis N, et al. Molecular signatures of antitumor neoantigen-reactive T cells from metastatic human cancers. *Science* 2022;375:877–84.
- [10] Calis JJ, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol* 2014;35:581–90.
- [11] Friedensohn S, Khan TA, Reddy ST. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol* 2017;35:203–14.
- [12] Pai JA, Satpathy AT. High-throughput and single-cell T cell receptor sequencing technologies. *Nat Methods* 2021;18:881–92.
- [13] Gielis S, Moris P, Bittremieux W, De Neuter N, Ogunjimi B, Laukens K, et al. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Front Immunol* 2019; 10:2820.
- [14] 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 TCR $\alpha$  and  $\beta$  sequence data. *Commun Biol* 2021;4:1060.
- [15] Levin N, Paria BC, Vale NR, Yossef R, Lowery FJ, Parkhurst MR, et al. Identification and validation of T-cell receptors targeting RAS hotspot mutations in human cancers for use in cell-based immunotherapy. *Clin Cancer Res* 2021;27:5084–95.
- [16] Yossef R, Tran E, Deniger DC, Gros A, Pasetto A, Parkhurst MR, et al. Enhanced detection of neoantigen-reactive T cells targeting unique and shared oncogenes for personalized cancer immunotherapy. *JCI Insight* 2018;3:e122467.
- [17] Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJDdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* 2018;46:419–27.
- [18] Bagaev DV, Vroomans R, Samir J, Stervbo U, Rius C, Dolton G, et al. VDJDdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res* 2020;48:1057–62.
- [19] Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* 2017;33:2924–9.
- [20] Chen SY, Yue T, Lei Q, Guo AY. TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. *Nucleic Acids Res* 2021;49:468–74.
- [21] Braunlein E, Lupoli G, Fuchsl F, Abualrous ET, de Andrade Kratzig N, Gosmann D, et al. Functional analysis of peripheral and intratumoral neoantigen-specific TCRs identified in a patient with melanoma. *J Immunother Cancer* 2021;9:e002754.
- [22] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–20.
- [23] de Sena Brandine G, Smith AD. Falco: high-speed FastQC emulation for quality control of sequencing data. *F1000Res* 2019; 8:1874.
- [24] Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* 2017;27:491–9.
- [25] Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* 2015;12:380–1.
- [26] Li DQ, Mei HH, Shen Y, Su S, Zhang WL, Wang JT, et al. ECharts: a declarative framework for rapid construction of web-based visualization. *Visual Informatics* 2018;2:136–46.
- [27] Yossef R, Tran E, Deniger DC, Gros A, Pasetto A, Parkhurst MR, et al. Enhanced detection of neoantigen-reactive T cells targeting unique and shared oncogenes for personalized cancer immunotherapy. *JCI Insight* 2018;3:e122467.
- [28] Ma L, Zou D, Liu L, Shireen H, Abbasi AA, Bateman A, et al. Database Commons: a catalog of worldwide biological databases. *Genomics Proteomics Bioinformatics* 2023;21:1054–8.